J. Mol. Evol. 11, 143–161 (1978)

Journal of Molecular Evolution

A Method for Detecting Distant Evolutionary Relationships Between Protein or Nucleic Acid Sequences in the Presence of Deletions or Insertions

T.C. Elleman

Division of Protein Chemistry, CSIRO, 343 Royal Parade, Parkville (Melbourne), Victoria 3052, Australia

Summary. A method for detecting homology between two protein or nucleic acid sequences which require insertions or deletions for optimum alignment has been devised for use with a computer. Sequences are assessed for possible relationship by Monte Carlo methods involving comparisons between the alignment of the real sequences and alignments of randomly scrambled sequences of the same composition as the real sequences, each alignment having the optimum number of gaps. As each gap is successively introduced into a comparison (real or random) a maximum score is determined from the similarity of the aligned residues. From the distribution of the maximum alignment scores of randomly scrambled sequences having the same number of gaps, the percentage of random comparisons having higher scores is determined, and the smallest of these percentage levels for each pair of sequences (real or random) indicates the optimum alignment. The fraction of the comparisons of random sequences having percentage levels at their optimum alignment below that of the real sequence comparison at its optimum estimates the probability that such an alignment might have arisen by chance. Related sequences are detected since their optimum alignment score, by virtue of a contribution from ancestral homology in addition to optimised random considerations, occupies a more extreme position in the appropriate frequency distribution of scores than do the majority of optimum scores of randomly scrambled sequences in their appropriate distributions.

Application of this 'optimum match' method of sequence comparison shows that the sensitivity of the 'maximum match' method of Needleman and Wunsch (1970) decreases quite dramatically with sequence comparisons which require only a few gaps for a reasonable alignment, or when sequences differ greatly in length. The 'maximum match' method as applied by Barker and Dayhoff (1972) has the additional disadvantage that deletions which have occurred in the longer of two homologous protein sequences further decrease the sensitivity of detection of relationship. The 'constrained match' method of Sankoff and Cedergren (1973) is seen to be misleading since large increments in the alignment score

0022–2844/78/0011/0143/$ 03.80

from added gaps do not necessarily result in a high total alignment score required to demonstrate sequence homology.

Key words: Homology - Gaps - Matrix - Protein - Nucleic acid - Sequence

Introduction

The sequences of proteins or nucleic acids can be used to establish relationships between organisms and to demonstrate similar genetic origins for macromolecules of different functions. Even for closely related molecules, however, when no test of significance is required to show a relationship, it is rarely possible, without using gaps, to align the sequences directly, so justifing the existence of these interruptions or gaps in alignments of homologous sequences. With less closely related molecules the probability of obtaining that degree of similarity by chance alone must be considered, and the evaluation of this probability when gaps have proven necessary for the optimum sequence alignment becomes a formidable task. This difficulty has been somewhat overcome either by examining the distribution of alignment scores of all small spans of the sequence, so minimizing the effect of a register shift (Fitch, 1966; McLachlan, 1971), or by allowing every possible sequence gap in forming the 'maximum match' (Neeleman and Wunsch, 1970), assessing the results in each case by Monte Carlo techniques. Variations on the latter method have shown how the sensitivity could be enhanced by: (a) a penalty factor subtracted from the 'maximum match' for each gap used to form it (Needleman and Wunsch, 1970); (b) a more discerning scoring system from 0 to 28 for all possible residue comparisons in proteins (Barker and Dayhoff, 1972); (c) assessing the individual significance of each gap added, by the increase in the number of correspondences (Sankoff and Cedergren, 1973). Whilst the former type of approach of Fitch (1966) and McLachlan (1971) is more versatile since it can be used to detect internal homology within a sequence or crossed homology between opposite ends of two sequences, the assessment of the significance of the distribution of span scores is difficult (Fitch, 1970). The latter types of approach based on the method of Needleman and Wunsch are only successful when regions of homology are mutually consistent in direction, but the significance of relationship is easily assessed, and generally much higher than established by the former method. The 'optimum match' method described here is one of the latter type, but assesses the possibility of relationship when all sequence comparisons have the optimum number of gaps in the alignment as determined by comparing the percentage levels reached by each sequence alignment in the frequency distribution of maximum alignment scores at each number of gaps.

In the 'optimum match' method, gaps are added individually and successively to each sequence comparison, and the maximum alignment score is determined at each addition. The same procedure applied to comparisons of randomly scrambled sequences of the same compositions as the real sequences yields a frequency distribution of alignment scores for each number of gaps, so that a percentage level for the proportion of alignment scores from random comparisons which exceeded the score of the real sequence comparison is obtained at each number of added gaps. The lowest percentage level reached in the comparison of real sequences then indicates the optimum

alignment, and provides a critical level which, in order that these sequences might be considered significantly better related than random sequences of the same composition, must be smaller than the lowest levels reached by all but a few comparisons of random sequences when these are assessed by the same procedure i.e. each random sequence at its own optimum alignment.

1. The Method of Sequence Comparison and the Introduction of Gaps

The maximum scores attainable by the alignment of two sequences allowing from 0 to n gaps are determined. The unit of comparison is a pair of amino acids or nucleotides, one from each sequence. When two sequences of length N and M are written on adjacent sides of a two-dimensional array, a comparison of all elements is possible, each cell of the matrix S(i,j) of N rows and M columns representing the alignment of the ith and jth residue from each sequence respectively[1]. Every possible alignment of the two sequences is represented by a pathway through the array which proceeds from cell to cell such that both indices increase simultaneously. If both indices increase by one, the residues from each sequence at i + 1 and j + 1 following the alignment of the ith and jth residues are also aligned, while if either or both indices increase by more than one, so breaking the continuity of the diagonal pathway, a gap or directly adjacent gaps respectively follow the aligned ith and jth residues.

A correspondence array is first used to assign scores to each cell of the matrix S(i,j), in accordance with the similarity of the ith and jth elements, and the maximum scores attainable for alignments containing from 0 to n gaps are then determined from transformations of this matrix. These scores are assessed later to determine the optimum alignment.

A zero gap score matrix A_0, to find the best alignment score without gaps is generated from the score matrix S. The cells of this array $A_0(i,j)$ show the maximum score attainable by aligning those portions of the sequences beginning with the ith and jth residue of each sequence and extending the alignment without a gap until the terminal residue of either of the chains is reached i.e.

$$A_0(i,j) = \sum_{k=0}^{\text{end of diagonal}} S(i+k, j+k)$$

The score matrix S is transformed to the zero gap score matrix A_0 by starting from the penultimate cell of the penultimate row and proceeding row-by-row from the penultimate cell of each row to S(1,1) replacing the value in each cell, S(i, j) by the total so far accumulated in $A_0(i + 1, j + 1)$ plus the value in S(i, j).

$$(1) \qquad A_0(i,j) = A_0(i+1, j+1) + S(i,j)$$

---

[1] Since peptide and oligonucleotide bonds are directional, proteins or nucleic acids are conventionally numbered from amino or 5' terminal residues. 'End' or 'terminal residue', then refer to the carboxyl or 3 terminal residue in the context of this manuscript.

The values in the first row and column then represent the simple frameshift scores, and the largest of these is stored for later conversion to a percentage level using the frequency distribution of the alignment scores of random sequences of the same composition and containing no gaps.

The zero gap score matrix A_0, is next converted to a score matrix A_1, each cell of which contains the maximum score attainable by aligning those portions of the sequences beginning with the ith and jth residues and permitting a single gap of any length, anywhere in the subsequent alignment. Since the score in each cell A_0 (i,j) of the zero gap score matrix is the maximum score attainable by alignment of those parts of the sequences beginning with the ith and jth residues and proceeding directly to the terminal residues without a gap, then the maximum score for an alignment originating at the ith and jth residues and containing a gap of any length immediately after the alignment of these residues is simply the sum of the correspondence score for these residues plus the maximum value in the zero gap score matrix A_0 for row (i + 1) with column values greater than j, or for column (j + 1) with row values greater than i. Thus starting at the penultimate cell of the penultimate row of A_0 and applying this search procedure to all cells converts the zero gap score matrix to an intermediate matrix containing in each cell B_1 (i, j) the score for an alignment originating at the ith and jth residues and containing a single gap of any length immediately after either residue.

$$B_1(i,j) = \max\left\{\max_{M \geqslant k > j}[A_0(i+1,k)],\ \max_{N \geqslant l > i}[A_0(l,j+1)]\right\} + S(i,j)$$

The transformation of this intermediate matrix to A_1 is accomplished by commencing at the penultimate cell of the penultimate row and proceeding row-by-row to the origin within the B_1 matrix, comparing each cell B_1(i, j) with the sum of B_1 (i + 1, j + 1) plus the simple correspondence score for the ith and jth residue, and replacing B_1 (i, j) with the larger of these values, i.e.:

$$B_1(i,j) \leftarrow \max\left\{B_1(i,j),\ B_1(i+1,j+1) + S(i,j)\right\}$$

Since the transformation procedure starts at the penultimate row of the matrix and proceeds through the whole matrix, this procedure results in the score in each cell being that for the pathway originating at that cell with the break in the diagonal pathway optimally placed. Thus the largest value in the first row or column becomes the maximum score attainable for alignment of the complete sequences with a single gap optimally placed, and the transformed matrix is the required matrix A_1.

The matrix A_1 now replaces the matrix for zero gaps, and the whole transformation procedure, when repeated, using A_1 in place of A_0, converts this matrix for a single gap to the matrix for 2 gaps and so on. The highest score in row 1 or column 1 of each matrix being the maximum alignment score possible for the number of gaps permitted.

The matrix transformations starting from A_0 may be briefly summarized for the introduction of n gaps (n = 1, 2, 3...) to two sequences of length M and N by the following set of recursive equations. Each step is performed for i from N-1 to 1 and for each decrement of i, j decreases from M-1 to 1.

$$(2)\quad B_n(i,j) = \max\left\{\max_{M \geqslant k > j} [A_{n-1}(i+1,k)],\ \max_{N \geqslant l > i} [A_{n-1}(l, j+1)]\right\} + S(i,j)$$

$$(3)\quad B_n(i,j) \longleftarrow \max\left\{B_n(i,j),\ B_n(i+1, j+1) + S(i,j)\right\}$$

$$(4)\quad A_n(i,j) = B_n(i,j)$$

The complete procedure is performed identically on both real and random sequences, and all scores are expressed in standard measure by subtracting the estimated mean for the number of gaps permitted and dividing by the estimated standard deviation.

To perform these calculations with a computer, it is not necessary to store the S (i, j) values which can be calculated from the correspondence array and the i and j values. Also, in order to perform the transformations within a single two-dimensional array, it is necessary to store B_n (i, j) values for each row (i) when applying equation 2, until calculations for the next row preceding this have been completed, since the unchanged A_{n-1} values of each succeeding row are required for calculating each row of B_n.

(cf. $\max_{M \geqslant k > j} [A_{n-1}(i+1, k)]$ of equation 2)

Likewise the largest value in each column for all rows in A_{n-1} greater than the current value of i is also stored and continually reassessed at each row as the search proceeds.

(cf. $\max_{N \geqslant l > i} [A_{n-1}(l, j+1)]$ of equation 2).

With correspondence arrays having 0 as the lowest value, alignments with directly adjacent gaps can fare no better than if these gaps were separated by a poor alignment. However, if the correspondence array contains negative values then alignments with directly adjacent gaps can fare better. Such gaps would correspond with overlapping deletion mutations in the sequences. In order to include this possibility of directly adjacent gaps in an alignment when such a correspondence array is used, it is necessary to calculate a second matrix, C_n, which contains in each cell C_n (i, j), the maximum alignment score from the ith and jth residue to the terminal residues allowing one gap immediately after either of the ith and jth residues but excluding the correspondence score for these residues i.e.

$$C_n(i,j) = \max\left\{\max_{M \geqslant k > j} [A_{n-1}(i+1,k)],\ \max_{N \geqslant l > i} [A_{n-1}(l, j+1)]\right\}$$

for $i > 1;\ j > 1$

The additional choice of C_n values is added to the selection of the maximum in equation (2) when n = 2, 3, 4 ... , i.e.:

$$(2a)\quad B_n(i,j) = \max\left\{D_n(i,j),\ E_n(i,j),\ F_n(i,j),\ G_n(i,j)\right\} + S(i,j)$$

where $$D_n(i,j) = \max_{M \geqslant k > j} [A_{n-1}(i+1, k)]$$

$$E_n(i,j) = \max_{N \geqslant l > i} [A_{n-1}(l, j+1)]$$

$$F_n(i,j) = \max_{M \geqslant k > j} [C_{n-1}(i+1, k)]$$

$$G_n(i,j) = \max_{N \geqslant l > i} [C_{n-1}(l, j+1)]$$

By the choice of C_{n-1} values as an alternative in calculating B_n, the procedure permits the normally compulsory single residue alignment between adjacent gaps, one from each sequence, to be excluded if negative, while still retaining the indices of this position as a reference point in the C_n matrices.

A maximum scoring pathway and the corresponding alignment of sequences may, if required, be found from the matrix A_n by locating the position where the maximum score occurs in row 1 or column 1, subtracting from this score the correspondence score of the elements at this position S(i, j), and searching for the resultant value in the following cell with each index increased by unity, then applying the same subtraction and search procedure from this new cell. Should a required value not be found, a gap follows the residue alignment corresponding with A_n (i, j), and the value must be sought in the previous matrix A_{n-1} at A_{n-1} (i + 1, k) for $M \geqslant k > j + 1$, or A_{n-1} (l, j + 1) for $N \geqslant l > i + 1$. Failure to find the value here indicates directly adjacent gaps, and the value should be searched for in A_{n-2} at A_{n-2} (l, k) for $N \geqslant l > i + 1$, $M \geqslant k > j + 1$. The subtraction and search procedure is then continued from the new cell of the last matrix.

The procedure thus far described suffers from an end effect problem since pathways through the matrices A_n (n = 1,2,3 ...) begin and end with cells corresponding to the compulsory alignment of either of the first residues and either of the last residues. This constraint is removed so that a gap may extend beyond first or last residues, by altering equation (2a) to read:

$$\text{(2b)}\quad B_n(i,j) = \max\left\{0, D_n(i,j), E_n(i,j), F_n(i,j), G_n(i,j)\right\} + \begin{cases} \max[0, S(i,j)] & \\ \text{if i or j} = 1, & \\ S(i,j) & \\ \text{otherwise} & \end{cases}$$

and the search procedure for a maximum scoring pathway must be altered to include this possibility, since the first subtraction and search procedure will fail when such a gap occurs at the origin of the alignment. To do this, an additional search must be made for the maximum value found at A_n (i,j) of row 1 or coulumn 1, in matrix A_{n-1} at A_{n-1} (i +1,k) for $M \geqslant k > j$ and A_{n-1} (l, j + 1) for $N \geqslant l > i$. This is done subsequent to an unsuccessful search in matrix A_{n-1}, and prior to the search in matrix A_{n-2}.

A maximum scoring alignment may be determined from a maximum scoring pathway by matching the residues corresponding to the cells in the pathway. Examples in Appendix 2 illustrate the procedures of this section.

2. Evaluating the Significance of Alignment Scores

The addition of each gap to an alignment of two sequences will increase the score until a saturation point is reached; each new alignment however, although of higher score, may not necessarily be an improvement, and the optimum alignment is that of which the score was exceeded by the smallest percentage of scores in the frequency distribution of alignment scores having the same number of gaps. This use of the percentage points in each frequency distribution prevents the optimum number of gaps in a sequence comparison being obscured by the continuous score increase with the addition of each gap, and provides a critical level from the optimum alignment of the real sequences which can be used to assess the significance of this alignment.

The number of gaps associated with each optimum alignment will differ for each pair of sequences, real or random, hence the critical percentage level at the optimum alignment of the real sequences cannot be used directly to demonstrate a relationship since it is a preferential value, although if too high it discredits a relationship (Haber and Koshland, 1970). However, the use of positions in distributions rather than direct scores allows an assessment to be made about the optimum alignment of the real sequence comparison in relation to optimum alignments of the randomly scrambled comparisons, even though they contain a different number of gaps. Thus, the proportion of randomly scrambled sequence comparisons having an upper percentage level smaller than that of the critical percentage level reached by the real comparison at the optimum alignment estimates the true significance level of the relationship between the sequences.

The distribution of maximum alignment scores for each number of gaps is bell-shaped, but shows skewness and kurtosis which vary with the number of gaps and the composition of the sequences, so that the number of standard deviates cutting off equal areas in the distributions of maximum alignment scores also varies for each number of gaps. This prevents the use of any approximation by a normal frequency distribution, but to determine accurately each percentage level by Monte Carlo methods would involve several thousand random alignments, all requiring the maximum scores computed at each number of gaps, and so the cost would be prohibitive. Therefore, these levels have been estimated using Pearson curves[2] (Johnson et al., 1963) fitted to the $\sqrt{\beta_1}$ and β_2 values determined at each number of gaps from a sample of ninety alignment scores (Xj) of random sequences of the same composition as the real sequences.

---

[2]A system of frequency curves which provides a means of estimating the percentage levels of a distribution from a sample. If the scores are standarized as $\frac{X_j - \bar{X}}{m_2}$ then the distribution may be expressed in terms of the two parameters $\sqrt{\beta_1}$ and β_2 which are independent of the dispersion of the distribution, and describe its shape. A Pearson curve with these $\sqrt{\beta_1}$ and β_2 values can then be used to estimate the percentage levels of the distribution.

$$\sqrt{\beta_1} = \left(\frac{m_3^2}{m_2^3}\right)^{1/2}, \qquad \beta_2 = \frac{m_4}{m_2^2}$$

where $$m_r = \frac{1}{90} \sum_{j=1}^{90} (X_j - \bar{X})^r \text{ for } r = 2, 3, 4 \text{ and}$$

$$\bar{X} = \frac{1}{90} \sum_{j=1}^{90} X_j$$

The alignment scores from the comparison of real sequences are converted to percentage levels using the Pearson curves obtained for each number of gaps. The smallest upper percentage level reached is then used to determine a critical standardized score corresponding to this percentage level in each distribution. The proportion of the comparisons of random sequences which exceed this critical standardized score in any of their alignments estimates the significance level of the relationship between the real sequences.

3. Correspondence Array Values

Using a scoring system of 1 and 0 for identity and non-identity, a system particularly well suited for nucleic acid sequences, the score represents the number of identical elements aligned. This system is poorly suited to comparisons of distantly related protein sequences since it is known from the homologous sequences in protein families that many substitutions of a conservative nature may occur, so preserving the structural integrity of the molecules, without necessarily maintaining a large proportion of identities.

The scoring system of Barker and Dayhoff (1972) is related to the 'information value' of amino acid replacements (see Appendix 1), and so is suited to detection of homologous sequences in proteins. From the frequency of accepted point mutations in their proposed phylogenetic trees and the relative mutability of amino acids they constructed a mutation probability matrix from which, by matrix multiplication, they calculated the probability that one amino acid will be replaced by another after a fixed large amount of evolutionary change. For each possible amino acid substitution they determined the ratio of the probability of the replacement of one amino acid by another in distantly related proteins to the probability of the random occurrence of the other amino acid. The logarithm of this ratio, the 'information value', is the basis of their scoring system. Substitutions which occur more frequently than expected on a random basis have positive values, whilst those occurring less frequently than expected have negative values. To these values Barker and Dayhoff added a constant to ensure all values are positive and so maintain a distinction between negative values, all of which would otherwise have been omitted by gaps in the maximum

match. This procedure unfortunately creates a bias against the addition of gaps to the longer sequence, since fewer total residue alignments are then possible. This system is scoring system 1.

The 'optimum match' method described can be used with negative values and so the original information values can be employed, although it is preferable to subtract from these the mean correspondence score based on the composition of each pair of proteins (de Haën et al., 1976), so that the mean correspondence score is then zero, and a gap is consequently without direct effect on the scoring, i.e.

$$S(i,j) \longleftarrow S(i,j) - \frac{\sum_{i,j} S(i,j)}{N.M.}$$

This system is scoring system 2.

Since changes in function will often necessitate radical changes in corresponding residues of homologous proteins, scoring system 3 is limited to a consideration of only those residue changes which could indicate homology, scoring zero for all negative information values. This system again has a bias towards aligning the maximum number of residues, although the effect is less pronounced than with system 1.

4. Application

a) KB Carcinoma 5S RNA and Pseudomonas 5S RNA

Sankoff and Cedergren (1973) claim to have shown that the 5S RNA molecules of *P. fluorescens* and human KB carcinoma cells are homologous by using a 'constrained match' method of similar type to this, although they give no overall probability of relationship. These sequences have been examined and the observation confirmed that a significant increment in the number of matched residues occurs by the addition of both the third and fourth gap ($P = 0.04$, $P < 0.01$ respectively). The number of matched residues was increased from 36 to 59 when four gaps were added, an increase significant at the 0.01 level compared with the increase in the number of matched residues for the first four gaps added to comparisons of random sequences of the same compositions.

Although these individual increases for the third and fourth gap are high, the direct interpretation of them as probabilities of relationship would be erroneous since such an interpretation applies only to a single sample taken at random, and since several gaps are added to an alignment, the significance of a high score increment for any individual gap must be lowered accordingly. Similarly the probability calculated for the aggregate score increase on adding four gaps is a somewhat preferential value, since four gaps would certainly not result in an optimum score increment in most random comparisons.

Even though a significant increase in score might be considered to have occurred as seems to be the case, this is not in itself an indication of homology, it only indicates that a more profitable comparison may be made by allowing these gaps; if the total alignment score is still poor, homology has not been demonstrated. Estimates from

random comparisons showed that if four gaps were allowed in all alignments, the alignment score of the 5S RNA sequences (59) was exceeded by 2.5% of the random alignment scores, and this poor value is obtained at the optimum number of gaps for the 5S RNA sequences i.e. allowing these sequences preferential treatment.

The statistics of some random comparisons further emphasize the poor homology. Of ninety random comparisons, three exceeded the alignment score of the 5S RNA sequences at four gaps with scores of 60, whilst one equalled it. Of these, two (runs 16 and 37 of Table 1) had scores of 49 ($P < 0.005$) and 46 ($P < 0.025$) without gaps being required, (cf. 36 for the real 5S RNA sequences, random mean 38.7) while the two others (runs 11 and 25 of Table 1) required only a single gap to score 52 ($P < 0.01$) and 54 ($P < 0.005$) respectively compared with 44 for the real 5S RNA sequences. In addition to these obviously superior random alignments, six others reached the 2.5% level of a frequency distribution for at least one number of gaps when this critical level was expressed as a critical standardized score in each distribution of standardized alignment scores (Table 1).

Thus from the total ninety comparisons of randomly scrambled sequences the alignment scores for ten comparisons were significant at the 2.5% level at their various optimal number of gaps, and so were superior to the alignment of the real 5S RNA sequences.

Although the ancestral homology of these 5S RNA molecules is not in doubt since both bear a high degree of similarity to the 5S RNA of *E. coli*, the high degree of sequence homology implied by the method of Sankoff and Cedergren could be misleading, since ten of ninety random comparisons surpassed the critical percentage level

Table 1. Selected values of the standardized scores $[(X_i - \overline{X})/m_2]$ for those ten comparisons of scrambled 5S RNA sequences which exceeded the critical standardized scores (values exceeding this score are shown underlined)

No. of gaps	0	1	4	10	20	30	Saturation
Critical Standardized score[a]	2.3	2.4	2.2	2.2	1.9	1.8	1.8
Random comparison No.							
1	0.5	-0.3	0.4	1.9	2.1	1.2	1.1
11	-0.2	2.9	2.6	2.3	1.8	1.6	1.5
16	3.7	2.1	2.6	1.5	1.3	0.7	0.7
25	-0.2	3.7	2.6	1.1	0.0	-1.0	-1.0
30	-0.6	0.1	0.9	1.1	2.2	1.6	1.5
31	1.9	0.9	1.7	1.9	1.8	2.2	1.9
34	1.5	0.1	0.0	1.5	1.3	1.8	1.5
37	2.6	2.5	2.1	2.7	2.2	0.7	0.7
78	0.8	0.1	0.4	1.5	2.1	1.6	1.5
89	-0.6	-0.3	0.0	0.6	0.9	1.6	1.9
Real Comparison	-1.0	-0.3	2.2	1.5	1.8	1.2	1.0

[a]expressed in standard measure and obtained from the Pearson curves fitted to $\sqrt{\beta_1}$ and β_2 values calculated from the 90 alignment scores of the random sequence comparisons at each number of gaps

of the real comparison. The probability of relationship using the 1,0 scoring system is thus estimated at 0.11 (10/90), while even if the maximum number of gaps in an alignment is restricted, this probability increases very little to a maximum of 0.04 (4/90) at 4 gaps.

If the number of gaps in the alignments of the 5S RNA sequences is increased still further, the score increases to 74 at 16 gaps which is still barely at the 2.5% level. Further gaps however increase the score at a slower rate than most random runs until the scores at saturation with gaps show the alignment of the 5S RNA molecules (score 78 with 26 gaps) is only 1.0 standard deviation from the mean maximum score, and 21 of the 90 random alignments equalled or exceeded the real alignment score. Statistics measured at this point are identical with those obtained by the method of Needleman and Wunsch (1970).

The observations that alignment scores from some comparisons of random sequences reach high levels of significance, with few or no gaps, and then fall to low values at saturation with gaps (Table 1) or conversely that scores from some comparisons may reach high significance levels only after large numbers of gaps have been added, cast doubt on the criterion used by Needleman and Wunsch to assess homology viz. scores at saturation with gaps. Is this score a reliable indication of homology? Certainly related proteins would be expected to fare favourably since they have a definite advantage over random comparisons, but the decision as to whether sequences are homologous when made at a point so remote from the optimum alignment is very susceptible to type II error, that is accepting the null hypothesis that the alignment is random and so failing to detect homology.

Needleman and Wunsch suggest the possible use of a variable penalty for a gap with their method, and this could help to alleviate errors of judgement since comparisons of homologous sequences generally need fewer gaps than random comparisons to reach the maximum match. However, since a set of deviations from the random mean is obtained, generally first increasing and then decreasing with increasing penalty, interpretation is difficult, since the optimum value is highly preferential. In the assessment of significance made using the optimum number of gaps in all comparisons as by the method proposed here, the real sequence comparison is not preferentially treated.

Barker and Dayhoff (1972) have adopted the method of Needleman and Wunsch, but add a constant to all elements in the correspondence array, gaps still scoring zero. This also decreases the number of gaps in the maximum match. When comparing the 5S RNA sequences they use a constant of 2 i.e.: instead of scoring 1 and 0 for identity and non-identity, they score 3 and 2. This method once again, however, yields a series of values depending on the constant chosen, and simply quoting the optimum value, is giving the real sequence comparison preferential treatment.

b) Eukaryote Cytochrome c and Cytochrome f of Algae and Bacteria

Cytochromes are a group of proteins widespread in living organisms, being involved in both respiration and photosynthesis in eukaryotes and prokaryotes.

The mitochondrial cytochromes c in eukaryotes occupy a position near the end of the mitochondrial respiratory electron transport chain where they are oxidized by an a-type cytochrome which is in turn oxidized by molecular oxygen. They form

a highly homologous group of the same molecular structure and are almost certainly evolutionary descendants of a common ancestor (Dickerson, 1972).

A small c-type cytochrome, cytochrome f, has been isolated from the chloroplasts of algal eukaryotes (Pettigrew, 1974) and the blue-green algal prokaryote (Ambler and Bartsch, 1975). This molecule is the terminal member of the photosynthetic electron transfer chain leading to chlorophyll photocentre I. Homology seems to be feasible since the regions of greatest similarity, viz. the N-terminal regions including the haem attachment site (residues 1 to 15 of cytochrome f and 5 to 19 of cytochrome c), and the latter part of the C-terminal regions (showing similarity from residues 59 of cytochrome f and 82 of cytochrome c) are brought together in the tertiary structure of cytochrome c to form the 'right channel' (Dickerson, 1972), and the conserved methionine of cytochrome f at residue 56 is within a single residue of the methionine at residue 80 of human cytochrome c, the sixth haem ligand.

Using the 'maximum match' method of Needleman and Wunsch, as applied by Barker and Dayhoff with their scoring system (system 1) no homology was apparent between human cytochrome c and Euglena cytochrome f, since the alignment score was only 1.2 standard deviations from the random mean ($P > 0.1$).

When examined by the 'optimum match' method, scoring system 1 yielded a percentage level of 2.5% at the optimum alignment, a level which six of the ninety random comparisons surpassed. With scoring systems 2 or 3 the first two gaps increased the non-significant simple frameshift scores by quite significant increments (system 2 $P = 0.02, 0.01$; system 3 $P < 0.01, 0.04$) and the percentage level at the optimum alignment of the real sequences (0.25%) surpassed those of all alignments of random sequences.

The method of Needleman and Wunsch thus failed to show homology (type II error) by using the maximum possible score at saturation with gaps, and the scoring system of Barker and Dayhoff (system I) is sub-optimal for this comparison since the suggested alignment based on structural considerations results in the last six residues of the shorter protein, cytochrome f, not being aligned.

Table 2 illustrates the need for a non -preferential estimate of the probability of a relationship, showing the extreme percentage levels reached by several comparisons of randomly scrambled sequences at their individual optimum alignments. Many random comparisons appear significant when examined with their optimum number of gaps and compared with the scores of other random comparisons at that number of gaps.

Table 2. The upper percentage levels at optimum alignment for ninety comparisons of randomly scrambled sequences of human cytochrome c and Euglena cytochrome f

Optimum percentage level	100% – 5%	5% – 0.5%	0.5% – 0.0
Scoring System			
1	81	6	3
2	74	13	3
3	80	9	1

c) Cytochrome c and Cytochrome c_{551}

These molecules occupy analogous positions in the respiratory chains of the mitochondria and bacteria. Dickerson (1971) suggested these molecules were homologous from both structural considerations and their more obvious similarities with cytochrome c_2 of *Rhodospirillum rubrum*. Using either scoring system, the percentage levels at the optimum alignments between human cytochrome c and cytochrome c_{551} of *Pseudomonas fluorescens* or *Pseudomonas aeruginosa* were below 0.25%, surpassing all levels from alignments of random sequences.

If only score increments for each gap were considered as suggested by Sankoff and Cedergren, homology would seem to be only a remote possibility for all the scoring systems.

d) Plant and Bacterial Ferredoxin

Using the method of Needleman and Wunsch, Barker and Dayhoff were unable to detect homology between plant and bacterial ferredoxins ($P > 0.1$) without first judiciously selecting part of the plant sequence. However, using either scoring system, the sequence alignments of representative members of each group, viz. *alfalfa* and *Clostridium pasteurianum* ferredoxins reached the 1% level at their optimum alignment and this critical percentage level was surpassed by 5, 5 and 3 alignments from random sequences using scoring systems 1, 2 or 3 respectively, so suggesting a distant homology, without having necessitated any manipulation of sequences.

e) Ferredoxins and Bovine Adrenodoxin

The addition of a single large gap in the bacterial ferredoxin *(Clostridium pasteurianum)* brought about a significant increase in the alignment score between this and bovine adrenodoxin (system 1, $P = 0.05$; system 2, $P = 0.01$; system 3, $P < 0.01$) and the alignment scores reached the 2.5% level using system 1 or 2 and the 1% level using system 3. However, these percentage levels were surpassed by 11, and 8 and 4 alignments respectively from 90 random comparisons, making any relationship appear either dubious or very remote.

Similarly, evidence of homology between the plant ferredoxin of *alfalfa* and bovine adrenodoxin was poor. Barker and Dayhoff had suggested these sequences were homologous using the 'maximum match' method and system 1, and the 'optimum match' method using system 1 demonstrates that a maximum scoring alignment is reached at saturation with 14 gaps, and this point coincides with the optimum alignment reaching a level of 1%. Three of ninety random comparisons were better than the critical 1% level at their optimum alignments when using scoring system 1.

The optimum alignment using system 2 reached a critical percentage level of 5% which was bettered by 15 alignments from random sequences, whilst no suggestion of homology was found using scoring system 3.

When all these considerations are taken into account, homology must appear more dubious than if simply assessed by the approach of Barker and Dayhoff, their possibly spurious result being due to the chance happening of saturation with gaps corresponding with the optimum alignment of the sequences.

f) Bovine Cytochrome b_5 and Horse Haemoglobin β Chain

Homology between these haem binding molecules was suggested by Ozols and Strittmatter (1967) in view of the numerous identities found when many gaps were

introduced. The alignment appeared feasible since the tertiary structure of horse haemoglobin was known, and the major deletions they proposed engulfed the α-helices. However, no evidence of homology between these molecules is evident using the 'optimum match' method' with either scoring system even when allowing preferential selection of only part of the cytochrome b_5 chain so that the proposed homology of Ozols and Strittmatter was not of the crossed type. The absence of any extensive structural homology was confirmed by the structural determination of cytochrome b_5 by Mathews, Levine and Argos (1972). More recently Rossmann and Argos (1975) have indicated a similarity of fold in the haem pockets. Provided neither gap nor frameshift is allowed in the randomly scrambled comparisons of those 48 residues considered structural equivalents, the proposed alignment is just significant at the 5% level. On this basis it is simpler to accept a hypothesis of distant evolutionary relationship rather than convergence of the structures.

Conclusion

Any amino acid sequence or pair of sequences can be examined from a variety of aspects and will probably possess some insignificant yet improbable feature. In assessing homology it is therefore essential to calculate not the *a posteriori* probability of the random occurrence of a particular event, but the *a priori* probability of a general event of that type. Thus care must be used in the interpretation of the increment data for gaps as criteria of homology (Sankoff and Cedergren, 1973) since a high increment for the addition of say a fourth gap in an alignment is a very particular event.

Any preferential treatment of the comparison of real sequences must be avoided when assessing the significance of a possible ancestral relationship, and this is achieved in the 'optimum match' procedure described by permitting all comparisons both real and random, the optimum number of gaps as assessed from the smallest upper percentage point in the frequency distributions of scores at each number of gaps. The significance level of the relationship is determined from the proportion of randomly scrambled sequence comparisons which reach a smaller upper percentage point at their optimum alignments than that of the real comparison at its optimum alignment. This use of the frequency distribution at each number of gaps to assess each score overcomes the higher scores associated with larger numbers of gaps, and so consequently the differing number of gaps introduced into the comparison of the real sequences and the various randomly scrambled sequence comparisons is, in effect, eliminated. This permits the determination of a significance level for the relationship of the real sequences using optimum alignments in all comparisons.

The 'maximum match' method fulfils the requirement of non-preferential treatment of the comparison of real sequences, but the use of maximum match scores is sub-optimal and so subject to greater errors of judgement than assessment at the optimum alignment - a fact recognized by Needleman and Wunsch who suggested a penalty for each gap they introduced in order to increase the significance of the maximum match - this assessment is then, however, preferential, and an alignment can only truly be assessed when random comparisons have been given an equal opportunity to reach an optimum alignment.

The use of the optimum alignments of both real and random sequences, and the ranked position of the critical percentage level of the real sequence alignment provides a sensitive criterion for assessing non-preferentially whether any relationship is present in sequences above the normal levels of random expectation. This is especially so when used in conjunction with scoring systems 2 or 3, whilst the large bias of system 1 against gaps in the longer sequence makes this generally unsuitable for use with the 'optimum match' method. The weakness in the method is that the frequency distributions of scores must be estimated from a finite number of random comparisons and the optimum alignment assessed from a finite number of random comparisons.

An alignment of homologous sequences has a number of similar or identical residues which are mutually, directionally consistent as a result of the sequences being derived from a single common ancestral sequence. The relationship may have deteriorated in places, especially when functional changes in the molecules have occurred, to levels of random expectation. The alignment score at the optimum alignment thus represents a contribution from both remnants of ancestral homology and optimised random considerations, and so generally, when the deleterious effect of gaps has been eliminated, the sequences should show greater similarity than most random sequences of the same compositions, so permitting the detection of ancestral relationships. Because of the random component, a claim that this optimum alignment represents the way in which the sequences are related is not justified. When homology has been established, alignments are best constructed from the maximum number of statistically significant gaps (Sankoff and Cedergren, 1973) together with any additional knowledge regarding functional residues or alignment with other sequences bearing a greater homology with both those compared.

Acknowledgments: I am grateful to Mr. John van der Touw for his stimulating discussion and critical reviewing of this manuscript.

References

Ambler, R.P., Bartsch, R.G. (1975). Nature **253**, 285-288

Barker, W.C., Dayhoff, M.O. (1972). In: Atlas of Protein Sequence and Structure, (Dayhoff, M.O. ed.), vol. 5, pp. 101-110. National Biomedical Research Foundation, Washington, USA

Haën, C. de, Swanson, E., Teller, D.C. (1976). J. Mol. Biol. **106**, 639-661

Dickerson, R.E. (1971). J. Mol. Biol. **57**, 1-15

Dickerson, R.E. (1972). Scientific American, vol. 226, No. 4, pp. 58-72

Fitch, W.M. (1966). J. Mol. Biol. **16**, 9-16

Fitch, W.M. (1970). J. Mol. Biol. **49**, 1-14

Haber, J.E., Koshland, D.E. (1970). J. Mol. Biol. **50**, 617-639

Johnson, N.L., Nixon, E., Amos, P.E. (1963). Biometrika **50**, 459-498

Mathews, F.S., Levine, M., Argos, P. (1972). J. Mol. Biol. **64**, 449-464

McLachlan, A.D. (1971). J. Mol. Biol. **61**, 409-424

Needleman, S.B., Wunsch, C.D. (1970). J. Mol. Biol. **48**, 443-453

Ozols, J., Strittmatter, P. (1967). Proc. Nat. Acad. Sci. Wash. **58**, 264-267

Pettigrew, G.W. (1974). Biochem. J. **139**, 449-459
Rossmann, M.G., Argos, P. (1975). J. Biol. Chem. **250**, 7525-7532
Sankoff, D., Cedergren, R.J. (1973). J. Mol. Biol. **77**, 159-164

Received October 10, 1977; Revised February 15, 1978

Appendix 1

The information values of amino acid replacements as a scoring system and their use in detecting sequence homologies.

Let p(E) be the probability of an event E, which occurs in conjunction with a second event H or $\bar{H}$ of which the latter are mutually exclusive and exhaustive; p(H) is the probability of the event H occurring; p(E/H) the probability of E occurring given H has occurred; p(H/E) the probability of H given E has occurred; and p(EH) the probability of the joint occurrence of both events, then:

$$p(H) \times p(E/H) = p(EH)$$

$$= p(E) \times p(H/E)$$

$$\frac{p(E/H)}{p(E)} = \frac{p(H/E)}{p(H)}$$ (Bayes's rule for the probability of causes if H and $\bar{H}$ are causes)

likewise

$$\frac{p(E/\bar{H})}{p(E)} = \frac{p(\bar{H}/E)}{p(\bar{H})}$$

where $p(\bar{H})$ i.e. (1-p(H)) is the probability of alternative hypotheses being correct if H is a hypothesis.

Hence

$$\frac{p(E/H)}{p(E/\bar{H})} = \frac{pH/E}{p\bar{H}/E} \div \frac{p(H)}{p(\bar{H})}$$

$$= \frac{O(H/E)}{O(H)}$$ where O is the odds on the hypothesis H against $\bar{H}$.

Hence $$\log_{10} \frac{p(E/H)}{p(E/\bar{H})} = \log_{10} O(H/E) - \log_{10} O(H)$$

N.B. If E is a series of independent events:

$$\frac{pE/H}{pE/\bar{H}} = \frac{pE_1.E_2 \ldots . E_n/H}{pE_1\, E_2 \ldots . E_n/\bar{H}} = \frac{p(E_1/H)}{p(E_1/\bar{H})} \times \frac{p(E_2/H)}{p(E_2/\bar{H})} \times \ldots\ldots \frac{p(E_n/H)}{p(E_n/\bar{H})}$$

The quantity $\log_{10} [p(E_i/H) / p(E_i/\bar{H})]$ is thus additive, each observed event adding to the weight of evidence in favor of H as opposed to $\bar{H}$, and this quantity is the information in E_i for discrimination between hypotheses H and $\bar{H}$, viz: $I\ (H:\bar{H};E_i)$.

In the context of the manuscript H is the hypothesis of relationships between two proteins, and each event E_i is an alignment of two amino acid residues. The odds of a relationship O(H) for two real sequences and all sequences of the same size and composition as these will be the

same prior to a consideration of the sequence alignments. The sum of $I\ (H{:}\bar{H};E_i)$ for all amino acid residues aligned in the real sequences can be compared with the distribution of this sum for alignments of the random sequences and this indicates how often that much additional information as to relationship from the sequence alone would be obtained, and hence provides a measure of sequence similarity additional to that of size and composition.

Appendix 2

The matrix transformations and search procedure as applied to two pairs of sequences.

The correspondence array of Fig 10-1 of Barker and Dayhoff (1972) is used to construct matrix S. Application of equation 1 to matrix S, followed by the recursive application of 2b, 3 and 4 determines the maximum alignment score possible for a pair of sequences as each gap is added.

Example 1 (see next page)

MVSK and DVSQK

The transformation of $B_1\ (2,2)$ to $A_1\ (2,2)$ illustrates how each new gap is optimally placed for maximum score by equation 3, viz:

$$A_1\ (2,2) = \max\ [B_1\ (2,2),\ B_1\ (3,3) + S\ (2,2)\ ]$$

i.e. The maximum value of $\begin{matrix} VSQK \\ V{-}SK \end{matrix}$ of score 7, or $\begin{matrix} VSQK \\ VS{-}K \end{matrix}$ of score 8.

The transformation of $A_1\ (1,1)$ to $B_2\ (1,1)$ and then $A_2\ (1,1)$ illustrates how the compulsory alignment of the first residue is avoided by the application of equation 2b rather than 2a viz:

$$B_2\ (1,1) = 8 + \max\ (0,-2)$$
$$= 8$$

The resulting maximum scores from the A_0, A_1 and A_2 matrices are thus 5, 6 and 8 for the alignments $\begin{matrix} DVSQK \\ MVSK \end{matrix}$, $\begin{matrix} DVSQK \\ MVS{-}K \end{matrix}$, and $\begin{matrix} DVSQK \\ M{-}VS{-}K \end{matrix}$ which allow 0, 1 and 2 gaps respectively. The sequence alignment of maximum score for n gaps may be determined by the search procedures described in the text using matrices $A_n, A_{n-1}, \ldots\ldots\ldots A_0$, eg. from $A_2\ (1,1)$, not in the pathway itself, is obtained a maximum scoring pathway, viz:

$A_1\ (2,2)$, $A_1\ (3,3)$, $A_0\ (5,4)$, and corresponding alignment

i = 1	2	3	4	5
D	V	S	Q	K
M	— V	S	—	K
j = 1	2	3		4

Example 2 (see next page)

SVMK and VDQKS

Example 1

$$S = \begin{array}{c} \\ D \\ V \\ S \\ Q \\ K \end{array}\!\!\begin{array}{c} \begin{array}{cccc} M & V & S & K \end{array} \\ \begin{bmatrix} -2 & -1 & 1 & 0 \\ 2 & 3 & -1 & -1 \\ -1 & -1 & 1 & 0 \\ -1 & -1 & 0 & 0 \\ -1 & -1 & 0 & 4 \end{bmatrix} \end{array} \rightarrow A_0 = \begin{bmatrix} 2 & -2 & 0 & 0 \\ \boxed{5} & 4 & -1 & -1 \\ -2 & 3 & 1 & 0 \\ -2 & -1 & 4 & 0 \\ -1 & -1 & 0 & 4 \end{bmatrix} \rightarrow B_1 = \begin{bmatrix} 4 & 4 & 5 & 0 \\ 5 & 7 & 3 & -1 \\ 4 & 3 & 5 & 0 \\ 4 & 3 & 4 & 0 \\ -1 & -1 & 0 & 4 \end{bmatrix} \rightarrow A_1 = \begin{bmatrix} \boxed{6} & 4 & 5 & 0 \\ 5 & 8 & 3 & -1 \\ 4 & 3 & 5 & 0 \\ 4 & 3 & 4 & 0 \\ -1 & -1 & 0 & 4 \end{bmatrix} \rightarrow B_2 = \begin{bmatrix} 8 & 5 & 5 & 0 \\ 7 & 8 & 3 & -1 \\ 4 & 3 & 5 & 0 \\ 4 & 3 & 4 & 0 \\ -1 & -1 & 0 & 4 \end{bmatrix} \rightarrow A_2 = \begin{bmatrix} \boxed{8} & 5 & 5 & 0 \\ 7 & 8 & 3 & -1 \\ 4 & 3 & 5 & 0 \\ 4 & 3 & 4 & 0 \\ -1 & -1 & 0 & 4 \end{bmatrix}$$

$$C_1 = \begin{bmatrix} * & * & * & * \\ * & 4 & 4 & 0 \\ * & 4 & 4 & 0 \\ * & 4 & 4 & 0 \\ * & 0 & 0 & 0 \end{bmatrix} \qquad C_2 = \begin{bmatrix} * & * & * & * \\ * & 5 & 4 & 0 \\ * & 4 & 4 & 0 \\ * & 4 & 4 & 0 \\ * & 0 & 0 & 0 \end{bmatrix}$$

Example 2

$$S = \begin{array}{c} \\ V \\ D \\ Q \\ K \\ S \end{array}\!\!\begin{array}{c} \begin{array}{cccc} S & V & M & K \end{array} \\ \begin{bmatrix} -1 & 3 & 2 & -1 \\ 1 & -1 & -2 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & -1 & -1 & 4 \\ 1 & -1 & -1 & 0 \end{bmatrix} \end{array} \rightarrow A_0 = \begin{bmatrix} 1 & 1 & \boxed{2} & -1 \\ -1 & 2 & -2 & 0 \\ -2 & -2 & 3 & 0 \\ -1 & -2 & -1 & 4 \\ 1 & -1 & -1 & 0 \end{bmatrix} \rightarrow B_1 = \begin{bmatrix} 2 & 6 & 6 & -1 \\ 4 & 2 & 2 & 0 \\ 4 & 3 & 3 & 0 \\ 0 & -1 & -1 & 4 \\ 1 & -1 & -1 & 0 \end{bmatrix} \rightarrow A_1 = \begin{bmatrix} 2 & 6 & \boxed{6} & -1 \\ 4 & 2 & 2 & 0 \\ 4 & 3 & 3 & 0 \\ 0 & -1 & -1 & 4 \\ 1 & -1 & -1 & 0 \end{bmatrix} \rightarrow B_2 = \begin{bmatrix} 4 & 7 & 6 & -1 \\ 5 & 3 & 2 & 0 \\ 4 & 3 & 3 & 0 \\ 0 & -1 & -1 & 4 \\ 1 & -1 & -1 & 0 \end{bmatrix} \rightarrow A_2 = \begin{bmatrix} 4 & \boxed{7} & 6 & -1 \\ 5 & 3 & 2 & 0 \\ 4 & 3 & 3 & 0 \\ 0 & -1 & -1 & 4 \\ 1 & -1 & -1 & 0 \end{bmatrix}$$

$$C_1 = \begin{bmatrix} * & * & * & * \\ * & 3 & 4 & 0 \\ * & 4 & 4 & 0 \\ * & 0 & 0 & 0 \\ * & 0 & 0 & 0 \end{bmatrix} \qquad C_2 = \begin{bmatrix} * & * & * & * \\ * & 4 & 4 & 0 \\ * & 4 & 4 & 0 \\ * & 0 & 0 & 0 \\ * & 0 & 0 & 0 \end{bmatrix}$$

Boxed scores in the A_n matrices (n = 0, 1, 2) indicate maximum score for n gaps

This illustrates the use of the C matrices for introducing adjacent gaps.

In the construction of matrix B_2 using equation 2b the larger values of the C_1 matrix are used rather than the values of A_1 when calculating $B_2(1,1)$, $B_2(1,2)$, $B_2(2,1)$, $B_2(2,2)$

eg. $B_2(1,2) = \max\ (0, A_1(2,3), A_1(2,4)\ A_1(3,3), A_1(4,3), A_1(5,3),$
$C_1(2,3), C_1(2,4), C_1(3,3), C_1(4,3), C_1(5,3)) + \max(0, S91,2))$
$= \max\ (0,\ 2,\ 0,\ 3\ -1, -1,\ 4,\ 0,\ 4,\ 0,\ 0) + \max(0,3)$
$= 7$

The choice of $C_1(2,3)$ permits the alignment $\begin{matrix} V - DQKS \\ VM--K \end{matrix}$, the construction relying on

a) the M/D alignment is scored as zero (i.e. as in the C_1 matrix).

b) A gap following M/D-which is implicit in the construction of the C_1 matrix.

c) A gap to the cell under consideration, which is being permitted in the construction of the new B_2 matrix.

From $A_2(1,2)$ a maximum scoring pathway may be found viz: $A_2(1,2)$, $A_0(4,4)$, and then the corresponding alignment.

```
j =     1     2  3  4  5
        V  -  D  Q  K  S
     S  V  M  -  -  K
j = 1  2  3        4
```

References

Barker, W.C., Dayhoff, M.O. (1972). In *Atlas of Protein Sequence and Structure,* (Dayhoff, M.O. ed.), vol. 5, pp. 101-110. National Biomedical Research Foundation, Washington, USA