J. Mol. Biol. (1972) **64**, 417–437

Repeating Sequences and Gene Duplication in Proteins

A. D. McLachlan

Medical Research Council
Laboratory of Molecular Biology
Hills Road, Cambridge
England

(*Received 23 October 1970, and in revised form 17 November 1971*)

The theory that proteins have evolved by repeated internal duplication of short segments of polypeptide chains has been tested by looking for repeats and near repeats in over 50 different proteins, many of them of known structure. The probability that the observed repeats could arise by chance has been calculated.

The search does not yield a single new example where the evidence for gene duplication is compelling. No protein shows a unique internally consistent pattern of repeats which both correlates with repeats in the structure and cannot be explained by chance. The evidence is discussed in detail for haemoglobin, chymotrypsin, subtilisin and carboxypeptidase. The evolution of complex large proteins from simple small ones has probably been a process of continuous growth in which chains have been gradually added to the outer surface surrounding an invariable core near the active centre.

1. Introduction

The idea has often been put forward that in the earliest stages of evolution of primitive forms of life the primitive proteins began with the production of repeating amino-acid sequences composed of a number of short similar segments. Thus, a large protein could be built up by repeating a small number of basic sequences which might form structural building blocks for the whole large structure. It is assumed that the repeating amino-acid sequence is coded for by a repeating DNA sequence, which is formed by duplicating again and again short pieces from the original sequence. Bacterial proteins then are expected to be more primitive and repetitive than those from mammals.

The bacterial ferredoxins (Benson, Mower & Yasunobu, 1967) have been cited as evidence for this theory (Jukes, 1966; Eck & Dayhoff, 1966), since the second half of the amino-acid sequence is an almost exact duplicate of the first, and each half has a regular periodic arrangement of cysteines. More recently Dus, Sletten & Kamen (1968) noticed repetitive pieces in the sequence of cytochrome c_2 from *Rhodospirillum rubrum*, and suggested that these features were survivals of a primitive ancestral repeating sequence. Other repeats have been noted in clupeine Z from the pacific herring (Black & Dixon, 1967), subtilisins BPN′ and Carlsberg (Markland & Smith, 1967; Wright, Alden & Kraut, 1969; Smith, DeLange, Evans, Landon & Markland, 1968), the haemoglobins (Cantor & Jukes, 1966; Fitch, 1966), and carboxypeptidase A (Bradshaw, Neurath & Walsh, 1969; Neurath, Bradshaw & Arnon, 1969). Thus, there appears to be a considerable body of circumstantial evidence in favour of the theory.

The object of the work described in this paper was to test the sequences of a large number of proteins, particularly those the structures of which are already known, for evidence that they could have been formed originally from a mosaic of repeating sequences. It is fairly straightforward to test whether two long halves of a long protein sequence have evolved from a common ancestor, by calculating the minimum number of base changes required, according to the genetic code (Fitch, 1966; Nolan & Margoliash, 1968; Needleman & Blair, 1969), or by looking for sequences of chemically similar amino acids (Haber & Koshland, 1970). To test whether a given sequence can have been formed from a relatively *large* number of *short* repeating pieces is much more difficult, but I believe that the methods which I have used are sufficiently sensitive to detect any underlying pattern if it exists. I have already (McLachlan, 1971) given reasons why the repeats in cytochrome *c* are not significant. We shall now see that this further investigation fails to show any convincing new examples of gene duplication or ancestral repetition, and supports Haber & Koshland's conclusion (1970) that many of the repeats noted in protein sequences may easily have occurred by chance.

If these proteins ever did arise from ancestral repeating sequences, no regularity in sequence or structure now remains which cannot reasonably be accounted for in other ways.

A second reason for searching for similar short segments in protein sequences is to see whether chemically similar segments have a similar three-dimensional structure (Low, Lovell & Rudko, 1968). We shall see that the structures of even very similar segments are often quite different, so that there is no simple relationship.

2. Gene Duplication

Parallel duplication of a gene to produce two distinct new genes is a well-known event in evolution (Bridges, 1936; Stephens, 1951; Jukes, 1966; Dixon, 1966). An example is the formation of the haemoglobin γ chains from the normal β chain (Ingram & Stretton, 1961). Series duplication, in which the new gene codes for a single polypeptide chain of twice the original length, has also occurred sometimes. The constant parts of the antibody heavy chains are good examples (Edelman *et al.*, 1969; Cohen & Milstein, 1967*a*,*b*; Singer & Doolittle, 1966; Milstein & Pink, 1970). Duplication has also occurred in ferredoxin (Tanaka, Nakashima, Benson, Mower & Yasunobu, 1966; Benson *et al.*, 1967; Eck & Dayhoff, 1966), haptoglobin (Smithies, Connell & Dixon, 1962; Dixon, 1966; Black & Dixon, 1968) and probably in the diheme cytochrome c_3 from *Desulphovibrio vulgaris* (Ambler, 1968). However, in spite of these examples, series duplication is a rare genetic event, and when duplication does begin to occur the extra piece may often be eliminated by looping out a portion of the DNA helix, so that it is cut out and the defect is repaired (Russell *et al.*, 1970). Because duplication is so rare one needs to look critically at the evidence in any case where it is postulated to explain a repeat in a protein sequence.

3. Tests for Significance

The usual method for testing whether two sections of a protein sequence are ancestrally related is to calculate the minimum number of base changes needed to convert one into another, according to the genetic code. This quantity is called the minimum mutation distance (Fitch, 1966,1970; Dayhoff, 1969). Two long proteins can be tested for homology by taking a long span of, say, twenty amino acids and comparing every possible segment of one protein with every possible segment of the other. If the proteins are unrelated the statistical distribution of the minimum mutation distances is approximately Gaussian. For related sequences there is an abnormally high number of short mutation distances, and Fitch's tests (Fitch & Margoliash, 1967; Fitch, 1970) can be used to detect deviations from a Gaussian curve.

For comparing *short* segments of sequence the minimum mutation distance is a poor test unless the pieces are closely related. The actual mutation distance can differ significantly from the minimum. Single-step mutations are not particularly significant by themselves because the genetic code (Crick, 1968) allows a large number of mutations to occur between amino acids which are completely dissimilar in size, shape and charge. Thus it is possible to pair off sections of sequences which are connected by many potential single-step mutations and identities, but could never conceivably have the same three-dimensional structure. Now the structure of a protein evolves much more slowly than the amino-acid sequence. For instance, in the haemoglobins even very distant species such as insect (Huber, Formanek & Epp, 1968), marine worm (Padlan & Love, 1968), lamprey (Braunitzer & Fujiki, 1969), carp and mammals (Dayhoff, 1969) share a common molecular architecture (Perutz, Muirhead, Cox & Goaman, 1968; Perutz, Kendrew & Watson, 1965). Hence a more discriminating test of distant evolutionary relationship is to ask whether two sequences could conceivably have a common structure; or whether corresponding pairs of amino acids are structurally similar in size, shape, polarity and so on (Thiebaux & Pattee, 1967). A good match between two sequences, either through mutation distances or chemical properties, only points to common ancestry if it has a low probability of occurring for other reasons; the degeneracy of the genetic code is so great that approximately 3^n nucleotide sequences can code for a given set of n amino acids, and the probability that even two short identical peptides correspond to the same DNA is very low.

The methods used in this paper have been described already (McLachlan, 1971). They are developed from those used by Fitch (1966), Cantor & Jukes (1966), Needleman & Blair (1969) and Haber & Koshland (1970).

The first step is to set up a measure of similarity for each pair of amino acids. In McLachlan (1971) this was based on the observed frequencies of amino-acid replacements in homologous proteins. However, in this work, which was done earlier, we used a more intuitive scoring scheme† based on polar or non-polar character, size, shape and charge (Sneath, 1966). To each pair of amino acids i, j we assign a similarity score $m(i, j)$ which ranges from 0 to 6 (see Table 1). The score $m(i, i)$ for matching an amino acid with itself is normally 5, but rises to 6 for the less common ones.

To compare two sequences A and B, in which the amino acids at positions p and q are a_{pA} and a_{qB}, one first sets up a score matrix $M(p,q) = m(a_{pA}, a_{qB})$ in which $m(a_{pA}, a_{qB})$ is the similarity between a_{pA} and a_{qB}. The next step is to assign a score to the match between two segments of protein centred on positions p and q. Here the

† The scores assigned to the repeating segments observed in this work are a little different if one uses the more objective scoring scheme of McLachlan (1971), and the matching probabilities are also affected, but the alterations do not affect any of the conclusions which we reach.

TABLE 1

Chemical similarity scores for the amino acids

Score	Pairs							
6	FF	MM	YY	HH	CC	WW	RR	GG
5	LL	II	VV	SS	PP	TT	AA	QQ
	NN	KK	DD	EE				
3	FY	FW	LI	LM	IM	ST	AG	QE
	ND	KR						
2	FL	FM	FH	IV	YH	YW	SC	HQ
	QN	DE						
1	FI	FV	LV	LP	LY	LW	IT	IY
	IW	MV	MY	MW	VP	SA	SN	SQ
	PT	PA	TA	TN	HN	HW	QK	QD
	NE							
0	All others, including unknowns and deletions							

One-letter code. F(Phe), L(Leu), I(Ile), M(Met), V(Val), S(Ser), P(Pro), T(Thr), A(Ala), Y(Tyr), H(His), Q(Gln), N(Asn), K(Lys), D(Asp), E(Glu), C(Cys), W(Trp), R(Arg), G(Gly), B(Asx), Z(Glx).

score for two segments of length s is taken to be a weighted sum of the successive M values, with weights W_h:

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

It is convenient to take s an odd number, $s = 2g + 1$. s should be large enough to make a high score statistically significant but short enough to avoid missing gaps. The weights W_h can be chosen at will. The matrix of weighted sums $C(p,q)$ is called the *comparison matrix* for the two sequences. If two sections of sequence are similar the comparison matrix shows a line of high scores running parallel to the main diagonal. A computer can be set to construct the matrix and print out suitable symbols to indicate scores which have different levels of significance, giving a correlation diagram for the sequences (Gibbs & McIntyre, 1970).

To judge the significance of a high score $C(p,q)$ one needs to know the probability distribution of the scores for a pair of random sequences of given compositions. The scores $C(p,q)$ and $C(p + r, q + r)$ will in fact be highly correlated if $r < s$. An exact calculation of probabilities is difficult, but we can calculate a related distribution exactly.

Consider the following 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 protein A or B. A set of s cards $a_1, a_2, \ldots, a_s$ are drawn in order from A and compared in turn with another set $b_1, b_2, \ldots, b_s$ drawn from pack B, the score for each pair being $m_r = m(a_r, b_r)$. The weighted score for the entire match is defined as:

$$M = \sum_r W_r m_r. \tag{2}$$

Then the *double matching probability* Q_{AB} (M), that the observed score in this experiment is greater than or equal to M, can be calculated (see McLachlan, 1971); so can the mean score and the standard deviation. If the number of cards is large the central part of the distribution is approximately Gaussian.

As an example, consider the comparison matrix for a haemoglobin α chain against the horse β chain, with span 5 and weights 1, 2, 3, 2, 1. Here the calculated mean and standard deviation are 6·12 and 6·22, respectively. In this case scores of 40, 30, 24 and 18 correspond to double matching probabilities of 9×10^{-5}, 3×10^{-3}, 2×10^{-2}, and 3×10^{-3}.

Another useful probability is the *single matching probability*. Suppose that we place a *given* set of s cards $a_1, a_2, \ldots, a_s$ in order on a table. Now draw cards $b_1, b_2, \ldots, b_s$ in turn from an infinite shuffled pack with the composition of protein B, and record each score $m_r = m(a_r, b_r)$. The single matching probability $R_{\alpha B}(M)$ is the probability that the sum $m_1 + m_2 + \cdots + m_s$ is greater than, or equal to, M. It depends on the compositions of the peptide α and of the entire protein B.

For a match to be statistically significant one requires a low value for the probabilities Q_{AB} or $R_{\alpha B}$. Since a comparison of every pair of short segments from two proteins of lengths n_A, n_B entails about $n_A n_B$ potential matches there are often as many as 10^4 to 10^5 entries in the comparison matrix. Thus, events with probabilities $R_{\alpha B}$ in the range 10^{-4} to 10^{-6} are likely to be observed reasonably often in pairs of random sequences.

So far we have considered only a single pair of sequences. Often one has families of proteins, such as the haemoglobins, in which the sequences from related animals form a homologous series. With such a family it is possible that remnants of an ancestral repeating pattern might persist in some members, but not in others, and yet still be detected by analysis of the entire family as a whole. With this object in mind we have set up a *family comparison matrix* $C_{\max}(p,q)$. Let $S_{1A}, S_{2A} \ldots$ be a set of j_A homologous sequences from a family A, and let $S_{1B}, S_{2B}, \ldots$ be a set of j_B sequences which form a second family B (A and B could be the same). Suppose that $C_{xy}(p,q)$ is the p,q element of the comparison matrix for S_{xA} and S_{yB}. Then $C_{\max}(p,q)$ is defined to be the maximum of the $C_{xy}(p,q)$ for all the pairs x,y (p,q being fixed). The family comparison matrix displays all the strongest matches between the two groups, and one can then search for any underlying pattern.

If an interesting match is found one can first try to calculate the probability that the observed regularity could have arisen by chance (Šorm & Knichal, 1958; Fitch, 1970). One fundamental difficulty is that any amino-acid sequence can be tested for an infinite number of different special features, and every sequence will therefore possess some unique but insignificant coincidental feature which may be exceedingly improbable (Šorm & Keil, 1962). By fastening on such features, which are suggested by the particular sequence in question, one can easily be persuaded that some subtle pattern of repeats exists (Urbain, 1969; Williams, Clegg & Mutch, 1961).

To avoid this danger we have restricted ourselves to one well-defined class of regularity, expressed as a high score on the comparison matrix. We have also done control experiments by comparing totally unrelated proteins such as carboxypeptidase and haemoglobin, or one protein sequence with another written in reverse order. These experiments confirm the results of the statistical calculations.

Suppose that several matches of various lengths and strengths have been found. If they are really remnants of a set of ancestral repeats, which may be interrupted by insertions or deletions, they must not merely be statistically improbable. They must satisfy the test of *mutual consistency*. Suppose for example that a segment *abcdefg* is related to two pieces $a'b'c'd'e'$ and $c''d''e''f''g''$. There should then be a match between $c'd'e'$ and $c''d''e'''$. Often, however, such a match is absent, and there may even be a weak match out of phase in which, say $a'b'c'd'e'$ resembles $c''d''e''f''g''$.

Once a mutually consistent set of repeats has been found a further test can be applied. This is to see whether the related segments share a common structure. An underlying regularity of structure would be good evidence that a protein had begun as a repeating sequence. On the other hand, similar sequences do not as a rule have similar conformations in different parts of a protein, and the structure may change in the course of evolution. Absence of a regular structure cannot therefore rule out the possibility that a sequence repeat is ancient, but the possibility becomes more remote.

The main features to be looked for in related segments of a protein which has evolved by repeated internal duplication of its sequence are therefore as follows:

(1) identical or chemically similar amino acids in weakly matched pieces; short mutation distances in strongly matched pieces;
(2) low probability that the observed repeats could occur by chance;
(3) mutually consistent sets of repeats;
(4) similar structures;
(5) persistence of a repeat in many members of a family of homologous proteins.

The next sections describe the correlations found in a variety of proteins. We use the word *correlation* to describe any repeating feature which appears to be interesting, without implying by this word that it is statistically significant in any way.

4. Haemoglobin and Myoglobin

The haemoglobins are a suitable family for testing because the structure is known and a great many sequences are available (Perutz *et al.*, 1965). The fact that the structure is nearly all built of α helices, many of which are of similar lengths, makes it seem possible that if any proteins have evolved in a repetitive way then haemoglobin could be one of them.

A preliminary study of myoglobin showed some promising similarities between helices A and B, B and G, and B and E. Also the C helix and the FG corner appeared to be related. The score matrix based on the genetic code showed several long runs of single-step mutations. Individual haemoglobin chains showed various short repeats, including some quite long correlations.

The next step was to make a family comparison matrix for ten homologous haemoglobin chains. One matrix, with span of 5, weights 1, 2, 3, 2, 1 and contour levels of 40, 30, 24, 18 showed all the short repeats with tripeptides and other fragments. Another, with a span of 11, gave the longer correlations which are listed in Table 2.

These correlations are found by comparing each of the ten sequences with each of the others in every possible registration. A correlation which is strong between one pair of species may be weak between others, because the sequences vary considerably. For each correlation Table 2 also lists the shift. This is the number of spaces which one has to slide the sequences past one another in order to align the two matched segments. The shift is used to test for consistency. For example, in Table 2 the shifts for m (relating helices E and H) and c (relating A and H) are 66 and 117. There should, therefore, be some relation between helices A and E with a shift of $(117-66) = 51$. The comparison matrix does show such a relation, but it is very weak, and insignificant compared with the other observed correlation (shift of 58).

The individual correlations, though never very long, are quite strong, in the sense that each pair of spans is at least as similar as the majority of diagonal spans in the comparison of two homologous but distantly related proteins. For example, the correlation between the E helices of sperm whale myoglobin and horse haemoglobin β is weaker than that between the two segments *m* in Table 2. But although the individual correlations are strong it proves to be impossible to relate the different fragments to one another in a meaningful way according to any evolutionary family tree. The correlations *b*, *c* and *m* discussed above illustrate the difficulty; one cannot reconcile them simultaneously and thus demonstrate a single common ancestral sequence for helices A, E and H. A search through the 19 correlations of Table 2, taken in threes, shows that there is no underlying pattern of repeats. Instead one must regard the correlations as the chance result of unrelated variations in different parts of the molecule.

Are these correlations related to the structure? It would be interesting if they brought out some general similarity between the different helical regions, or some common features of the corners. The most permanent feature of the haemoglobin sequences (apart from the haem-linked histidines) is the persistence of non-polar side chains at certain internal positions (Perutz *et al.*, 1965). The long correlation *m* between helices E and H, which has already been commented on by Fitch (1966) and Cantor & Jukes (1966), makes use of the fact that both helices have a similar distribution of internal side chains, with non-polar sites at positions 4, 8, 11, 12, 15 and 19. The distribution of non-polar sites in the other helices is less regular and cannot be matched to that of E or H except over short sections. In fact there is little or no relation between the structure and the other correlations. For example, in the correlation *g*, helix B matches the GH corner and the first half of the H helix. The most that can be said is that since haemoglobin has such a high helix content (even the corners contain helical fragments), many pairs of short segments have some local structural similarity. As a further test we compared the haemoglobin sequences with carboxypeptidase and chymotrypsin to see whether the helical regions in these latter proteins would be picked out. No significant relationship was detected. Haemoglobin was also tested against the sequence of horse α chain in reverse order, and gave several quite long correlations. This reinforces the view that most of the correlations in Table 2 are random.

These tests cannot, of course, prove that haemoglobin did not originally arise from a primitive repeating sequence. But they do demonstrate that no trace of such a sequence remains detectable beneath the correlations which naturally arise by chance in any long protein chain.

5. Chymotrypsin

The X-ray analyses of α-chymotrypsin and elastase (Birktoft, Blow, Henderson & Steitz, 1970; Matthews, Sigler, Henderson & Blow, 1967; Shotton & Watson, 1970; Watson, Shotton, Cox & Muirhead, 1970) show that both molecules share a common structural framework. This consists principally of two independent hydrogen-bonded substructures which rest upon one another, forming two sides of the active site. The substructures have a similar pattern of antiparallel pleated sheets linked by hydrogen bonds, with loops closed by disulphide bridges. There is therefore a possibility that the two substructures might have been formed by gene duplication.

Birktoft & Blow examined the amino-acid sequence to see whether there was any

Table 2

Strongest repeats in haemoglobin and myoglobin

Pair	Shift	Regions	Sequences	Species
a	42	A3-A13	A D K T N V K A A W G	Human α
		CD5-D7	G D L S N A K A V M A	Horse β
b	58	A8-B1	V L H V W G K V G A H	Myoglobin/human α
		E11-EF1	V L H S F G K A V G H	Horse β/rabbit α
c	117	A7-B2	A V L A L W D K V E A D V	Horse β/myoglobin
		H3-H15	A V H A S L D K F L A D V	Rabbit α
d	72	A7-AB1	N L K G T F A K L S	Kangaroo β
		EF4-F5	N V K A A W S K V G	Horse α
e	44	A13-B3	G K V G A H A G	Human α
		E2-E9	P K V L A H G A	Kangaroo β
f	40	B1-B9	D V A G H G Q D I	Human α
		E3-E11	Q V K A H G K K V	Horse α
g	99	B1-B13	H A G E Y G A E A L E R M	Human α
		GH1-H8	H P G N F G A D A Q G A M	Myoglobin
h	37	CD2-D6	D H F G D L S N A K V M	Kangaroo β
		EF4-F7	D L P G A L S D L S N L	Horse α
i	40	CD3-D3	S F G D L S D P	Horse β
		EF8-F7	T F A Q L S E L	Human γ
j	30	CD4-E1	F K H L S N A K A V M A N	Myoglobin/kangaroo β
		E19-F3	I K H L D D L K G T F A Q	Human γ

k	33	CD4-D7	F G D L S S A D A I L	Human γ
		EF2-F4	L D D L K G A F A S L	Cow γ
l	46	E8-E16	G V T V L H S F G E G	Myoglobin/horse β
		G12-GH2	L V T V L H S R H P G	Human γ/myoglobin
m	66	E2-EF1	P K V K A H G K K V L G A F S D G L A H	Human β
		H2-H21	P E L Q A S Y Q K V V A G V A N A L A H	Horse β
n	17	E12-EF1	L T S L G D A I K H	Human γ
		F1-FG1	L S T L S D L H A H	Rabbit α
o	10	E15-EF6	L G D A I K N L D N L K	Kangaroo β
		EF5-F8	M P N A L S A L S D L H	Human α
p	37	E19-F3	V G H L D D L P G A L S T	Rabbit α
		G14-H2	V G I M F Y L P G D F P P	Carp α
q	3	EF6-F4	P G A L S D L	Horse α
		F1-F7	F A A L S E L	Horse β
r	28	G6-G17	K L L G N V L V T V L A	Human γ
		H10-H21	K F L A D V S T V L T S	Rabbit α
s	4	G7-GH1	L L G N V L A L V V A R H F	Horse β
		G11-GH5	V L V T V L A I H F G K E F	Human γ

These 19 repeats are the strongest selected from a family comparison matrix, using a span of 11 with weights of 1, 2, 2, 3, 3, 3, 3, 3, 2, 2, 1. Those correlations were selected in which there were 7 or more consecutive spans with a score of at least 60. Ten sequences were used. Haemoglobin α: horse, carp, rabbit, human. Haemoglobin β: horse, human, kangaroo. Haemoglobin γ: cow, human. Myoglobin: sperm whale. The total number of spans compared was approximately 1,000,000. Sequences were taken from Dayhoff (1969) except for the kangaroo β chain (Air, Thompson, Richardson & Sharman, 1971). The sequences of lamprey and insect haemoglobins show no additional correlations.

TABLE 3

Repeats in the chymotrypsin family

a	19–27	*e*	128–135
	39–47		179–185
b	12–20	*f*	132–139
	142–150		178–185
c	15–24	*g*	138–145
	201–230		170–175
d	39–46	*h*	141–148
	179–185		192–199

Positions are numbered as in chymotrypsinogen (Blow, Birktoft & Hartley, 1969; Shotton & Hartley, 1970; Brown & Hartley, 1966). Correlations are taken from a comparison matrix with span of 11, weights 1, 2, 2, 3, 3, 3, 3, 3, 2, 2, 1. They have at least 7 consecutive spans with a score $\geq$ 60, using the scores of Table 1. Five sequences were used: chymotrypsinogen A, B (cow), elastase (pig), trypsin (cow) and fragments of trypsinogen (dogfish). (References: Shotton & Hartley, 1970; Bradshaw, Neurath, Tye, Walsh & Winter, 1970.)

repeat, but found none (Birktoft *et al.*, 1970). We searched for repeats, using the family comparison matrix for five sequences (Table 3). The matrix shows eight repeats of similar strength to those in haemoglobin, but they bear no relation to the pattern of hydrogen bonding or the disulphide bridges. The strongest repeat:

G	L	S	R	I	V	N	G	E	(12–20) chymotrypsin
G	L	T	R	-	T	N	G	Q	(142–150) elastase

relates the junctions between sections A, B and sections B, C of the chain, which are cut when the enzyme is activated. It is not statistically significant, since the matching probability for these two segments is of the order of 1 in 100,000.

The active sites of chymotrypsin and subtilisin are similar, since they each contain histidine, serine and aspartic acid residues arranged in almost the same relative orientations in space. These amino acids do not occur in the same order in the two sequences, and the structural organization of the two proteins is quite different (Wright *et al.*, 1969), so that there can be no evolutionary connection between them. Instead these enzymes are a good example of convergent evolution. The comparison matrix shows a few weak correlations which bear no relation to the active site. There is also one strong repeat:

A	N	T	V	P	Y	Q	V	S	(24–32) trypsin
A	Q	S	V	P	Y	G	V	S	(1–9) subtilisin BPN′.

The matching probability is approximately 3×10^{-6}, which is not very low. It is interesting that two such similar fragments have entirely different structures. The first contains a corner between two pieces of β structure. The second lies within the first two helices of subtilisin.

6. Lysozyme and Ribonuclease

The structures of lysozyme (Blake, Mair, North, Phillips & Sarma, 1967) and ribonuclease (Wyckoff *et al.* 1967, 1970; Kartha, Bello & Harker, 1967) are broadly similar: both have two wings and a central cleft; both have an irregular structure made up of α helices and β structure; both have an active site in the cleft which binds a sugar molecule; the sequences are almost the same length. In 1967, Manwell (1967) noticed some correlations between their sequences which he considered to be statistically significant in the light of the genetic code. Hence he argued that there might be a distant evolutionary link between the two enzymes. Later investigations by Needleman & Blair (1969) and Haber & Koshland (1970) have shown that there is no statistically significant relation between the amino-acid sequences, even if one introduces several deletions, and our comparison matrix confirms their conclusions. Nevertheless there are a number of short peptides which are identical in both proteins (Low *et al.*, 1968). These suggested the possibility that if both proteins arose by repeating a number of short pieces of sequence, then lysozyme and ribonuclease might be formed from the same set of pieces, put together in a different pattern.

A family comparison matrix for three lysozyme sequences and three ribonucleases shows several short correlations, of which the strongest are shown in Table 4.

TABLE 4

Similarity between lysozyme and ribonuclease

a	A	A	K	F	E	S	N	F	L(31- 38)	Hen
	A	A	K	F	E	R	Q	H	R(8–15)	Cow
b	M	K	R	H	G	L			L(11–16)	Hen
	M	K	R	Q	G	M			R(33–38)	Cow
c	D	V	Q	A					L(121–124)	Hen
	D	V	Q	A					R(56–59)	Rat
d	A	L	C	S	E	K			L(114–119)	Lactalbumin
	A	I	C	S	Q	K			R(59–64)	Rat
e	S	S	N	I	C	N			L(72–77)	Lactalbumin
	S	S	N	Y	C	N			R(25–30)	Cow

Lysozymes: hen (Canfield, 1963), part of duck (Jollès, 1969), lactalbumin (Brew, Vanaman & Hill, 1967). Ribonucleases: cow (Smyth, Stein & Moore, 1963), rat, part of horse (Beintema & Gruber, 1967).

Each has a matching probability between 10^{-5} and 10^{-6}. However, these fragments cannot be fitted into any consistent pattern, and one is forced to conclude that they arise by chance.

Some of the pairs of segments in Table 4 have similar structures. Both segments of *a*, *c* and *d* are α-helical. The partners in *b* and *c* fold differently.

Comparisons between hen lysozyme and phage T4 lysozyme (Tsugita & Inouye, 1968), and between rat ribonuclease and nuclease T1 (Takahashi, 1965) showed no significant correlations.

7. Subtilisin

The sequences of subtilisins BPN′ (Markland & Smith, 1967) and Carlsberg (Smith *et al.*, 1968) are much more repetitive than haemoglobin or chymotrypsin, and the most prominent repeats have already been noted (Smith *et al.*, 1968; Wright *et al.*, 1969). Table 5 shows the longest segments which are picked out on a comparison matrix which uses the genetic code. Almost the same pieces are picked out by using the chemical similarity test. Here a_2, a_3 and a_4 are closely related to a_1, with chemical matching probabilities of the order 10^{-4}, 10^{-5} and 10^{-5}, respectively. Segments a_3 and a_4 are also related, with a probability of 10^{-4}. The probabilities for matching

Table 5

Repeats in subtilisin

Section											Sequence												Position
a_5	S	I	G	V	L	G	V	A	P	S	S	A	L	Y									(78–91)
a_4				N	V	K	V	A	V	I	D	S	G	I	D	S	S	H	P	N	L		(26–42)
a_3				N	L	K	V	A	G	G	A	S	M	V	P	S	E	T	P	N	F		(41–58)
a_2						H	V	A	G	T	V	A	A	L									(67–75)
a_1		M	A	S	P	H	V	A	G	A	A	A	L	I	L	K	S	H	P	N	W		(221–241)
b_1		V	A	S	G	V	V	V	V	A	A	A	G	N	Q	G	S	T	G	S	S	S	(143–163)
b_2											A	L	H	S	Q	G	Y	T	G	S			(15–25)
b_3		(P	S	S	A	L)	Y	A	V	K	V	L	G	N	A	G	S	(Q	G	Y	S)		(91–101)
b_4					(S	V	I	A)	V	G	A	V	D	S	S	N	Q	R	A	S	F	S	(177–190)

These segments are taken from a family comparison matrix, span 19, using a scoring system based on the genetic code: 5 for an identity, 1 for a single-step mutation. Regions are selected if the sum of the scores is greater than 36 in at least 8 consecutive spans. Segment a_5 is related to a_1 rather more weakly. a_2 is selected because of its chemical similarity to a_1. In the matrix a_1 is related to the first 11 amino acids of b_1 and to each of a_2–a_5. a_3 is also related to a_4. b_1 is related to b_2, b_3 and b_4.

each of a_1 and b_2 to b_1 are also about 10^{-4}, while b_3 and b_4 are more weakly related to b_1.

These repeats are longer and show a greater degree of internal consistency than those in other proteins, but matching probabilities as high as 10^{-5} are not very significant in a comparison of about 100,000 spans. Many of the repeats depend on local concentrations of alanine, valine and glycine residues. We have examined their structures on a model of subtilisin BPN′ built at Cambridge by Dr C. Wright.

Segments a_1 and a_2 are interesting because both contain histidines close to the active site (Wright *et al.*, 1969). Also, both histidines are near the beginning of two long helices (63–73) and (223–238) which rest upon one another and run parallel to one another through the centre of the molecule. On the other hand, segment b_1 is totally unlike a_1: it consists of the end of a helix (143–145) followed by a long internal piece of extended chain (146–154) which then runs along the surface (155–159) and leads into an irregular loop (160–164). It is also interesting that several of the other segments include long pieces of internal extended chain: a_3 (45–50), a_4 (26–32), a_5 (80–85) and (89–95), b_3 (89–95) and b_4 (137–180).

The statistical evidence for gene duplication is weak, since the probability that the tetrapeptide HVAG should be repeated somewhere in the sequence is greater than 10^{-2}. The structural similarity of the two helices (a_1 and a_2) and the pieces of extended chain is suggestive but not compelling. There is no sign of regularity in the molecular structure as a whole. Of the two long correlations noted by Haber & Koshland (1970), one (133–164 with 212–243) includes the relation between a_1 and b_1 above; the other (5–50 with 113–158) embraces no structurally similar regions and does not appear prominently in our comparison matrix.

8. Carboxypeptidase

The most interesting repeat in carboxypeptidase A is the tripeptide IHS at 68–70 and 195–197. Both histidines are zinc ligands at the active site (Bradshaw, Ericsson, Walsh & Neurath, 1969; Lipscomb *et al.*, 1968; Lipscomb, Reeke, Hartsuck, Quiocho & Bethge, 1970), and both occur at the ends of two parallel strands of extended chain which run from 61–68 and 190–196 through the core of the structure. There is, however, no continuing correspondence of either structure or sequence on either side of this pair of regions. In carboxypeptidase B the first histidine tripeptide becomes FHA (Bradshaw, Neurath & Walsh, 1969). The matching probability for the tripeptide IHS to be repeated is of the order of 10^{-3} and is not statistically significant. If gene duplication has occurred in this region it must have been overlaid by very extensive later changes in the three-dimensional structure. Neurath *et al.* (1969) have also proposed that a repeated sequence near arginine 145 has duplicated:

```
N R L W K T R S - - - V T S S S L C    (123–138)
N R N W D A G F G K A G A S S S P C    (144–161).
```

Here there is no discernible structural similarity in the two pieces, and the correlation is again not statistically significant. Finally there is a weak correlation linking two helical regions 18–26 and 99–107.

9. Papain

Papain (Drenth, Jansonius, Koekoek, Swen & Wolthers, 1968; Husain & Lowe, 1969; Light, Frater, Kimmel & Smith, 1964) contains a repeated tetrapeptide PVKN at 15–18 and 209–212 which occurs in both places as a piece of extended chain in the surface. There are also three interesting strongly related segments, with matching probabilities of about 10^{-5}:

G	I	I	K	I	R	T	G	N	L	N	Q	Y	(36–48)
G	Y	I	L	I	K	N	S	W	G				(169–178)
G	Y	I	R	I	K	R	G	T	G	N	S	Y	(185–197).

The structure of the first, which includes a section of the helical core (26–40) of the second lobe of the structure, is totally unlike the latter pair. These include two antiparallel strands of β structure (169–175 and 187–190) which are linked to one another in the surface of the first lobe.

10. Histones

The sequence of calf thymus histone IV has several short repeated sections (DeLange, Famborough, Smith & Bonner, 1969). The longest of these:

L	G	K	G	G	A	K	R	H	(10–18)
L	A	R	G	G	V	K	R	R	(37–44)

has a matching probability of about 10^{-4}, so it is not statistically significant. There is no evidence of any regular pattern of repeats. The lysine-rich histone (Iwai, Ishikawa & Hayashi, 1970) appears to be quite unrelated in sequence and contains a repeat (13–19 with 114–120) which is too short to be statistically significant.

11. The Significance of Repeats

The proteins which we have studied, including several others which do not merit special mention (tobacco mosaic virus coat, phage f2 coat, lysozyme T4, azurins, penicillinases, staphylococcal nuclease, tryptophan synthetase alpha and glyceraldehyde 3-phosphate dehydrogenases), do not yield a single example of long mutually consistent regular repeats which are strong enough to be statistically significant evidence for an ancestral repeating sequence. In those examples where a few short repeating segments are found to be well correlated, the structures of corresponding pieces are, as often as not, quite different. There is no sign yet that bacterial proteins, or ancient proteins such as the histones, are consistently more repetitive than proteins which have evolved more recently. The only two examples where a sequence repeat appears suggestive are the pair of central helices in subtilisin and the two central strands of β structure in carboxypeptidase. Here, however, the evidence is not statistically significant, and one has to rely on the fact that the structural core of each protein adjacent to the active site contains two similar segments which each carry a chemically active histidine.

None of the proteins studied shows a large-scale repeat in its structural organization, accompanied by a repeat in the sequence, which cannot be accounted for by chance.

There is, therefore, no necessity to postulate that gene duplication, either in the form of regular repeating sequences, or a mosaic of short repeated pieces of various kinds, has been a dominant influence in the recent evolution of proteins. If events of this kind did occur in the earliest stages of evolution of the most primitive proteins, their traces have been almost completely obscured by later changes in the sequence and structure.

Chromosomes of many higher animals (e.g. mice) contain large amounts of highly repetitive satellite DNA near the centromeres. This DNA contains repeat periods as short as six nucleotides (Southern, 1970) and does not code for any known protein. Even if satellite DNA is a source for the random evolution of new proteins, it can bear little relevance to the processes of evolution in bacteria.

It is worth emphasising that surprisingly strong short repeats arise fairly often by chance. They can easily appear significant when taken by themselves, especially if the three-dimensional structure of the protein is unknown. The case against the significance of repeats relies more on the careful examination of the mutual consistency of the observed correlations and their relationship with the folding of the protein than on statistical arguments. The early work of Cantor & Jukes (1966) and Fitch (1966) thus suggested possibilities that were of great interest and potential importance, which could not be dismissed on the evidence then available to them.

There is an apparent weakness in our arguments because it has been assumed throughout that two sections of protein sequence which share a common ancestor must fold similarly. At the same time many examples have been pointed out where similar short sequences have quite different structures. Why should there not be duplication of short pieces of sequences which then form different structures? To this there are two answers. One is that no statistically significant evidence for this kind of duplication event yet exists. The other is that ancestral similarities in sequence only persist for a long time during evolution if the structures to which they belong remain the same. Hence any weak repeat which does not correspond to a structural repeat is far less likely to be a long-standing ancestral feature than one which does.

The occurrence of very similar segments with quite different structures in so many proteins suggests that local sequence may be even less important in protein folding than is usually thought, and that the balance of molecular forces is exceedingly delicate.

12. Evolution of Large Proteins from Smaller Ones

During the course of evolution, proteins have tended to acquire successively more complicated and sophisticated functions. These new functions often make use of larger structures, either with longer chains or with several interacting subunits. Thus, many protein chains which now possess 200 to 400 amino acids have probably evolved from shorter proteins of 50 to 100 amino acids. There is probably also a critical length of about 50 amino acids below which it is difficult to form a protein structure which is stable in solution under normal conditions.

In the very earliest stages of evolution, after the setting up of the genetic code, the rate of error in protein synthesis would tend to control the length of a protein. Two extremes are conceivable: large imprecisely folded molecules which were relatively inefficient, but which could tolerate many sequence errors without losing activity; or shorter, precisely formed, and highly intolerant structures which could function very effectively provided they were free of error. For example, if we require that the protein synthesis apparatus must give a 90% yield of perfect sequences, the tolerable error rates per amino acid for chains of lengths 20, 100 and 400 are 0·002, 0·0004 and 0·0001, respectively. If the *proportion* of errors, rather than their total number, determined whether primitive proteins were acceptable, the argument for short proteins disappears, since long chains would have a better chance of forming stable structures.

In any case there are certainly a large number of proteins of the second, intolerant type which have increased in size during more recent evolution; and one can ask what processes may have led to their growth.

There are three principal possibilities.

(a) *Multiple repeats followed by consolidation of the whole*

A single short sequence of, say, 10 to 15 amino acids was exactly copied *n* times over at one time to produce a primitive repeating sequence having approximately the same length as the final protein (Eck & Dayhoff, 1966). The fundamental sequence was capable of forming some simple kind of stable structural unit—a helix or a loop of β structure—which was repeated several times. The units might then aggregate together to form a stable whole, forming a protein with a repetitive secondary structure, folded into a tertiary structure the chief features of which were determined once and for all. Proteins which happened to be able to bind other molecules might then later acquire some activity as an enzyme without undergoing large changes in tertiary structure. On this view the primitive repeating protein would have no biological function until after its over-all structure was decided.

(b) *Random duplication and consolidation of segments*

The initial short sequence, which has a primitive biological function, grows by adding a segment of 10 to 15 amino acids at long intervals of time. Each segment is a copy of some existing portion of the chain, and may be added anywhere in the sequence. After a piece is added there is a rapid series of small evolutionary changes to consolidate its position in the new structure. The protein would maintain some biological function which changes in small steps. A rarer process would be the duplication of an entire protein chain, to produce a molecule built out of two similar substructures in contact. This event could lead to a discontinuous change of function or a large improvement in biological activity. Two identical substructures tend to fit together well, because pieces which are of similar structure are more likely to fit one another in complementary fashion than are two dissimilar pieces (Monod, Wyman & Changeux, 1965).

The sequences and structures studied in this paper give no support at all to the idea of multiple repeats, and little, if any, to the idea of limited duplication. The theory of multiple repeats also has the drawback that it requires a large structure to evolve before it acquires any well-defined biological function. Thus we are led to a third possibility.

(c) *Piecemeal growth*

The protein would begin as a short chain with some biological activity, centred on an active site, and then alter its structure, gradually for the most part, inserting or deleting one amino acid at a time at points on the outer surface in such a way as to conserve the structural core round the active site. In this way successively more complicated surface loops and supporting structures could be added to strengthen the original molecular framework and improve the activity. During these changes the biological function would always exist, although it might change its character, as happens in going from chymotrypsin to thrombin (Magnusson, 1968,1970) or from lysozyme to α-lactalbumin (Brew & Campbell, 1967; Brew *et al.*, 1967; Browne *et al.*, 1969). The sequences of cytochrome *c* (Nolan & Margoliash, 1968) or of myoglobin, lamprey haemoglobin and mammalian haemoglobins (Dayhoff, 1969; Braunitzer & Fujiki, 1969) show that protein chains can easily lengthen or contract at either end and at corners. Elastase and thrombin yield examples where growth occurs at bends in the middle of a chain (Shotton & Hartley, 1970) while haemoglobin Gun Hill (Bradley, Wohl & Rieder, 1967) provides one where several deletions may have occurred at the same time. It is more difficult to imagine several simultaneous insertions, because the new amino acids are unlikely to fit well into the existing structure. Gene duplication would only occur as a rare and atypical event.

Two important features are common to many enzyme structures (Blow & Steitz, 1970): the presence of buried polar groups in the active site, and the way in which these polar groups are attached to different lobes or sub-assemblies of the structure which come together on different sides of the active centre. One reason for this, suggested by Blow & Steitz, is that the folding of the protein must supply sufficient free energy to compensate for the *electronic strain* energy which is used to abstract the polar groups from the surrounding solution: one simple way is to use the free energy of adhesion of several large rigid substructures. A second reason may be that small changes in the mutual packing of the substructures during evolution allow very precise adjustment of the positions of the active groups to optimise the catalytic activity. Figure 1 illustrates a typical scheme of gradual piecemeal evolution which embodies these features. Here the active site begins with three lobes which gradually grow and become more extensively interlinked. Later a fourth outer lobe is added with a new active group which modifies the catalytic function.

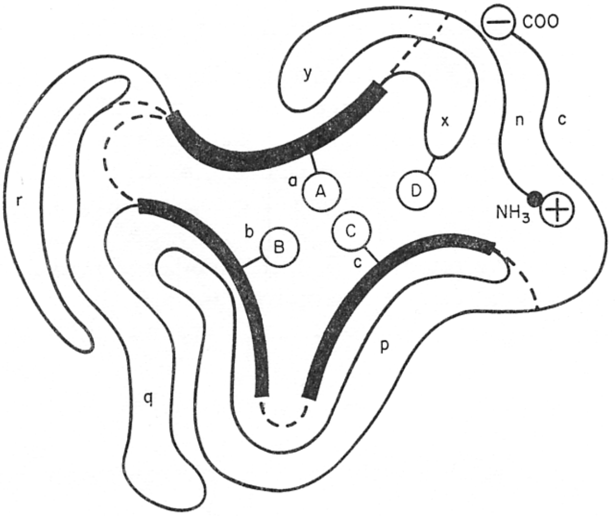

FIG. 1. Scheme for gradual growth of an enzyme chain. Initially the active site consists of three lobes, a, b and c, which support the active groups A, B, C. Later the closed loops p, q and r grow successively on the outer surface. A new loop x, later supported by the section y, introduces a new group D into the active centre, modifying the activity. The amino and carboxyl ends grow at n and c to link together the ends of the chain.

According to this type of scheme the final large structure gradually builds up from the interior in successive layers centred on the active site. The path by which the enzyme folds would evolve too, but if there is any orderly series of stages there should be a tendency for the folding of the parts which evolved earliest to guide the folding of the later and more peripheral sections. For example, in the final structure of the protein in Figure 1 the sections p, q and r could not achieve their final conformation until a, b and c were assembled. This idea of a hierarchy of folding events set up during evolution of a large chain, and each dominating the folding of successive portions of the structure, is related to the idea that a protein contains nucleation points about which it folds. However, it goes further, since it suggests a way in which a very large structure can "learn" to fold as it evolves, and suggests that distantly related proteins of similar structure but widely different chain lengths should have whole loops of peripheral structure inserted or deleted, while conserving other parts of the core almost unchanged.

I thank Mrs P. M. E. Altham, Dr F. H. C. Crick, Dr M. F. Perutz, Dr L. E. Orgel and Dr D. M. Blow for discussions.

APPENDIX

Substitution Frequencies in Proteins

Here are the data used in Figure 1 of the preceding paper (McLachlan, 1971). They are taken from 17 homologous families of proteins: 2 subtilisins, 13 haemoglobins, 8 cytochromes *c*, 2 penicillinases, 14 antibody light-chain variable regions, 12 antibody constant regions, 5 tobacco mosaic viruses, 6 azurins, 3 glyceraldehyde-3-phosphate dehydrogenases, 6 chymotrypsin enzymes, 2 cytochromes c_3, 3 cytochromes c_{551}, 4 lysozymes, 3 ribonucleases, 17 insulins, 3 bacterial ferredoxins, and 3 plant ferredoxins.

In each family we have counted $N(i,j)$, the number of *positions* at which amino acids i and j occur at least once as alternatives, and $n(i)$, the number of positions at which amino acid i ever occurs. The total number of substitutions for each amino acid is defined as:

$$N(i) = \sum N(i,j), \qquad j \neq i. \tag{A1}$$

If the substitutions were all equally probable, the expected values of $N(i,j)$ and $N(i)$ would be:

$$E(i,j) = \alpha n(i)n(j), \tag{A2}$$

$$E(i) = \sum E(i,j), \qquad j \neq i, \tag{A3}$$

where

$$\alpha = N_1/N_2;\ N_1 = \sum N(i,j),\ N_2 = \sum n(i)n(j),\ i \neq j. \tag{A4}$$

The relative frequencies of the substitutions are defined to be:

$$f(i,j) = N(i,j)/E(i,j) \quad \text{and} \quad f(i) = N(i)/E(i).$$

In Table A1 the diagonal elements are $N(i)$ and $f(i)$. The off-diagonal elements are $N(i,j)$ and $f(i,j)$. The row below the main body of the Table gives $n(i)$.

TABLE A1

Frequencies of amino-acid substitutions in proteins

AMINO ACID REPLACEMENTS IN 17 FAMILIES OF PROTEINS

	V	L	I	M	F	W	Y	G	A	P	S	T	C	H	R	K	Q	E	N	D
VAL	662	82	102	28	27	5	19	35	79	25	50	46	6	12	18	41	21	27	18	21
	0.94	1.74	2.77	1.66	1.19	0.64	0.81	0.67	1.09	0.84	0.68	0.85	0.46	0.67	0.71	0.73	0.66	0.72	0.43	0.47
LEU	82	545	72	44	43	7	18	18	43	12	38	35	2	11	17	34	21	12	18	18
	1.74	0.93	2.40	3.19	2.33	1.10	0.95	0.43	0.73	0.49	0.63	0.79	0.19	0.75	0.82	0.74	0.80	0.39	0.53	0.49
ILE	102	72	411	19	18	5	14	13	31	11	31	33	2	7	7	13	6	10	10	7
	2.77	2.40	0.89	1.77	1.25	1.01	0.94	0.39	0.67	0.58	0.66	0.96	0.24	0.61	0.43	0.36	0.30	0.42	0.38	0.25
MET	28	44	19	227	12	1	5	5	20	3	17	20	3	6	3	10	9	6	8	8
	1.66	3.19	1.77	1.04	1.82	0.44	0.73	0.33	0.95	0.34	0.79	1.26	0.79	1.15	0.40	0.61	0.96	0.55	0.66	0.61
PHE	27	43	18	12	249	10	49	2	17	5	18	12	0	10	5	6	1	4	4	6
	1.19	2.33	1.25	1.82	0.86	3.28	5.36	0.10	0.60	0.43	0.62	0.56	0.0	1.43	0.50	0.27	0.08	0.27	0.25	0.34
TRP	5	7	5	1	10	85	16	4	4	0	9	5	1	2	3	4	3	3	1	2
	0.64	1.10	1.01	0.44	3.28	0.84	5.09	0.57	0.41	0.0	0.91	0.68	0.57	0.83	0.88	0.53	0.70	0.59	0.18	0.33
TYR	19	18	14	5	49	16	247	6	13	3	24	13	2	10	7	13	5	9	11	10
	0.81	0.95	0.94	0.73	5.36	5.09	0.83	0.29	0.44	0.25	0.81	0.59	0.38	1.38	0.68	0.57	0.39	0.59	0.66	0.55
GLY	35	18	13	5	2	4	6	535	78	26	73	37	5	11	20	46	21	41	46	48
	0.67	0.43	0.39	0.33	0.10	0.57	0.29	0.84	1.20	0.97	1.10	0.76	0.43	0.68	0.88	0.91	0.73	1.22	1.24	1.20
ALA	79	43	31	20	17	4	13	78	878	52	135	87	6	19	24	68	43	68	44	47
	1.09	0.73	0.67	0.95	0.60	0.41	0.44	1.20	1.02	1.39	1.46	1.28	0.37	0.85	0.75	0.97	1.07	1.45	0.85	0.84
PRO	25	12	11	3	5	0	3	26	52	331	46	27	2	8	12	27	16	28	8	20
	0.84	0.49	0.58	0.34	0.43	0.0	0.25	0.97	1.39	0.88	1.21	0.96	0.30	0.87	0.92	0.93	0.97	1.45	0.37	0.87
SER	50	38	31	17	18	9	24	73	135	46	1039	144	12	25	48	87	53	62	98	69
	0.68	0.63	0.66	0.79	0.62	0.91	0.81	1.10	1.46	1.21	1.18	2.08	0.72	1.09	1.48	1.22	1.30	1.30	1.86	1.21
THR	46	35	33	20	12	5	13	37	87	27	144	751	8	22	25	67	37	47	43	43
	0.85	0.79	0.96	1.26	0.56	0.68	0.59	0.76	1.28	0.96	2.08	1.13	0.65	1.31	1.05	1.27	1.23	1.33	1.10	1.02
CYS	6	2	2	3	0	1	2	5	6	2	12	8	68	3	3	2	1	1	4	5
	0.46	0.19	0.24	0.79	0.0	0.57	0.38	0.43	0.37	0.30	0.72	0.65	0.40	0.74	0.52	0.16	0.14	0.12	0.43	0.50
HIS	12	11	7	6	10	2	10	11	19	8	25	22	3	249	18	25	14	8	20	18
	0.67	0.75	0.61	1.15	1.43	0.83	1.38	0.68	0.85	0.87	1.09	1.31	0.74	1.08	2.29	1.44	1.41	0.69	1.56	1.30
ARG	18	17	7	3	5	3	7	20	24	12	48	25	3	18	359	65	30	20	21	13
	0.71	0.82	0.43	0.40	0.50	0.88	0.68	0.88	0.75	0.92	1.48	1.05	0.52	2.29	1.11	2.64	2.13	1.21	1.15	0.66
LYS	41	34	13	10	6	4	13	46	68	27	87	67	2	25	65	701	46	52	53	42
	0.73	0.74	0.36	0.61	0.27	0.53	0.57	0.91	0.97	0.93	1.22	1.27	0.16	1.44	2.64	1.02	1.48	1.43	1.32	0.97
GLN	21	21	6	9	1	3	5	21	43	16	53	37	1	14	30	46	450	56	33	34
	0.66	0.80	0.30	0.96	0.08	0.70	0.39	0.73	1.07	0.97	1.30	1.23	0.14	1.41	2.13	1.48	1.11	2.70	1.44	1.37
GLU	27	12	10	6	4	3	9	41	68	28	62	47	1	8	20	52	56	568	35	79
	0.72	0.39	0.42	0.55	0.27	0.59	0.59	1.22	1.45	1.45	1.30	1.33	0.12	0.69	1.21	1.43	2.70	1.21	1.30	2.73
ASN	18	18	10	8	4	1	11	46	44	8	98	43	4	20	21	53	33	35	534	59
	0.43	0.53	0.38	0.66	0.25	0.18	0.66	1.24	0.85	0.37	1.86	1.10	0.43	1.56	1.15	1.32	1.44	1.30	1.03	1.84
ASP	21	18	7	8	6	2	10	48	47	20	69	43	5	18	13	42	34	79	59	549
	0.47	0.49	0.25	0.61	0.34	0.33	0.55	1.20	0.84	0.87	1.21	1.02	0.50	1.30	0.66	0.97	1.37	2.73	1.84	0.99
POSN	394	321	250	115	154	53	159	353	493	203	502	370	89	122	173	382	218	255	282	304

N1 = 9438 N2 = 25290138 ALPHA = 0.00037

REFERENCES

Air, G. M., Thompson, E. O. P., Richardson, B. J. & Sharman, G. B. (1971). *Nature*, **229**, 391.

Ambler, R. P. (1968). *Biochem. J.* **109**, 47.

Beintema, J. J. & Gruber, M. (1967). *Biochim. biophys. Acta*, **147**, 612.

Benson, A. M., Mower, H. F. & Yasunobu, K. T. (1967). *Arch. Biochem. Biophys.* **121**, 563.

Birktoft, J. J., Blow, D. M., Henderson, R. & Steitz, T. A. (1970). *Phil. Trans. Roy. Soc. Lond.* B, **257**, 67.

Black, J. A. & Dixon, G. H. (1967). *Nature*, **216**, 152.

Black, J. A. & Dixon, G. H. (1968). *Nature*, **218**, 736.

Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *Proc. Roy. Soc. Lond.* B, **167**, 365.

Blow, D. M., Birktoft, J. J. & Hartley, B. S. (1969). *Nature*, **221**, 337.

Blow, D. M. & Steitz, T. A. (1970). *Ann. Rev. Protein Chem.* **39**, 63.

Bradley, T. B., Wohl, R. C. & Rieder, R. F. (1967). *Science*, **157**, 1581.

Bradshaw, R. A., Ericsson, L. H., Walsh, K. A. & Neurath, H. (1969). *Proc. Nat. Acad. Sci., Wash.* **63**, 1389.

Bradshaw, R. A., Neurath, H., Tye, R. W., Walsh, K. A. & Winter, W. P. (1970). *Nature*, **226**, 237.

Bradshaw, R. A., Neurath, H. & Walsh, K. A. (1969). *Proc. Nat. Acad. Sci., Wash.* **63**, 406.

Braunitzer, G. & Fujiki, H. (1969). *Naturwiss.* **56**, 322.

Brew, K. & Campbell, P. N. (1967). *Biochem. J.* **102**, 258.

Brew, K., Vanaman, T. C. & Hill, R. L. (1967). *J. Biol. Chem.* **242**, 3747.

Bridges, C. B. (1936). *Science*, **83**, 210.

Brown, J. R. & Hartley, B. S. (1966). *Biochem. J.* **101**, 214, 229.

Browne, W. J., North, A. C. T., Phillips, D. C., Brew, K., Vanaman, T. C. & Hill, R. L. (1969). *J. Mol. Biol.* **42**, 65.

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

Cantor, C. & Jukes, T. H. (1966). *Proc. Nat. Acad. Sci., Wash.* **56**, 172.

Cohen, S. & Milstein, C. (1967*a*). *Advanc. Immunology*, **7**, 1.

Cohen, S. & Milstein, C. (1967*b*). *Nature*, **214**, 449.

Crick, F. H. C. (1968). *J. Mol. Biol.* **38**, 367.

Dayhoff, M. O. (1969). *Atlas of Protein Sequence and Structure 1969*. Silver Spring, Maryland: National Biochemical Research Foundation.

DeLange, R. J., Famborough, D. M., Smith, E. L. & Bonner, J. (1969). *J. Biol. Chem.* **244**, 319.

Dixon, G. H. (1966). *Essays in Biochemistry,* vol. 2, ed. by P. N. Campbell & G. D. Greville. New York: Academic Press.

Drenth, J., Jansonius, J. N., Koekoek, R., Swen, H. M. & Wolthers, B. G. (1968). *Nature*, **218**, 929.

Dus, K., Sletten, K. & Kamen, M. (1968). *J. Biol. Chem.* **243**, 5507.

Eck, R. V. & Dayhoff, M. O. (1966). *Science*, **152**, 363.

Edelman, G. M., Cunningham, B. A., Gall, W. E., Gottlieb, P. D., Rutishauser, U. & Waxdal, M. J. (1969). *Proc. Nat. Acad. Sci., Wash.* **63**, 78.

Fitch, W. M. (1966). *J. Mol. Biol.* **16**, 1, 8, 17.

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.

Huber, R., Formanek, H. & Epp, O. (1968). *Naturwiss.* **2**, 75.

Husain, S. S. & Lowe, G. (1969). *Biochem. J.* **114**, 279.

Ingram, V. M. & Stretton, A. O. W. (1961). *Nature*, **190**, 1079.

Iwai, K., Ishikawa, K. & Hayashi, H. (1970). *Nature*, **226**, 1057.

Jollès, P. (1969). *Angewandte Chemie* (Internat. edn.), **8**, 227.

Jukes, T. H. (1966). *Molecules and Evolution*. New York: Columbia University Press.

Kartha, G., Bello, J. & Harker, D. (1967). *Nature*, **213**, 862.

Light, A., Frater, R., Kimmel, J. R. & Smith, E. L. (1964). *Proc. Nat. Acad. Sci., Wash.* **52**, 1276.

Lipscomb, W. N., Hartsuck, J. A., Reeke, G. N., Quiocho, F. A., Bethge, P. H., Ludwig, M., Steitz, T. A., Muirhead, H. & Coppola, J. C. (1968). *Structure, Function and Evolution in Proteins*, Brookhaven Symposia in Biology. **21**, 23.

Lipscomb, W. N., Reeke, G. N., Hartsuck, J. A., Quiocho, F. A. & Bethge, P. H. (1970). *Phil. Trans. Roy. Soc. Lond.* B, **257**, 177.

Low, B. W., Lovell, F. M. & Rudko, A. D. (1968). *Proc. Nat. Acad. Sci., Wash.* **60**, 1515.

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

Magnusson, S. (1968). *Biochem. J.* **110**, 25.

Magnusson, S. (1970). In *Structure–Function Relationships of Proteolytic Enzymes*, p. 138, ed. by P. Desnuelle, H. Neurath & M. Ottesen. Copenhagen: Munksgaard.

Manwell, J. (1967). *J. Comp. Biochem. Physiol.* **23**, 383.

Markland, F. S. & Smith, E. L. (1967). *J. Biol. Chem.* **242**, 5198.

Matthews, B. W., Sigler, P. B., Henderson, R. & Blow, D. M. (1967). *Nature*, **214**, 652.

Milstein, C. & Pink, J. R. L. (1970). *Progress in Biophysics & Molecular Biology*, **21**, 209, ed. by J. A. V. Butler & D. Noble. Oxford: Pergamon Press.

Monod, J., Wyman, J. & Changeux, J. P. (1965). *J. Mol. Biol.* **12**, 88.

Needleman, S. B. & Blair, T. T. (1969). *Proc. Nat. Acad. Sci., Wash.* **63**, 1227.

Neurath, H., Bradshaw, R. A. & Arnon, R. (1969). *International Symposium on Structure–Function Relationships of Proteolytic Enzymes*. Copenhagen: Munksgaard.

Nolan, C. & Margoliash, E. (1968). *Ann. Rev. Biochem.* **37**, 727.

Padlan, E. A. & Love, W. E. (1968). *Nature*, **220**, 376.

Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). *J. Mol. Biol.* **13**, 669.

Perutz, M. F., Muirhead, H., Cox, J. M. & Goaman, L. G. C. (1968). *Nature*, **219**, 139.

Russell, R. L., Abelson, J. N., Landy, A., Gefter, M. L., Brenner, S. & Smith, J. D. (1970). *J. Mol. Biol.* **47**, 1.

Shotton, D. M. & Hartley, B. S. (1970). *Nature*, **225**, 802.

Shotton, D. M. & Watson, H. C. (1970). *Nature*, **225**, 811.

Singer, S. J. & Doolittle, R. F. (1966). *Science*, **153**, 13.

Smith, E. L., DeLange, R. J., Evans, W. H., Landon, M. & Markland, F. S. (1968). *J. Biol. Chem.* **243**, 2184.

Smithies, O., Connell, G. E. & Dixon, G. H. (1962). *Nature*, **196**, 232.

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

Sneath, P. H. A. (1966). *J. Theoret. Biol.* **12**, 157.

Šorm, F. & Keil, B. (1962). *Advanc. Protein Chem.* **17**, 167.

Šorm, F. & Knichal, V. (1958). *Collection Czech. Chem. Commun.* **23**, 1575.

Southern, E. M. (1970). *Nature*, **227**, 794.

Stephens, S. G. (1951). *Advanc. Genetics*, **4**, 247.

Takahashi, K. (1965). *J. Biol. Chem.* **240**, pc 4117.

Tanaka, M., Nakashima, T., Benson, A., Mower, H. & Yasunobu, K. T. (1966). *Biochemistry*, **5**, 1666.

Thiebaux, H. J. & Pattee, H. H. (1967). *J. Theoret. Biol.* **17**, 121.

Tsugita, A. & Inouye, M. (1968). *J. Mol. Biol.* **37**, 201.

Urbain, J. (1969). *Biochemical Genetics*, **3**, 249.

Watson, H. C., Shotton, D. M., Cox, J. M. & Muirhead, H. (1970). *Nature*, **225**, 806.

Williams, J., Clegg, J. B. & Mutch, M. O. (1961). *J. Mol. Biol.* **3**, 532.

Wright, C. S., Alden, R. A. & Kraut, J. (1969). *Nature*, **221**, 235.

Wyckoff, H. W., Hardman, K. D., Allewell, N. M., Inagami, T., Johnson, L. N. & Richards, F. M. (1967). *J. Biol. Chem.* **242**, 3749.

Wyckoff, H. W., Tsernoglu, D., Hanson, A. W., Knox, J. R., Lee, B. & Richards, F. M. (1970). *J. Biol. Chem.* **245**, 305.