M15214 HO#2

J. Mol. Biol. (1970) 48, 443–453

A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

Saul B. Needleman and Christian D. Wunsch

Department of Biochemistry, Northwestern University, and
Nuclear Medicine Service, V. A. Research Hospital
Chicago, Ill. 60611, U.S.A.

(*Received 21 July 1969*)

A computer adaptable method for finding similarities in the amino acid sequences of two proteins has been developed. From these findings it is possible to determine whether significant homology exists between the proteins. This information is used to trace their possible evolutionary development.

The maximum match is a number dependent upon the similarity of the sequences. One of its definitions is the largest number of amino acids of one protein that can be matched with those of a second protein allowing for all possible interruptions in either of the sequences. While the interruptions give rise to a very large number of comparisons, the method efficiently excludes from consideration those comparisons that cannot contribute to the maximum match.

Comparisons are made from the smallest unit of significance, a pair of amino acids, one from each protein. All possible pairs are represented by a two-dimensional array, and all possible comparisons are represented by pathways through the array. For this maximum match only certain of the possible pathways must be

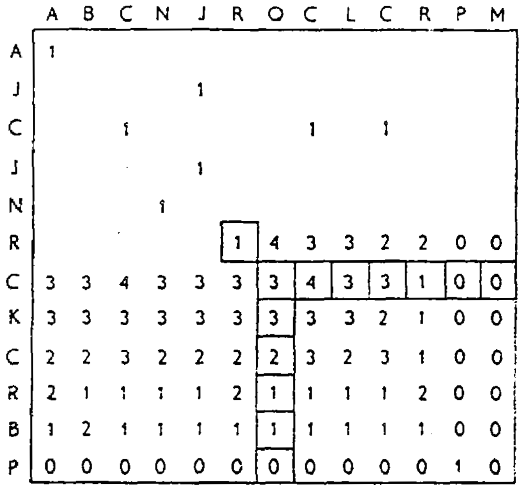
[redacted]

The maximum match can be determined by representing in a two-dimensional array, all possible pair combinations that can be constructed from the amino acid sequences of the proteins, A and B, being compared. If the amino acids are numbered from the N-terminal end, Aj is the jth amino acid of protein A and Bi is the ith amino acid of protein B. The Aj represent the columns and the Bi the rows of the two-dimensional array, MAT. Then the cell, MATij, represents a pair combination that contains Aj and Bi.

Every possible comparison can now be represented by pathways through the array. An i or j can occur only once in a pathway because a particular amino acid cannot occupy more than one position at one time. Furthermore, if MATmn is part of a pathway including MATij, the only permissible relationships of their indices are $m > i$, $n > j$ or $m < i$, $n < j$. Any other relationships represent permutations of one or both amino acid sequences which cannot be allowed since this destroys the significance of a sequence. Then any pathway can be represented by MATab . . . MATyz, where $a \geqslant 1$, $b \geqslant 1$, the i and j of all subsequent cells of MAT are larger than the running indices of the previous cell and $y \leqslant K$, $z \leqslant M$, the total number of amino acids comprising the sequences of proteins A and B, respectively. A pathway is signified by a line connecting cells of the array. Complete diagonals of the array contain no gaps. When MATij and MATmn are part of a pathway, $i - m \neq j - n$ is a sufficient, but not necessary condition for a gap to occur. A necessary pathway through MAT is defined as one which begins at a cell in the first column or the first row. Both i and j must increase in value; either i or j must increase by only one but the other index may increase by one or more. This leads to the next cell in a MAT pathway. This procedure is repeated until either i or j, or both, equal their limiting values, K and M, respectively. Every partial or unnecessary pathway will be contained in at least one necessary pathway.

In the simplest method, MATij is assigned the value, one, if Aj is the same kind of amino acid as Bi; if they are different amino acids, MATij is assigned the value, zero. The sophistication of the comparison is increased if, instead of zero or one, each cell value is made a function of the composition of the proteins, the genetic code triplets representing the amino acids, the neighboring cells in the array, or any theory concerned with the significance of a pair of amino acids. A penalty factor, a number subtracted for every gap made, may be assessed as a barrier to allowing the gap. The penalty factor could be a function of the size and/or direction of the gap. No gap would be allowed in the operation unless the benefit from allowing that gap would exceed the barrier. The maximum-match pathway then, is that pathway for which the sum of the assigned cell values (less any penalty factors) is largest. MAT can be broken up into subsections operated upon independently. The method also can be expanded to allow simultaneous comparison of several proteins using the amino acid sequences of n proteins to generate an n-dimensional array whose cells represent all possible combinations of n amino acids, one from each protein.

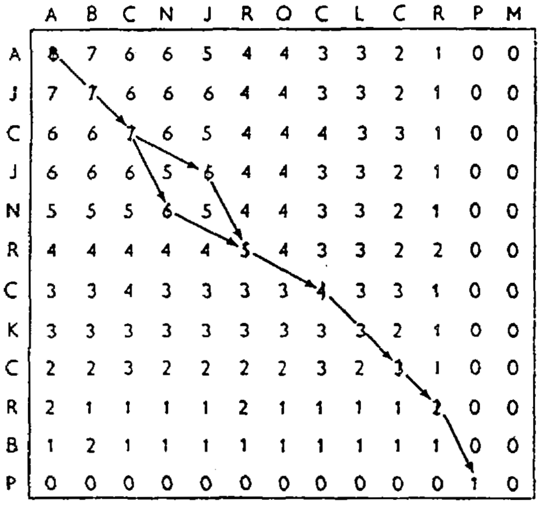
The maximum-match pathway can be obtained by beginning at the terminals of the sequences ($i = y, j = z$) and proceeding toward the origins, first by adding to the value of each cell possessing indices $i = y - 1$ and/or $j = z - 1$, the maximum value from among all the cells which lie on a pathway to it. The process is repeated for any pathway at that cell and the largest number in that row or column is equal to the maximum match; the maximum-match pathway in any row or column must begin at this number. The operation of successive summations of cell values is illustrated in Figures 1 and 2.



	A	B	C	N	J	R	O	C	L	C	R	P	M
A	1												
J					1								
C			1					1		1			
J					1								
N				1									
R						1	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

FIG. 1. The maximum-match operation for necessary pathways.

The number contained in each cell of the array is the largest number of identical pairs that can be found if that cell is the origin for a pathway which proceeds with increases in running indices. Identical pairs of amino acids were given the value of one. Blank cells which represent non-identical pairs have the value, zero. The operation of successive summations was begun at the last row of the array and proceeded row-by-row towards the first row. The operation has been partially completed in the R row. The enclosed cell in this row is the site of the cell operation which consists of a search along the subrow and subcolumn indicated by borders for the largest value, 4 in subrow C. This value is added to the cell from which the search began.



	A	B	C	N	J	R	O	C	L	C	R	P	M
A	8	7	6	6	5	4	4	3	3	2	1	0	0
J	7	7	6	6	6	4	4	3	3	2	1	0	0
C	6	6	7	6	5	4	4	4	3	3	1	0	0
J	6	6	6	5	6	4	4	3	3	2	1	0	0
N	5	5	5	6	5	4	4	3	3	2	1	0	0
R	4	4	4	4	4	5	4	3	3	2	2	0	0
C	3	3	4	3	3	3	3	4	3	3	1	0	0
K	3	3	3	3	3	3	3	3	3	2	1	0	0
C	2	2	3	2	2	2	2	3	2	3	1	0	0
R	2	1	1	1	1	2	1	1	1	1	2	0	0
B	1	2	1	1	1	1	1	1	1	1	1	0	0
P	0	0	0	0	0	0	0	0	0	0	0	1	0

FIG. 2. Contributors to the maximum match in the completed array.

The alternative pathways that could form the maximum match are illustrated. The maximum match terminates at the largest number in the first row or first column, 8 in this case.

It is apparent that the above array operation can begin at any of a number of points along the borders of the array, which is equivalent to a comparison of N-terminal residues or C-terminal residues only. As long as the appropriate rules for pathways are followed, the maximum match will be the same. The cells of the array which contributed to the maximum match, may be determined by recording the origin of the number that was added to each cell when the array was operated upon.

3. Evaluating the Significance of the Maximum Match

A given maximum match may represent the maximum number of amino acids matched, or it may just be a number that is a complex function of the relationship between sequences. It will, however, always be a function of both the amino acid compositions of the proteins and the relationship between their sequences. One may ask whether a particular result found differs significantly from a fortuitous match between two random sequences. Ideally,one would prefer to know the exact probability of obtaining the result found from a pair of random sequences and what fraction of the total possibilities are less probable, but that is prohibitively difficult, especially if a complex function were used for assigning a value to the cells.

As an alternative to determining the exact probabilities, it is possible to estimate the probabilities experimentally. To accomplish the estimate one can construct two sets of random sequences, a set from the amino acid composition of each of the proteins compared. Pairs of random sequences can then be formed by randomly drawing one member from each set. Determining the maximum match for each pair selected will yield a set of random values. If the value found for the real proteins is significantly different from the values found for the random sequences, the difference is a function of the sequences alone and not of the compositions. Alternatively, one can construct random sequences from only one of the proteins and compare them with the other to determine a set of random values. The two procedures measure different probabilities. The first procedure determines whether a significant relationship exists *between* the real sequences. The second procedure determines whether the relationship *of* the protein used to form the random sequences *to* the other proteins is significant. It bears reiterating that the integral amino acid composition of each random sequence must be equal to that of the protein it represents.

The amino acid sequence of each protein compared belongs to a set of sequences which are permutations. Sequences drawn randomly from one or both of these sets are used to establish a distribution of random maximum-match values which would include all possible values if enough comparisons were made. The null hypothesis, that any sequence relationship manifested by the two proteins is a random one, is tested. If the distribution of random values indicates a small probability that a maximum match equal to, or greater than, that found for the two proteins could be drawn from the random set, the hypothesis is rejected.

4. Cell Values and Weighting Factors

To provide a theoretical framework for experiments, amino acid pairs may be classified into two broad types, identical and non-identical pairs. From 20 different amino acids one can construct 180 possible non-identical pairs. Of these, 75 pairs of amino acids have codons (Marshall, Caskey & Nirenberg, 1967) whose bases differ at only one position (Eck & Dayhoff, 1966). Each change is presumably the result of a single-point mutation. The majority of non-identical pairs have a maximum of only one or zero corresponding bases. Due to the degeneracy of the genetic code, pair differences representing amino acids with no possible corresponding bases are uncommon even in randomly selected pairs. If cells are weighted in accordance with the maximum number of co:responding bases in codons of the represented amino acids, the maximum match will be a function of identical and non-identical pairs. For comparisons in general, the cell weights can be chosen on any basis.

If every possible sequence gap is allowed in forming the maximum match, the significance of the maximum match is enhanced by decreasing the weight of those pathways containing a large number of gaps. A simple way to accomplish this is to assign a penalty factor, a number which is subtracted from the maximum match for each gap used to form it. The penalty is assigned before the maximum match is formed. Thus the pathways will be weighted according to the number of gaps they contain, but the nature of the contributors to the maximum match will be affected as well. In proceeding from one cell to the next in a maximum-match pathway, it is necessary that the difference between each cell value and the penalty, be greater than the value for a cell in a pathway that contains no gap. If the value of the penalty were zero, all possible gaps could be allowed. If the value were equal to the theoretical value for the maximum match between two proteins, it would be impossible to allow a gap and

1967). The UGA, UAA and UAG codons were not used, but UUG was used as a codon for leucine. The subsequent data cards indicated the numerical values for a variable set.

The two-dimensional comparison array was generated row-by-row. The amino acid code numbers for *Ai* and *Bj* referenced the correspondence array to determine the type of amino acid pair constituted by *Ai* and *Bj*. The type number referenced a short array, the variable set, containing the type values, and the appropriate value from that set was assigned to the appropriate cell of the comparison array. The maximum match was then determined by the procedure of successive summations.

Following the determination of the maximum match for the real proteins, the amino acid sequence of only one member of the protein pair was randomized and the match was repeated. The sequences of β-hemoglobin and ribonuclease were the ones randomized. The randomization procedure was a sequence shuffling routine based on computer-generated random numbers. A cycle of sequence randomization–maximum-match determination was repeated ten times in all of the experiments in this report, giving the random values used for comparison with the real maximum-match. The average and standard deviation for the random values of each variable set was estimated.

6. Results and Discussion

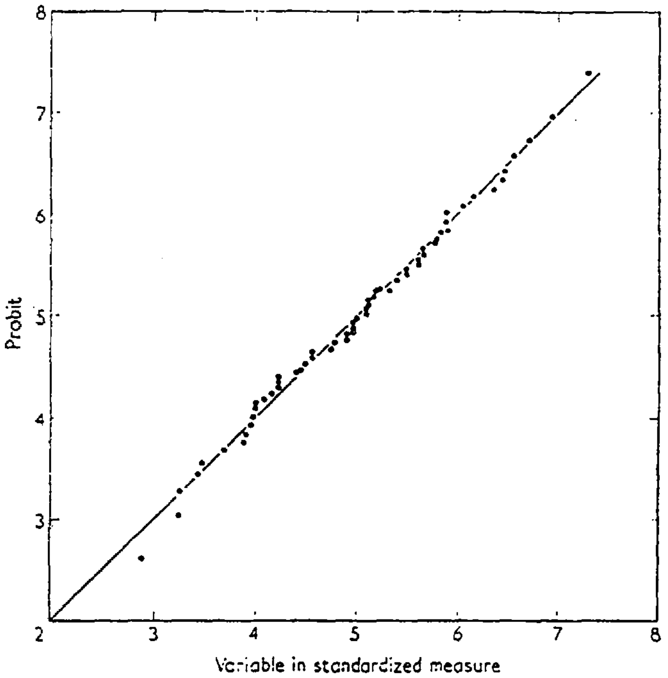


FIG. 3. Probit plot for six grouped random samples.

The solid line indicates the plot that would result from a probit analysis on an infinite number of samples from a normally-distributed population. The points represent the results of probit calculations on 60 random maximum match values that were assumed to have come from one population.

TABLE 1

β-Hemoglobin–myoglobin maximum matches

Variable set	Match values for pair types		Penalty	Maximum-match value sum		s	Real X	Minimum deletions	
	2	1		Real	Random†			Real	Random†
1	0	0	0	63·00	55·60	1·80	4·11	35	36·2
2	0	0	1·00	38·00	27·80	2·09	4·88	4	5·5
3	0·67	0·33	0	97·00	91·47	1·55	3·57	18	24·3
4	0·67	0·33	1·03	89·63	80·25	1·11	8·46	1	3·6
5	0·25	0·05	0	71·55	64·78	1·59	4·27	46	45·0
6	0·25	0·05	1·05	51·95	40·54	1·46	7·80	3	7·5
7	0·25	0·05	25	47·30	33·80	1·52	8·87	0	0

s is the estimated standard deviation: X, tho standardized value, (real – random)/s, of the maximum match of the real proteins. The values for type 3 and type 0 pairs were 1·0 and 0, respectively, in each variable set.

† An average value from 10 samples.

TABLE 2

Ribonuclease–lysozyme maximum matches

Variable set	Match values for pair types		Penalty	Maximum-match value sum		s	Real X	Minimum deletions	
	2	1		Real	Random†			Real	Random†
1	0	0	0	48·00	44·20	2·56	1·48	34	29·2
2	0	0	1·00	23·00	22·00	1·73	0·58	5	5·2
3	0·67	0·33	0	78·33	76·17	0·82	2·64	21	18·8
4	0·67	0·33	1·03	67·93	67·37	1·27	0·43	2	2·2
5	0·25	0·05	0	56·00	52·26	2·12	1·77	35	35·5
6	0·25	0·05	1·05	33·70	33·02	1·66	0·41	8	6·8
7	0·25	0·05	25	28·15	27·67	1·75	0·22	0	0

s is the estimated standard deviation; X, the standardized value, (real−random)/s, of the maximum match of the real proteins. The values for type 3 and type 0 pairs were 1·0 and 0, respectively in each variable set.

† An average value from 10 samples.

amino acids in β-hemoglobin and myoglobin can be matched. To attain this match, however, it is necessary to permit at least 35 gaps. In contrast, when two gaps are allowed according to Braunitzer (1965), it is possible to match only 37 of the amino acids. Curiously, when this variable set was used for comparing *human* myoglobin (Hill, personal communication) with human β-hemoglobin, the maximum match obtained was not significant. Differences between real and random values were highly significant, however, when other variable sets were used.

Variable set 2 attaches a penalty equal to the value of one identical amino acid pair to the search for identical amino acid pairs. This penalty will exclude from consideration any possible pathway that leaves and returns to a principal diagonal, thereby needing two gaps, in order to add only one or two amino acids to the maximum match. This set results in a total of $30 + 4 = 42$ amino acids matched (the maximum-match value plus the number of gaps is reduced to four) and the significance of the result relative to set 1 appears to be increased. Braunitzer's comparison would have a value of $37 - 2 = 35$ using this variable set, hence it was not selected by the method.

Variable sets 3 and 4 have an interesting property. Their maximum-match values can be related to the minimum number of mutations needed to convert the selected parts of one amino acid sequence into the selected parts of the other. The minimum number of mutations concept in protein comparisons was first advanced by Fitch (1966). If the type values for these sets are multiplied by three, they become equal to their pair type and directly represent the maximum number of corresponding bases in the codons for a given amino acid pair. Thus the maximum match and penalty factors may be multiplied by three, making it possible to calculate the maximum number of bases matched in the combination of amino acid pairs selected by the maximum-match operation.

β-Hemoglobin, the smaller of the two proteins, contains 146 amino acids; consequently the highest possible maximum match (disregarding integral amino-acid composition data) with myoglobin is $146 \times 3 = 438$. Insufficient data are available to analyze the result from set 3 on the basis of mutations. If it is assumed that the gap in set 4 does not exclude any part of β-hemoglobin from the comparison, this set has a maximum of $3(89{\cdot}63 + 1{\cdot}03) = 272$ bases matched, indicating a minimum of $438 - 272 = 166$ point mutations in this combination. Using this variable set and placing gaps according to Braunitzer, a score of 88·6 was obtained, thus his match was not selected. Again it may be observed that the penalty greatly enhanced the significance of the maximum match.

Variable sets 5 and 6 have no intrinsic meaning and were chosen because the weight attached to type 2 and type 1 pairs is intermediate in value with respect to sets 1 and 2 and sets 3 and 4. The maximum match for set 6 is seen to have a highly significant value.

The data of set 7 are results that would be obtained from using the frame-shift method to select a maximum match; the penalty was large enough to prevent any gaps in the comparisons. The slight differences in significance found among the maximum-match values of β-hemoglobin and myoglobin resulting from use of sets 4, 6 and 7 are probably meaningless due to small sample size and errors introduced by the assumptions about the distribution functions of random values. Finding a value in set 7 that is approximately equal to those from sets 4 and 6 in significance is not surprising. A larger penalty factor would have increased the difference from the mean in sets 4 and 6 because almost every random value in each set was the result of more gaps than were required to form the real maximum match. Further, the gaps that are allowed are at the N-terminal ends so that about 85% of the comparison can be made without gaps. If an actual gap were present near the middle of one of the sequences, it would have caused a sharp reduction in the significance of the frame-shift type of match.

Set 3 is the only variable set in Table 2 that shows a possible difference. Assuming the value is accurate, other than chance, there is no simple explanation for the difference. A small but meaningful difference in any comparison could represent evolutionary divergence or convergence. It is generally accepted that the primary structure of proteins is the chief determinant of the tertiary structure. Because certain features of tertiary structure are common to proteins, it is reasonable to suppose that proteins will exhibit similarities in their sequences, and that these similarities will be sufficient to cause a significant difference between most protein pairs and their corresponding randomized sequences, being an example of submolecular evolutionary convergence. Further, the interactions of the protein backbone, side chains, and the solvent that determine tertiary structure are, in large measure, forces arising from the polarity and steric nature of the protein side-chains. There are conspicuous correlations in the polarity and steric nature of type 2 pairs. Heavy weighting of these pairs would be expected to enhance the significance of real maximum-match values if common structural features are present in proteins that are compared. The presence of sequence similarities does not always imply common ancestry in proteins. More experimentation will be required before a choice among the possibilities suggested for the result from set 3 can be made. If several short sequences of amino acids are common to all proteins, it seems remarkable that the relationship of ribonuclease to lysozyme in six of the seven variable sets appears to be truly a random one. It should be noted, however, that the standard value of the real maximum-match is positive in each variable set in this comparison.

This method was designed for the purpose of detecting homology and defining its nature when it can be shown to exist. Its usefulness for the above purposes depends in part upon assumptions related to the genetic events that could have occurred in the evolution of proteins. Starting with the assumption that homologous proteins are the result of gene duplication and subsequent mutations, it is possible to construct several hypothetical amino-acid sequences that would be expected to show homology. If one assumes that following the duplication, point mutations occur at a constant or variable rate, but randomly, along the genes of the two proteins, after a relatively short period of time the protein pairs will have nearly identical sequences. Detection of the high degree of homology present can be accomplished by several means. The use of values for non-identical pairs will do little to improve the significance of the results. If no, or very few, deletions (insertions) have occurred, one could expect to enhance the significance of the match by assigning a relatively high penalty for gaps. Later on in time the hypothetical proteins may have a sizable fraction of their codons changed by point mutations, the result being that an attempt to increase the significance of the maximum match will probably require attaching substantial weight to those pairs representing amino acids still having two of the three original bases in their codons. Further, if a few more gaps have occurred, the penalty should be reduced to a small enough value to allow areas of homology to be linked to one another. At a still later date in time more emphasis must be placed on non-identical pairs, and perhaps a very small or even negative penalty factor must be assessed. Eventually, it will be impossible to detect the remaining homology in the hypothetical example by using the approach detailed here.

From consideration of this simple model of protein evolution one may deduce that the variables which maximize the significance of the difference between real and random proteins gives an indication of the nature of the homology. In the comparison of human β-hemoglobin to whale myoglobin, the assignment of some weight to type 2 pairs considerably enhances the significance of the result, indicating substantial evolutionary divergence. Further, few deletions (additions) have apparently occurred.

It is known that the evolutionary divergence manifested by cytochrome (Margoliash, Needleman & Stewart, 1963) and other heme proteins (Zuckerkandl & Pauling, 1965) did not follow the sample model outlined above. Their divergence is the result of *non-random* mutations along the genes. The degree and type of homology can be expected to differ between protein pairs. As a consequence of the difference there is no *a priori* best set of cell and operation values for maximizing the significance of a maximum-match value of homologous proteins, and as a corollary to this fact, there is no best set of values for the purpose of detecting only slight homology. This is an important consideration, because whether the sequence relationship between proteins is significant depends solely upon the cell and operation values chosen. If it is found that the divergence of proteins follows one or two simple models, it may be possible to derive a set of values that will be most useful in detecting and defining homology.

The most common method for determining the degree of homology between protein pairs has been to count the number of non-identical pairs (amino acid replacements) in the homologous comparison and to use this number as a measure of evolutionary distance between the amino acid sequences. A second, more recent concept has been to count the minimum number of mutations represented by the non-identical pairs. This number is probably a more adequate measure of evolutionary distance because it utilizes more of the available information and theory to give some measure of the number of genetic events that have occurred in the evolution of the proteins. The approach outlined in this paper supplies either of these numbers.

This work was supported in part by grants to one of us (S.B.N.) from the U.S. Public Health Service (1 501 FR 05370 02) and from Merck Sharp & Dohme.

REFERENCES

Braunitzer, G. (1965). In *Evolving Genes and Proteins*, ed. by V. Bryson & H. J. Vogel, p. 183. New York: Academic Press.

Canfield, R. (1963). *J. Biol. Chem.* **238**, 2698.

Eck, R. V. & Dayhoff, M. O. (1966). *Atlas of Protein Sequence and Structure*. Silver Spring, Maryland: National Biomedical Research Foundation.

Edmundson, A. B. (1965). *Nature*, **205**, 883.

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

Konigsberg, W., Goldstein, J. & Hill, R. J. (1963). *J. Biol. Chem.* **238**, 2028.

Margoliash, E., Needleman, S. B. & Stewart, J. W. (1963). *Acta Chem. Scand.* **17**, S 250.

Marshall, R. E., Caskey, C. T. & Nirenberg, M. (1967). *Science*, **155**, 820.

Needleman, S. B. & Blair, T. H. (1969). *Proc. Nat. Acad. Sci., Wash.* **63**, 1127.

Smyth, D. G., Stein, W. G. & Moore, S. (1963). *J. Biol. Chem.* **238**, 227.

Zuckerkandl, E. & Pauling, L. (1965). In *Evolving Genes and Proteins*, ed. by V. Bryson & H. J. Vogel, p. 97. New York: Academic Press.