J. Mol. Biol. (1971) **61**, 409–424

Tests for Comparing Related Amino-acid Sequences. Cytochrome *c* and Cytochrome c_{551}

A. D. McLachlan

Medical Research Council Laboratory of Molecular Biology
University Postgraduate Medical School
Hills Road, Cambridge CB2 2QH, England

(*Received 26 August 1970, and in revised form 9 March 1971*)

An improved method for testing similarities or repeats in protein sequences is described. It includes three features: a measure of similarity for amino acids, based on observed substitutions in homologous proteins; a search procedure which compares all pairs of segments of two proteins; new statistical tests which estimate the probabilities that observed correlations could have occurred by chance. Calculations show that gene duplication has probably not occurred in plant ferredoxins; phage Qβ and f2 coat proteins may be homologous; and repeats in cytochrome *c* are not statistically significant. The method predicted an alignment of cytochrome *c* and c_{551} sequences which later appeared consistent with Dickerson's atomic model of horse cytochrome *c*.

1. Introduction

This paper describes improved methods for comparing the amino-acid sequences of proteins to see whether they are likely to be related. The methods are based on those developed by Fitch (1966), Cantor & Jukes (1966), Needleman & Blair (1969) and Haber & Koshland (1970), but have important new features. The first is to use a measure of similarity between every pair of amino acids, which is based on observed substitutions in homologous groups of proteins, rather than on the genetic code or intuitive notions of chemical similarity. The second is a search procedure which presents all possible comparisons between portions of two proteins on a matrix in graphical form, giving an immediate estimate of the statistical significance of any correlation. A single chart shows all the comparisons between proteins in a family; say ten cytochromes *c* with three cytochromes c_{551}. The third is a pair of new probability distributions which are useful for estimating significance. These distributions agree with experimental ones derived by comparing proteins either with unrelated ones or with random sequences, whereas a Gaussian distribution does not.

The methods suggest that the gene duplication in bacterial ferredoxins is not present in the vegetable ones; that the coat proteins of bacteriophages Qβ and f2 could be related, and their RNA sequences are significantly similar; that repeats in the sequence of cytochrome *c* are not significant.

There is a statistically significant similarity between the sequences of cytochrome *c* and cytochrome c_{551}, and the best alignment predicted theoretically, before the

structure of cytochrome c was known, appears to agree well with the observed structure.

2. Genetic and Structural Similarity

The first test, which was used to decide whether two apparently similar proteins have evolved from a common ancestor, was founded on the genetic code. Fitch (1966) and Cantor & Jukes (1966) calculated the minimum mutation distance; i.e. the minimum number of base changes required to convert one sequence into the other. If the proteins are closely related this method is particularly useful and can be used for constructing evolutionary family trees (Doolittle & Blomback 1964; Fitch & Margoliash, 1967). But if the relationship is more distant the test is not very discriminating, as single base changes are so common.

The fundamental assumption of the present approach is that if the amino-acid sequences of two proteins are so alike that their similarity is very unlikely to have happened by chance, then they will have the same three-dimensional structure and be ancestrally related. This is based on the finding from X-ray studies that homologous proteins have very similar three-dimensional structures, so that observed amino-acid substitutions usually conserve the folding of the polypeptide chain (Perutz, Kendrew & Watson, 1965). Thus, related proteins remain structurally similar even if the mutation distances are large. In those proteins which are known to be homologous and share a common structure the following features are often found: (a) many identical amino acids and single base changes; (b) side chains of similar size, shape and charge; (c) a similar pattern of internal hydrogen bonds; (d) a common pattern of non-polar side chains at internal sites; (e) similar structure and functional groups at the active site. The more distant members of the family lose common features in the order (a) to (e), beginning with individual amino acids and ending with the functional groups. Thus it seems logical to assess the relationships between proteins by looking for sequences of chemically similar amino acids, as has been done by Sneath (1966), Needleman & Blair (1969) and Haber & Koshland (1970).

One could object to the fundamental assumption, on the grounds that convergent evolution is likely to lead to precisely these kinds of accidental similarities between unrelated proteins. There is not sufficient evidence yet to exclude this possibility. However, no example is yet known where convergent evolution has led to similarities of structure or sequence which approach those found repeatedly in homologous proteins. Rather, the existence of unrelated lysozymes or nucleases, the irregular and apparently random structural features of many proteins, and the large variety of amino-acid substitutions in homologous families of proteins, all suggest that the number of conceivable ways of evolving an enzyme to perform a given function is astronomically large. Thus, convergent evolution is unlikely to repeat more than a few of the many fine details of structure and sequence in any pair of proteins.

Our object then, is to test whether two protein sequences possess the kind of similarities which are known to occur in homologous proteins, bearing in mind the reservation that one may detect an example of convergent evolution. This requires a quantitative measure of the similarity of any two amino acids, based on experiment. Sneath's (1966) methods are valuable but do not settle the weights to be attached to different factors, such as shape, charge and so on. A more logical approach is to use the frequencies of observed amino-acid substitutions in homologous proteins as a guide, setting up a numerical score $m(i,j)$ for each pair of amino acids i,j.

3. Substitution Frequencies

Figure 1 shows the observed pattern of amino-acid substitutions at corresponding positions in families of ancestrally related proteins. Dayhoff (1969) has already counted the 'accepted point mutations' at branch points of the phylogenetic trees of several families, and constructed a frequency table which applies to closely related proteins. Here we are most interested in examples where the genetic relationship may be distant, but the spatial structure is conserved, so we have counted substitutions the following way.

(1) Count the number $N(i,j)$, with $i \neq j$, of *positions* in each family of sequences at which amino acids i and j occur as alternatives. For example, let the amino acid at position r in six homologous proteins be A, L, V, A, V, V. This counts as *one* occurrence for each of the pairs AL, LV, VA. Count also the number $n(i)$ of positions at which acid i appears, and the sum $N(i)$ of all the $N(i,j)$ which involve acid i. Also count the total number N_1 of substitutions and N_2 the sum of products $n(i)n(j)$, both summed over all pairs $i \neq j$. N_1 is the sum of all the $N(i,j)$.

(2) If the substitutions are random, the expected value of $N(i,j)$ is $E(i,j) = \alpha n(i)n(j)$, where $\alpha = N_1/N_2$. The expected value $E(i)$ for $N(i)$ is then the sum of $E(i,j)$ with $i \neq j$.

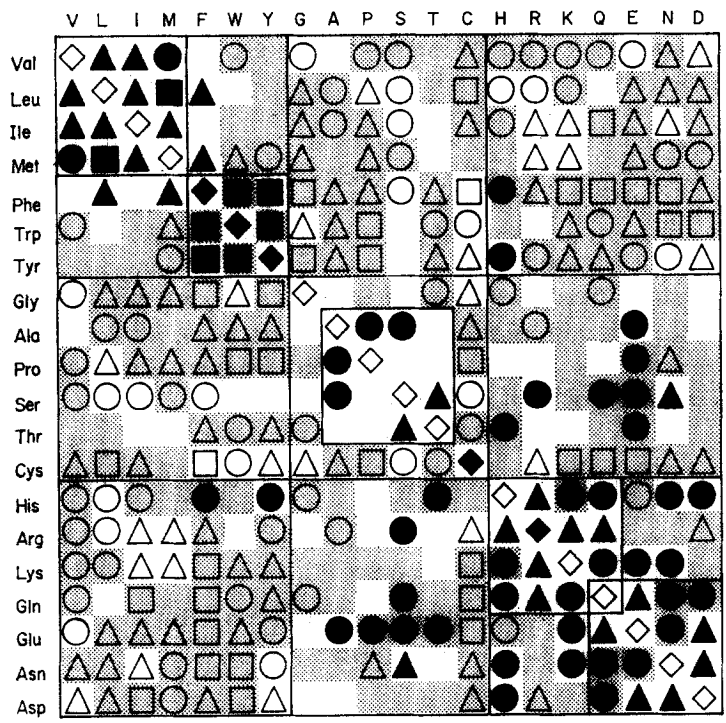


FIG. 1. Relative substitution frequencies $f(i,j)$ in homologous proteins.
●, ■, ▲, Values which are higher than average: ■, over 3·0; ▲, 3·0 to 1·75; ●, 1·74 to 1·30; blank squares, 1·29 to 0·80. ○, □, △, Values below average: ○, 0·79 to 0·60; △, 0·59 to 0·33; □, below 0·33. Shaded squares show mutations which are forbidden because they require more than one base change.

(3) The relative pair substitution frequency is defined as $f(i,j) = N(i,j)/E(i,j)$ and the relative variability of each amino acid as $f(i) = N(i)/E(i)$.

The results shown in Figure 1 are taken from a sample of 16 families† and include $N_1 = 9280$ substitutions. The value of $N(i)$ varies from 68 for Cys to 1019 for Ser, and 164 $N(i,j)$ values out of 380 are greater than or equal to 20. The substitution Ser-Thr is the most common, and occurs at 141 positions. The $f(i)$ values are all between 0·84 and 1·22, with the exception of Cys ($f = 0{\cdot}43$) which is seldom replaced. Of the other amino acids Glu, Ser, Thr and Gln have $f > 1{\cdot}1$ and Gly, Trp, Tyr, Phe, Ile, Pro have $f < 0{\cdot}9$.

The frequencies of substitutions $f(i,j)$ form a pattern like Dayhoff's except for a few surprising changes, notably that Ala-Gly and Cys-Met are not particularly common. The substitutions with the highest frequencies are Phe-Tyr ($f = 5{\cdot}41$), and Trp-Tyr ($f = 5{\cdot}11$), which occur at 48 and 16 positions, respectively. The pattern deviates strongly from the pattern of single base changes allowed by the genetic code.

A score matrix $m(i,j)$ to measure the similarity of pairs of amino acids was set up as follows: scores of 6, 5, and 4 for the most frequent substitutions (black squares, triangles and circles in Fig. 1); 3 for a neutral substitution; 2, 1, 0 for the less frequent substitutions (open circles, triangles and squares); 8 for an identity. A score of 9 was given for identities involving Phe, Tyr, Trp, Cys. Initially Arg was also given a score of 9, but in later work this was reduced to 8. Thus a score of 3 is average, and scores between 4 and 9 show varying degrees of similarity. The score for a gap was 0.

4. The Comparison Matrix

Suppose that we wish to search for similar sequences in two proteins A, B of lengths n_A, n_B. Let the amino acids at position p of A and position q of B be a_p, b_q, and let $M(p,q) = m(a_p,b_q)$ be their similarity scores. If two stretches of sequence in this region are similar, one will see a line of high scores on the *score matrix* $M(p,q)$ running parallel to a main diagonal. If there is a deletion or insertion in one sequence the high scores will continue on a neighbouring diagonal after a break. Gibbs & McIntyre's diagram (1970) uses this idea, taking $M(p,q) = 1$ if $a_p = b_q$ and 0 otherwise. We first tried printing out the similarity scores in symbolic form, and the names of identical amino acids, on the matrix, but the diagram becomes too confusing. The next stage, therefore, was to calculate a weighted sum of scores for spans of s amino acids along a diagonal, calculating a *comparison matrix*

$$C(p,q) = \sum W_h\, M(p + h, q + h), \qquad h = -g \ldots + g. \tag{1}$$

Here W_h are weights, which can be chosen at will. It is convenient to take s an odd number, $s = 2g + 1$. As an illustration take the sequences around the distal histidine of human haemoglobin α chain and human haemoglobin β chain with span 5 and weights 1,2,3,2,1. The similarity scores give:

Sequence A	Asp	Leu	His	Ala	His
Sequence B	Glu	Leu	His	Cys	Asp
Score	5	8	8	1	4
Weight	1	2	3	2	1
Total for span	5 +	16 +	24 +	2 +	4 = 51

† 2 subtilisin, 13 haemoglobin, 8 cytochrome *c*, 2 penicillinase, 14 antibody light-chain variable regions, 12 antibody constant regions, 5 tobacco mosaic virus coat, 6 azurin, 3 glyceraldehyde-3-phosphate dehydrogenase, 6 chymotrypsin enzymes, 2 cytochrome *c*3, 3 cytochrome c_{551}, 4 lysozyme, 3 ribonuclease, 3 bacterial ferredoxin, and 3 plant ferredoxin chains.

The mean expected score for two randomly chosen pentapeptides chosen from these sequences can be calculated as 25·5 with a standard deviation of 8·5, as we shall see later; also the probability of getting a score over 50 by chance from random segments of these two sequences is 0·0068, so that these pieces are closely similar. Rather than print out the numerical scores on a diagram it is clearer to use a set of contour symbols which show whether the score lies above a series of threshold values. The thresholds are chosen to give a 'probability map' of the comparisons. For example, with the span and weight chosen above for haemoglobin we might choose to set four threshold levels I, II, III, IV such that the probabilities of exceeding them by chance were respectively 2×10^{-4}, 10^{-3}, 10^{-2}, 5×10^{-2}. The elements of the matrix which exceed the corresponding thresholds, calculated as 62, 52, 46, 42, are printed with symbols (X), (0), (+) and (,) leaving all other elements blank. The comparison diagram now appears in a convenient form where one can detect gaps and alternative alignments by eye.

The procedure is very flexible since the scores $m(i,j)$, the span, weights and thresholds are all adjustable. If the score is based on mutation distances and the weights set to unity, one recovers Fitch's (1966) and Cantor & Jukes's (1966) methods. A set of weights which dies away smoothly at each end of a span is helpful, as it is easier to see where gaps begin and end (see Fig. 2). However, the comparison matrix cannot deal with gaps so systematically as Needleman & Wunsch's ingenious computer search (1970). In practice it is useful to do two matrices: one with a span of 11 and weights 1,2,2,3,3,3,3,3,2,2,1 to pick out weak persistent similarity, and another with span 5 to detect local details.

A FORTRAN program written for an IBM 360 44 computer calculates these comparison matrices in a few minutes. It can compare two different sequences or compare a protein with itself to search for internal repeats.

Two extensions of the comparison matrix are useful. Suppose that one believes that two families of proteins A and B are distantly related, and one knows the sequences of

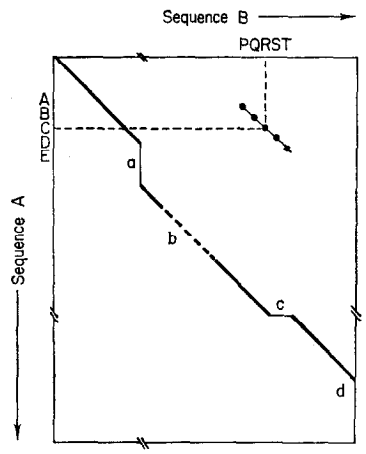


FIG. 2. Sketch of a comparison matrix for two sequences A and B.
The two spans ABCDE and PQRST of length 5 are compared to give a single entry in the matrix. The diagonal line shows a correlation between the sequences which requires insertions in A at a and in B at c. The dotted line b represents a weak correlation. The terminal region of A has extra amino acids added.

j_A different species of A and j_B species of B. It would be inconvenient to form $j_A j_B$ different comparison matrices C_{xy} for each A species x and B species y and look at them all. We define two types of *family comparison matrix:*

(1) the maximal family score C_{Max} (p,q) which is the largest of the C_{xy} (p,q) for all pairs xy; or (2) the averaged family score C_{Av} (p,q) is the average over all pairs. Each of the matrices shows all the interesting similarities in the family on one diagram. It would take too much space to illustrate one here, but diagrams have been printed for many proteins and show clearly all those similarities which have been previously pointed out.

A similar matrix has been used to search for base pairs in RNA sequences, scoring for G·C, A·U and G·U pairs and summing over antiparallel segments. The averaged family comparison matrix displays the common clover-leaf pattern in transfer RNA sequences, for example.

5. Matching Probabilities

What is the expected probability distribution for the scores in the comparison matrix for two random protein sequences A and B, with given spans, weights and score matrix m? An exact calculation is difficult because the numbers in the $n_A \times n_B$ array of scores are not statistically independent, but one can calculate a related distribution exactly.

Consider the following *double matching experiment.* Two infinite packs of cards A and B are shuffled. The cards represent amino acids, and the composition of each pack is in the same proportions as the proteins A, B. A set of s cards $a_1,a_2 \ldots a_s$ are drawn in succession from A, and compared in turn with another set $b_1,b_2 \ldots b_s$ drawn from pack B. For each pair the score $m_r = m(a_r,b_r)$ is recorded and the weighted score for the set of s matches is defined as

$$M = \sum W_r m_r \quad . \tag{2}$$

The *double matching probability* $Q_{AB}(M)$ is defined as the probability that the score in this experiment is greater than or equal to M, and is obtained by summing the probabilities $P_{AB}(M')$ that the score is *exactly* M', with $M' \geq M$. Let $n_A(i)$, $n_B(i)$ be the number of times amino acid i occurs in each sequence, while $f_A(i) = n_A(i)/n_A$ and $f_B(i) = n_B(i)/n_B$ are the fractional compositions. Then the coefficient of x^m in the expression

$$G(x) = \sum_{i,j} f_A(i) f_B(j) x^{m(i,j)} \tag{3}$$

gives the probability that $m(a_r,b_r)$ has the value m, for each r, and the coefficient of x^m in $G(x^{W_r})$ gives the probability that $W_r m_r$ has the value m. Since the m_r are statistically independent, the probability $P_{AB}(M)$ is easily seen to be the coefficient of x^M in the expression

$$F_{AB}(x) = G(x^{W_1})\ G(x^{W_2}) \ldots G(x^{W_s}). \tag{4}$$

These results are trivial extensions of the binomial probability distribution, and a computer can calculate the complete distribution in a few seconds. The probability distribution depends only on the composition of the two proteins.

The mean score M_1 and mean square score M_2 are found from the relations:

$$S_1 = \sum_{i,j} f_A(i) f_B(j) m(i,j), \qquad S_2 = \sum_{i,j} f_A(i) f_B(j) [m(i,j)]^2, \tag{5}$$

$$W_1' = \sum_h W_h, \qquad W_2' = \sum_h W_h^2, \tag{6}$$

$$M_1 = S_1 W_1', \qquad M_2 = S_2 W_2'. \tag{7}$$

These formulae apply to all elements of the comparison matrix except for those near the edges if two different proteins are being compared. If a protein is compared with itself more complicated expressions apply to the diagonal elements $C(p,p)$ and to certain special elements near the main diagonal. Another important quantity is the statistical correlation between consecutive spans along the same diagonal. If $h < s$, the span length, one finds that

$$\overline{C(p+h, q+h)\, C(p,q)} = M_1^2 + (S_2 - S_1^2) w, \tag{8}$$

where

$$w = \sum_r W_r W_{r+h}. \tag{9}$$

This leads to a persistence effect: if the span is long a high score generated at one place in the comparison matrix tends to raise the scores at about s adjacent positions along the same diagonal.

Another probability, which I call the *single matching probability* is also useful. Suppose that on comparing two proteins a certain peptide $a_1 a_2 \ldots a_s$ of A matches closely with a peptide $b_1 b_2 \ldots b_s$ of B, the score being M. If the amino acids $a_1 a_2 \ldots a_s$ are all very uncommon in A the match is much less likely to arise by chance than if a set of common amino acids gave the same score. Therefore, we consider the following imaginary experiment. A *given* set α of s cards $a_1 a_2 \ldots a_s$ is placed on a table in order. Then cards $b_1 b_2 \ldots b_s$ are drawn in turn from an infinite shuffled pack with the composition of protein B, and each b_r matched with a_r. Let $P_{\alpha B}(M)$ be the probability that the total score is M—the weights, for simplicity, all being unity. We define polynomials $g_i(x)$ for each amino acid:

$$g_i(x) = \sum_j f_B(j) x^{m(i,j)}, \tag{10}$$

then $P_{\alpha B}(M)$ is the coefficient of x^M in the product

$$F_{\alpha B}(x) = g_{a1}(x)\, g_{a2}(x) \ldots g_{as}(x). \tag{11}$$

The single matching probability $R_{\alpha B}(M)$ is defined as the probability that the score for one pair should be greater than or equal to M. It is the sum of the $P_{\alpha B}(M')$ for $M' \geq M$. It depends on the composition of the peptide α and the entire protein B.

These probabilities are also easy to calculate, and they confirm that not all of the apparently striking regularities in protein sequences (Sorm & Keil, 1962; Sorm & Knichal, 1958; Urbain, 1969) are significant (Williams, Clegg & Mutch, 1961).

6. Comparison with Experiment

Two questions arise at once. Do the theoretical probability distributions agree with experiment (Williams *et al.*, 1961), and do they resemble a Gaussian distribution with the same mean and standard deviation?

Comparison matrices were calculated for two haemoglobin sequences, human α and horse β, aligned in the usual way, with 148 amino acids including gaps. The span was 11, weights 1,2,2,3,3,3,3,3,2,2,1, and the score matrix as described above. The expected mean and standard deviations for the scores were 70·70 and 15·52. The threshold levels I, II, III, IV for printing out the matrix were set at 150, 139, 127, 112, and the calculated probabilities for exceeding these values by chance, estimated from $Q_{AB}(M)$, were 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, respectively. A histogram was made of all the $(148)^2$ scores†, and the counts converted into 'experimental' probabilities to compare with the calculated $Q_{AB}(M)$ distribution and a theoretical Gaussian $G_{AB}(M)$.

First the α chain was compared with a randomly shuffled β chain. The results, in Figure 3, show that:

(1) the experimental and theoretical cumulative distributions match almost perfectly right out to events which occur only 1 in 10^{-4} times;

(2) a Gaussian distribution fits adequately to within 2 to 3 standard deviations of the mean;

(3) the distribution of scores greater than 3 standard deviations from the mean deviates strongly from the Gaussian. The Gaussian underestimates the probabilities;

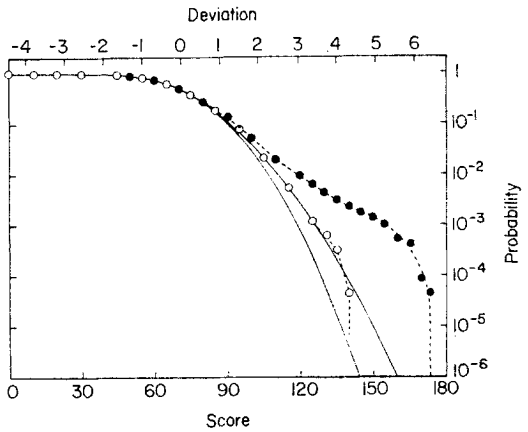


FIG. 3. Cumulative probability distributions of scores in the $(148)^2$ spans in a comparison matrix for haemoglobin α human and haemoglobin β horse.

Upper smooth curve is predicted distribution, equation (4). Lower is a Gaussian with same mean and standard deviation. ○, Observed results for shuffled sequences; ●, comparison of real sequences. Upper scale gives score as a fraction of standard deviation.

(4) the observed mean and standard deviation of 68·67 and 18·10 are close to the expected values.

Next, the real α and β chain sequences were compared. The distribution again fits both the theoretical and the Gaussian ones out to 2 or 3 standard deviations, but now there is a large 'shoulder' at high scores which deviates from theory. This is due to the large number of high scores on the main diagonal.

Table 1 shows that these diagonal elements account for all the deviations from the expected distribution.

† The $(148)^2$ spans include the short spans at the ends of the sequences.

TABLE 1

Frequency of high scores in the comparison matrix for two haemoglobin chains, human α and horse β

	I	II	III	IV
Diagonal elements (real sequences)	31	18	38	34
Off-diagonal elements (real sequences)	0	4	28	214
All elements (shuffled sequences)	0	2	17	166
Expected (random sequences)	0	2	20	197

The Table gives the number of scores which exceed a given level, but not the next higher one. The levels I to IV are set at 150, 139, 127, 122 and the probabilities for exceeding them in a comparison of random sequences are calculated to be 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}.

A further test was made by comparing the same haemoglobin α chain with the unrelated sequence of subtilisin BPN', first using the real haemoglobin sequence and then the shuffled one. Both experimental probability distributions of scores agreed closely with the predicted one. These and other tests show that the double matching probability gives a correct probability distribution in many cases, and is better than the Gaussian approximation.

7. Matching of Whole Sequences

The double and single matching probabilities also provide a test to check whether two entire sequences of the same length are more similar than would be expected by

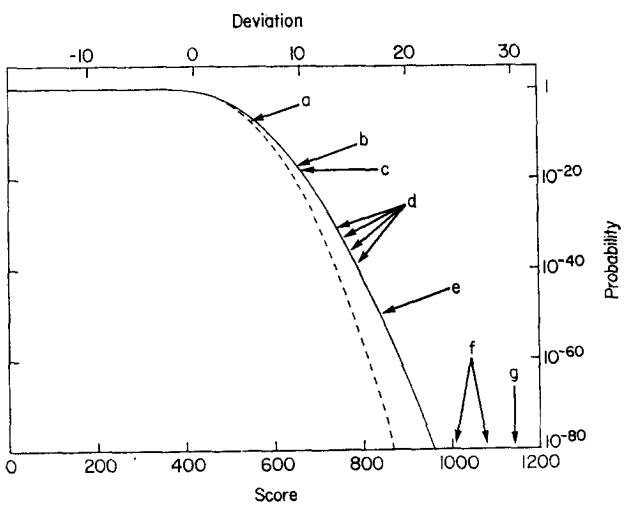


FIG. 4. Similarity scores and matching probabilities for haemoglobins compared with human α chain.

(———) Theoretical double matching probability; (– – –) Gaussian with same mean and standard deviation. Arrows show scores found for chains as follows: (a) *Chironomus thummi*, (b) sperm whale myoglobin, (c) lamprey, (d) mammalian β chains, (e) carp α chain, (f) mammalian α chains, (g) maximum score, human α chain.

chance. Let us compare the corresponding amino acids a_r, b_r one by one and form the *similarity score*

$$L = \sum m(a_r, b_r), \qquad r = 1 \ldots N, \tag{12}$$

with weights of unity.

The probabilities $Q_{AB}(L)$ or $R_{AB}(L)$ are easily calculated. For example, the score of human α and horse β haemoglobin chains aligned in the usual way is 781 out of a maximum 1186 (β against itself); the predicted mean and standard deviations are 418 and 24. The matching probability Q_{AB} is 10^{-52}. However, the score of human α chain against sperm whale myoglobin is only 558 and Q_{AB} rises to 2×10^{-9}. These and other comparisons are plotted on Figure 4. It is important to stress that these probabilities only apply to *this particular alignment* of the sequences. The calculation of matching probabilities when gaps are introduced in various ways is very difficult (Fitch, 1969), and the only estimates available at present are the experimental results of Needleman & Blair (1969) and Haber & Koshland (1970) taken from random sequences.

8. Applications

(a) *Ferredoxins*

The well-known internal doubling (Eck & Dayhoff, 1966) in the sequences of the bacterial ferredoxins from *Micrococcus aerogenes*, *Clostridium pasteurianum* and *C. butyricum* (Tsunoda, Yasunobu & Whitely, 1968; Tanaka, Nakashima, Benson, Mower & Yasunobu, 1966; Benson, Mower & Yasunobu, 1966) shows clearly on the comparison matrices, as a long series of spans with scores above the 10^{-3} probability threshold. The similarity score for matching positions 1 to 26 with 27 to 55 (one gap after position 5, and weights of unity) is 150 out of a possible 213, and the single matching probability for this score is $1{\cdot}5\times10^{-9}$. By using another scoring system, based on the genetic code (2 for an identity, 1 for a single base change, 0 for anything else) one obtains a score of 32/52 with a probability of $1{\cdot}4\times10^{-6}$. The difference between the probabilities of $1{\cdot}5\times10^{-9}$ and $1{\cdot}4\times10^{-6}$ illustrates how the score table based on the observed substitutions in proteins gives a more discriminating test of relatedness than the minimum mutation distance.

To test whether plant ferredoxins are homologous with the bacterial ones and show the same internal doubling a maximal family comparison matrix was printed for the three bacterial *versus* three plant ferredoxins; spinach, alfalfa and scenedesmus (Matsubara & Sasaki, 1968; Keresztes-Nagy, Perini & Margoliash, 1969; Sugeno & Matsubara, 1968), and the plant ferredoxins against themselves (span 11, weights and thresholds as for haemoglobin above). A weak repeat appears if positions (7 to 50) are matched against (51 to 59; 60 to 75; 76 to 88) with two gaps, but the score of 164/360 corresponds to a single matching probability of 6×10^{-4} which is too high to be convincing. The genetic scoring system gives 33/88 with a high probability of 2×10^{-2}.

The comparison between the plant and bacterial chains gives some support to Dayhoff & Eck's (1969) alignment; the single matching probability for fitting P(1 to 42) to S(32 to 75) is $1{\cdot}4\times10^{-8}$. But the fit of P(43 to 55) to S(79 to 91) is poor, with a probability of 0·11. Here P stands for *C. pasteurianum* and S for spinach. These tests suggest that if the two types of ferredoxin are related it is only in the first halves of

their sequences. However Fitch (1970) has obtained some evidence that parts of these sequences may be related by a genetic frameshift.

(b) *Bacteriophage coat proteins*

The 5′ ends of the RNA sequences of bacteriophage R17 and Qβ (Adams & Cory, 1970; Billeter, Dahlberg, Goodman, Hindley & Weissman, 1969) show two striking similarities:

GGGACCCC--UUUGGGGGUC	R17(5 to 22)
GGGACCCCCCUUUAGGGGUC	Qβ(2 to 21)
UAAUGCCAUUUUUAAUGUCUUUAGCGAG	R17(44 to 71)
UAAUGAAAUUCUUAAUGAUUUUCAGGAG	Qβ(101 to 128)

The matching probabilities for these pairs, based on a score of 1 for each identical base and 0 for different ones are 8×10^{-10} and 4×10^{-7}, suggesting that these RNA sequences have a common ancestor. It is, therefore, interesting that Konigsberg, Maita, Katze & Weber (1970) have proposed that the amino-acid sequences of the coat proteins of the phages f2, which differs by one amino acid from R17, and Qβ are homologous. A comparison matrix with span 11, weights 1,2,2,3,3,3,3,3,2,2,1 and thresholds 139, 127, 112, 103, shows a good fit, with scores over 127 for their match of the regions surrounding Qβ positions 29 to 35, 48 to 54 and 62 with the corresponding parts of f2. But many stretches are dissimilar and there is no evidence at all for any similarity between the second halves of the sequence. The matching probability for the first halves (positions 1 to 76) in their alignment is 6×10^{-9}, but for positions 77 to 136 it rises to 2×10^{-3}. A probability of 6×10^{-9} for matching two such long segments which contain 7 gaps is interesting, but probably not low enough to be conclusive evidence for a homology.

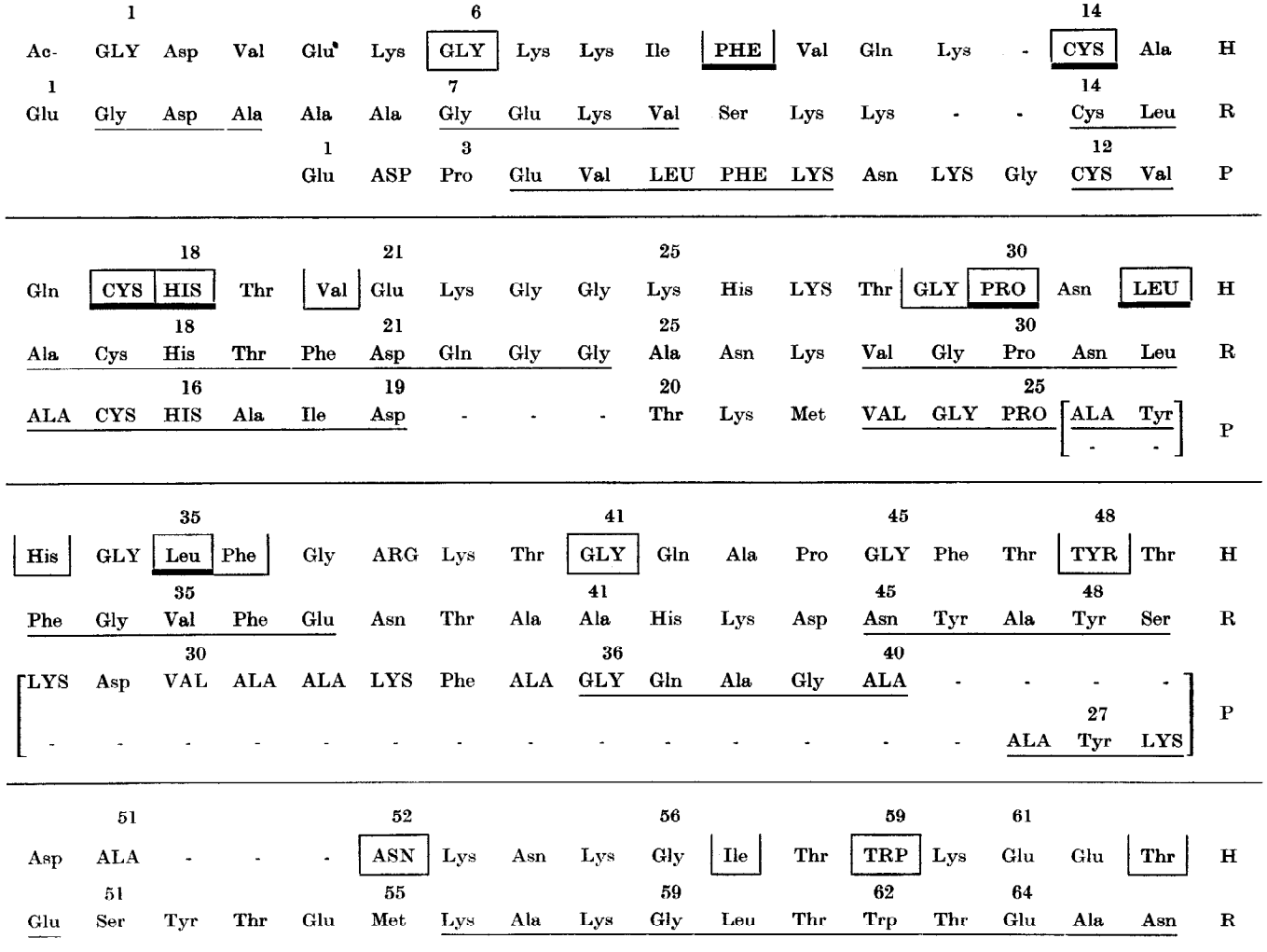
(c) *Internal repeats in cytochromes*

Cantor & Jukes (1966) noticed repeats in the sequences of cytochrome c_{551} and *Neurospora crassa* cytochrome *c*. Later Dus, Sletten & Kamen (1968) noticed further repeats in cytochrome c_2 and proposed that all cytochromes had arisen from an ancestral repeating sequence. Table 2 shows the strongest repeats (*a* to *g*) found in a maximal family comparison matrix $C_{\mathrm{Max}}(p,q)$ which searched for all possible comparisons between the sequences of six cytochromes *c*, three c_{551} and one c_2. For each pair, the single and double matching probabilities were calculated. The pair (*h*) was noted by Cantor & Jukes (1966) and is clearly not significant when judged by these tests. None of the other correlations is judged to be significant either, for the following reasons.

(1) In approximately 10^6 spans one expects to find scores with matching probabilities of the order 10^{-6} and 10^{-7}.

(2) If all repeats are ancestral they are mutually incompatible. Thus (a), (b), (c) are mutually inconsistent.

(3) The three-dimensional structure of cytochrome *c* is as irregular as other proteins and shows no repeating pattern to match the sequence repeats. The only exception is the pair (e) where positions (64 to 72) form a highly distorted α-helix, while (91 to 99) are in a regular helix.


1 6 14
Ac- GLY Asp Val Glu Lys GLY Lys Lys Ile PHE Val Gln Lys - CYS Ala H
1 7 14
Glu Gly Asp Ala Ala Ala Gly Glu Lys Val Ser Lys Lys - - Cys Leu R
1 3 12
Glu ASP Pro Glu Val LEU PHE LYS Asn LYS Gly CYS Val P
18 21 25 30
Gln CYS HIS Thr Val Glu Lys Gly Gly Lys His LYS Thr GLY PRO Asn LEU H
18 21 25 30
Ala Cys His Thr Phe Asp Gln Gly Gly Ala Asn Lys Val Gly Pro Asn Leu R
16 19 20 25
ALA CYS HIS Ala Ile Asp - - - Thr Lys Met VAL GLY PRO ALA Tyr P
- -
35 41 45 48
His GLY Leu Phe Gly ARG Lys Thr GLY Gln Ala Pro GLY Phe Thr TYR Thr H
35 41 45 48
Phe Gly Val Phe Glu Asn Thr Ala Ala His Lys Asp Asn Tyr Ala Tyr Ser R
30 36 40
LYS Asp VAL ALA ALA LYS Phe ALA GLY Gln Ala Gly ALA - - - - P
27
ALA Tyr LYS
51 52 56 59 61
Asp ALA - - - ASN Lys Asn Lys Gly Ile Thr TRP Lys Glu Glu Thr H
51 55 59 62 64
Glu Ser Tyr Thr Glu Met Lys Ala Lys Gly Leu Thr Trp Thr Glu Ala Asn R

														41				
-	-	-	-	-	-	-	-	-	-	-	-	-	-	Glu	Ala	Glu		P
	30			31					36			39						
Asp	VAL	-	-	ALA	ALA	LYS	Phe	ALA	GLY	Gln	Ala	Gly	ALA					
			67						73		74		76					
Leu	Met	Glu	TYR	LEU	Glu	ASN	PRO	LYS	LYS	-	TYR	ILE	PRO	-	-	-		H
			70						76		77		79					
Leu	Ala	Ala	Tyr	Val	Lys	Asn	Pro	Lys	Ala	-	Phe	Val	Leu	Glu	Lys	Ser		R
			47						53		55		57					
LEU	ALA	Gln	Arg	ILE	LYS	Asn	GLY	Ser	Gln	Gly	Val	TRP	GLY	-	-	-		P
					77			80					85					
-	-	-	-	-	GLY	THR	LYS	MET	Ile	PHE	Ala	GLY	Ile	Lys	LYS	Lys		H
					88			91					95					
Gly	Asp	Pro	Lys	Ala	Lys	Ser	Lys	Met	Thr	Phe	Lys	-	Leu	Thr	Lys	Asp		R
					58			61					65					
-	-	-	-	-	PRO	ILE	PRO	MET	PRO	PRO	ASN	-	Ala	VAL	Ser	Asp		P
		91						97					102		104			
Thr	Glu	ARG	Glu	Asp	Leu	Ile	Ala	Tyr	Leu	Lys	Lys	Ala	Thr	Asn	Glu			H
		101						107					112					
Asp	Glu	Ile	Glu	Asn	Val	Ile	Ala	Tyr	Leu	Lys	Thr	Leu	Lys	-	-			R
		71						77					82					
Asp	Glu	ALA	Gln	Thr	LEU	ALA	Lys	TRP	Val	LEU	SER	Gln	LYS	-	-			P

FIG. 5. Alignment of cytochrome c (horse), cytochrome c_2 (*R. rubrum*) and c_{551} (*P. fluorescens* P6009).
Regions which fit the horse sequence well are underlined. Boxed amino acids are internal and partly boxed ones are almost buried. Those which are in contact with the haem group are heavily underlined. Capital letters indicate amino acids which are invariant in their group of cytochromes.

TABLE 2

Repeats in cytochrome sequences

	Position	Species	Peptides	Score	Probability
a	(−2)–10	*Neurospora crassa*	FS-AGDSKKGANLF	78/106	4×10^{-7}
	20–33	*R. rubrum*	FDQGGANKVGPNLF		5×10^{-7}
b	3–8	Horse	VEKGKK	43/48	6×10^{-5}
	20–25	Horse	VEKGGK		1×10^{-5}
c	21–27	Wheat	DAGAGHK	45/56	3×10^{-4}
	3–9	*R. rubrum*	DAAAGEK		7×10^{-5}
d	20–28	Wheat	VDAGAGHKQ	44/72	4×10^{-3}
	36–44	*R. rubrum*	FENTAAHKD		2×10^{-3}
e	64–72	*R. rubrum*	EANLAAYVK	56/73	5×10^{-6}
	91–99	Wheat	RADLIAYLK		5×10^{-7}
f	55–63	*Candida krusei*	AGVEWDENT	41/73	8×10^{-3}
	44–55	Screw-worm fly	AGFAYTNAN		8×10^{-3}
g	20–28	*P. fluorescens* C18	TKMVGPALK	59/72	1×10^{-6}
	78–86	Wheat	TKMVFPGLK		4×10^{-8}
h	(−3)–11	*Neurospora crassa*	GFSAGDSKKGANLFK	59/122	3×10^{-2}
	12–26	*Neurospora crassa*	TRCAECHGEGGNLTQ		3×10^{-2}

Sequences are given in one-letter code (Fig. 1). The numbering system for cytochrome c is based on the horse sequence in Fig. 5. The sequences of cytochrome c_2 and c_{551} are numbered as in Fig. 5. The similarity score 78/106 means 78 out of a maximum of 106 and is calculated from the Table in Fig. 1 with weights of unity. The first probability in each pair is the single matching probability. The second is the double matching probability.

(4) On the averaged comparison matrix $C_{Av}(p,q)$ all the repeats appear very weak, showing that they are the result of random variations in the sequences rather than underlying ancestral repeats in the DNA sequences.

(d) *Cytochrome* c, c_2 *and* c_{551}

There is a very close similarity between cytochrome c_2 from *Rhodospirillum rubrum* (R) (Dus *et al.*, 1968) and the cytochrome *c* of animals and plants (Margoliash & Schechter, 1966; Nolan & Margoliash, 1968), of which horse (H) is typical. A family comparison matrix between *R. rubrum* and eight representative cytochrome *c*'s shows an excellent fit except near the regions H(12 to 13), H(37 to 46), H(50 to 52) and H(74 to 77). The score for the two entire aligned horse and *R. rubrum* sequences shown in Figure 5 is 540/829, with a double matching probability of 1×10^{-26}.

Dus *et al.* (1968) also noticed a similarity between the sequences of *R. rubrum* and cytochrome c_{551} from *Pseudomonas* (P), which has only 82 amino acids (Ambler, 1963). The maximal family comparison matrix (span 11) between the cytochromes *c* and three c_{551} sequences (*P. aeruginosa* P6009, *P. fluorescens* C18, and *P. Stutzerii* (Ambler, 1971)) shows many scores above the 10^{-3} probability threshold for the match of P(1 to 25) with H(1 to 30). The total score for these pieces is 120/195, and the matching probability is 2×10^{-6}. In the rest of the matrix there are only two spans

which rise above the 10^{-3} probability threshold. However, the matrix suggests possible ways of aligning the rest of the sequence. In particular, four matches seem reasonable: (a) P(26 to 40) with H(31 to 45); (b) P(26 to 40) with H(47 to 60); (c) P(41 to 50) with H(61 to 70); (d) P(69 to 82) with H(89 to 102). The matching probability for (c) is $1{\cdot}0 \times 10^{-3}$ but the others are all higher. The regions around the methionines P(61) and H(80) appear completely different. These alignments were made before the structure of horse cytochrome *c* was known (Dickerson *et al.*, 1967,1971). The alignments suggested that: (1) cytochromes *c* and c_{551} might well be homologous, since they both have similar sequences around the essential Cys-X-X-Cys-His grouping, and both possess a methionine near the end of the sequence; (2) one of the two alignments shown in Figure 5 represents a reasonable match for the rest of the sequence, whereas the matrix gives no support to the match of tryptophans P(56) with H(59) suggested by Needleman & Blair (1969).

Later, Dickerson's atomic model was built, which gave the opportunity to see whether the matches were structurally possible (Dickerson, 1971). The *R. rubrum* alignment looks good because there are sharp corners, with hydrogen bonds running from residue $(n + 3)$ to n (Venkatachalam, 1968) at positions H(9), H(49), and H(75) where gaps or insertions occur. Two further corners at H(51 to 52) and H(76 to 77) rest on one another, so that the two insertions here appear to be correlated. The other substitutions appear to be structurally feasible.

For cytochrome c_{551} the predicted alignment appears to agree well with the structure up to position P(25). In particular the Val-Gly-Pro segment supports the edge of the haem group and keeps histidine H(18) in position. The matches (c) and (d) above also fit the structure well. H(61 to 70) includes a highly distorted α helix and H(89 to 102) is part of the C-terminal helix. The chief uncertainty is how to match P(26 to 40). Choice (a) deletes the loop H(46 to 60) taking a short cut across the edge of the haem group and leaving most of the internal groups which touch the haem undisturbed. Choice (b) matches Tyr P(27) to H(48), which is linked to a haem propionic group, but a difficulty is that P(27) is Leu or Phe in the two other strains of *Pseudomonas*.

I thank Dr Patricia Altham for much help and advice on the statistical calculations. I also thank Dr R. E. Dickerson and Dr D. Eisenberg for many valuable discussions and for the opportunity of testing the possible structures of cytochrome c_{551} on their atomic model of horse cytochrome *c*. I am indebted to a referee for suggestions about the scoring system.

REFERENCES

Adams, J. M. & Cory, S. (1970). *Nature*, **227**, 570.
Ambler, R. P. (1963). *Biochem. J.* **89**, 349.
Ambler, R. P. (1971). in the press.
Benson, A. M., Mower, H. F. & Yasunobu, K. T. (1966). *Arch. Biochem. Biophys.* **121**, 563.
Billeter, M. A., Dahlberg, J. E., Goodman, H. M., Hindley, J. & Weissman, C. (1969). *Nature*, **224**, 1083.
Cantor, C. & Jukes, T. (1966). *Proc. Nat. Acad. Sci., Wash.* **56**, 177.
Dayhoff, M. O. (1969). In *Atlas of Protein Sequence and Structure*, p. 85. Silver Spring, Maryland: National Biomedical Research Foundation.
Dickerson, R. E. (1971). *J. Mol. Biol.* **57**, 1.
Dickerson, R. E., Kopka, M. L., Weinzierl, J. E., Varnum, J. C., Eisenberg, D. & Margoliash, E. (1967). *J. Biol. Chem.* **242**, 3015.

Dickerson, R. E., Takano, T., Eisenberg, D., Kallai, O. B., Samson, L. & Margoliash, E. (1971). *J. Biol. Chem.* **246**, 1511.
Doolittle, R. F. & Blomback, B. (1964). *Nature*, **202**, 147.
Dus, K., Sletten, K. & Kamen, M. (1968). *J. Biol. Chem.* **243**, 5507.
Eck, R. V. & Dayhoff, M. O. (1966). *Science*, **152**, 363.
Fitch, W. M. (1966). *J. Mol. Biol.* **16**, 1, 8, 17.
Fitch, W. M. (1969). *Biochem. Genet.* **3**, 99.
Fitch, W. M. (1970). *J. Mol. Biol.* **49**, 1, 15.
Fitch, W. M. & Margoliash, E. (1967). *Science*, **155**, 279.
Gibbs, A. J. & McIntyre, G. A. (1970). *Europ. J. Biochem.* **16**, 1.
Haber, J. E. & Koshland, D. (1970). *J. Mol. Biol.* **50**, 617.
Keresztes-Nagy, S., Perini, F. & Margoliash, E. (1969). *J. Biol. Chem.* **244**, 981.
Konigsberg, W., Maita, T., Katze, J. & Weber, K. (1970). *Nature*, **227**, 271.
Margoliash, E. & Schechter, A. (1966). *Advanc. Protein Chem.* **21**, 114.
Matsubara, H. & Sasaki, R. M. (1968). *J. Biol. Chem.* **243**, 1732.
Needleman, S. B. & Blair, T. T. (1969). *Proc. Nat. Acad. Sci., Wash.* **63**, 1227.
Needleman, S. B. & Wunsch, C. D. (1970). *J. Mol. Biol.* **48**, 443.
Nolan, C. & Margoliash, E. (1968). *Ann. Rev. Biochem.* **37**, 727.
Perutz, M. F., Kendrew, J. C. & Watson, H. M. (1965). *J. Mol. Biol.* **13**, 669.
Sneath, P. H. A. (1966). *J. Theoret. Biol.* **12**, 157.
Sorm, F. & Keil, B. (1962). *Advanc. Protein Chem.* **17**, 167.
Sorm, F. & Knichal, V. (1958). *Collection Czech. Chem. Commun.* **23**, 1575.
Sugeno, K. & Matsubara, H. (1968). *Biochem. Biophys. Res. Comm.* **32**, 951.
Tanaka, M., Nakashima, T., Benson, A., Mower, H. & Yasunobu, K. T. (1966). *Biochemistry*, **5**, 1666.
Tsunoda, J. N., Yasunobu, K. T. & Whitely, H. R. (1968). *J. Biol. Chem.* **243**, 6262.
Urbain, J. (1969). *Biochem. Genet.* **3**, 249.
Venkatachalam, C. M. (1968). *Biopolymers*, **6**, 1255.
Williams, J., Clegg, J. B. & Mutch, M. O. (1961). *J. Mol. Biol.* **3**, 532.