J. Mol. Evol. 12, 219–236 (1979)

Journal of Molecular Evolution

Two Types of Amino Acid Substitutions in Protein Evolution

Takashi Miyata, Sanzo Miyazawa, and Teruo Yasunaga

Department of Biology, Kyushu University, Fukuoka 812, Japan

Summary. The frequency of amino acid substitutions, relative to the frequency expected by chance, decreases linearly with the increase in physico-chemical differences between amino acid pairs involved in a substitution. This correlation does not apply to abnormal human hemoglobins. Since abnormal hemoglobins mostly reflect the process of mutation rather than selection, the correlation manifest during protein evolution between substitution frequency and physico-chemical difference in amino acids can be attributed to natural selection. Outside of 'abnormal' proteins, the correlation also does not apply to certain regions of proteins characterized by rapid rates of substitution. In these cases again, except for the largest physico-chemical differences between amino acid pairs, the substitution frequencies seem to be independent of the physico-chemical parameters. The limination of the substituents involving the largest physico-chemical differences can once more be attributed to natural selection. For smaller physico-chemical differences, natural selection, if it is operating in the polypeptide regions, must be based on parameters other than those examined.

Key words: Amino acid substitution – Physico-chemical difference – Conservative – Low-constraint – Protein evolution

Introduction

Hitherto, several workers have suggested that during the evolution of proteins amino acid substitutions producing relatively little physico-chemical changes are much more frequent than those involving relatively large changes, i.e., the amino acid substitutions are conservative (Zuckerkandl and Pauling, 1965; Epstein, 1967; Clarke, 1970; McLachlan, 1971; Dayhoff et al., 1972 a; Grantham, 1974; Hasegawa and Yano, 1975). By analysing the relationship between the frequencies of particular substitutions and the physico-chemical properties of the amino acids, Epstein (1967) and Clarke (1970) have inferred that natural selection has acted in the evolution of proteins to favour the substitutions that would be compatible with the retention of the existing conformation

0022-2844/79/0012/0219/$ 03.60

of the proteins. Dayhoff et al. (1972a) have examined the patterns in the rate of substitutions relative to the rate expected by chance, by using the data derived from the substitutions observed in many families of proteins, and they have pointed out that amino acids fall into groups and subgroups whose members have similar substitutional characteristic and these groupings are understandable on the basis of the physico-chemical properties of the amino acids. Grantham (1974) has presented a formula for physico-chemical differences between changing residues, by combining the three properties of the amino acids, i.e., polarity, volume and composition, and has shown quantitatively that the difference values calculated from this formula correlate best with the substitution frequencies of amino acids observed in many proteins.

The amino acid replacements in abnormal hemoglobins permit close examination of the effect of mutation upon the structure and function of protein molecules (Perutz and Lehmann, 1968). The amino acid replacements observed in abnormal hemoglobins are not the substitutions that have spread to fixation in a population during the course of evolution, but are only the result of mutations which have not yet been eliminated from the population. It seems appropriate, therefore, to examine the pattern of amino acid changes observed in abnormal hemoglobins in order to investigate the extent of the structural and functional disruptions due to amino acid replacements and to detect any possible influences of natural selection on proteins. In this report, we compare the pattern of amino acid replacements in conservative substitutions with that of amino acid replacements observed in abnormal hemoglobins quantitatively and show that natural selection operates to favour those amino acid substitutions that tend to maintain the original conformation of proteins through conservative substitution.

The pattern of the amino acid substitutions studied so far has been concerned with those that are observed in the proteins which have been preserving their existing structures throughout the course of evolution. In order to investigate the pattern of substitutions observed in the protein regions in which fewer structural constraints are imposed or the structure is diverging, we extend our analysis to the substitutions which arise in the hypervariable regions of immunoglobulin variable domains and fibrinopeptides and insulin c-peptides. From this analysis, we can show that there exists another type of substitution which depends less on the extent of physico-chemical properties of substituted amino acids. Hereafter, we will call this type of substitution low-constraint substitution, distinguishing it from conservative substitution.

Amino Acid Pair Distance

For the quantitative analysis of the substitution pattern of amino acids, several quantities for the objective classification of the physico-chemical properties of amino acids have been proposed (Sneath, 1966; Epstein, 1967; Clarke, 1970; Grantham, 1974; Goodman and Moore, 1977). On the basis of the three physico-chemical characters, polarity, volume and composition, Grantham (1974) has introduced the amino acid pair differences, d_{ij}, indicating the degree of the difference of physico-chemical properties of the amino acids a_i and a_j. The three dimensional conformation of proteins is mainly determined by weak interactions between the amino acid residues. These interactions are determined by some of the properties the amino acids have, such as the ability to form hydrophobic interactions, hydrogen bonds, van der Waals interactions and salt bridges. These specific properties of the amino acids may be represented mainly

by the two physico-chemical factors, volume and polarity. That is, the polarity and volume are the main representatives of the specific properties the amino acids have and they are the primary determinants for the three dimensional conformation of protein molecules. We therefore define the amino acid pair distance, d_{ij}, indicating the degree of the physico-chemical properties of the amino acids a_i and a_j, only by polarity p_i and volume v_i for simplicity as follows:

$$(1)\quad d_{ij} = \sqrt{(\triangle p_{ij}/\sigma_p)^2 + (\triangle v_{ij}/\sigma_v)^2}$$

where $\triangle p_{ij}$ and $\triangle v_{ij}$ represent polarity and volume difference induced as the result of amino acid substitution respectively, (i.e., $\triangle p_{ij} = |p_i - p_j|$ and $\triangle v_{ij} = |v_i - v_j|$), and σ_p and σ_v are standard deviations of $\triangle p_{ij}$ and $\triangle v_{ij}$ respectively. Our definition is somewhat different from Grantham's in that the polarity and volume differences are divided by their standard deviations respectively so as to fit the scale of polarity and volume differences to each other. The values for polarity and volume are from the data of Grantham (1974).

According to Grantham (1974), it is important to include the third parameter, composition, in his formula for obtaining the best correlation with observed substitution frequency. In comparing the physico-chemical difference with the substitution frequency obtained by McLachlan (1971), the amino acid pairs resulting from more than two steps were also included in his analysis. By a one step pair we mean an amino acid pair one member of which is changeable into the other by one base substitution in the sense of minimum base mutation (Fitch, 1966). Presumably this may be the main reason why he failed to obtain the good correlation by only the two parameters, polarity, and volume. According to his argument, the substitutions are expected to be observed with the same frequency for amino acid pairs whose differences are the same. But, in general, the substitutions between amino acid pairs of one step must be observed more frequently than those of two- or three-step pairs in a relatively short evolutionary time interval, even if they have the same physico-chemical differences. Therefore, the comparison must be made separately for amino acid pairs of one step and those of more than two steps. As we can see below, the amino acid difference based only on polarity and volume correlates well with observed substitution frequency in one-step pairs.

Table 1 shows the pair distance, d, calculated by the above formula. The distance d ranges from the value 0.06 of the most similar pair, alanine and proline, to 5.13 of the most disimilar pair, glycine and tryptophan. All the amino acids are classified into well defined six groups by the distance. The classification is: amino acids in group 2 are small and have neutral polarity, group 3A amino acids are hydrophilic and relatively small, group 3B amino acids are hydrophilic and relatively large, group 4A amino acids are hydrophobic and relatively small, group 4B amino acids are hydrophobic and relatively large, and group 1 amino acid is special. Each pair within a group has a distance less than unity and mean distances between groups always exceed unity. Table 2 shows the classification of amino acids and mean distances within and between groups. This classification of amino acids will not be used in the subsequent analysis, but is shown only to illustrate the correspondence between the amino acid pair and the distance.

Table 1. Amino acid pair distance

CYS	PRO	ALA	GLY	SER	THR	GLN	GLU	ASN	ASP	HIS	LYS	ARG	VAL	LEU	ILE	MET	PHE	TYR	TRP	
	1.33	1.39	2.22	1.84	1.45	2.48	3.26	2.83	3.48	2.56	3.27	3.06	0.86	1.65	1.63	1.46	2.24	2.38	3.34	CYS
		0.06	0.97	0.56	0.87	1.92	2.48	1.80	2.40	2.15	2.94	2.90	1.79	2.70	2.62	2.36	3.17	3.12	4.17	PRO
			0.91	0.51	0.90	1.92	2.46	1.78	2.37	2.17	2.96	2.92	1.85	2.76	2.69	2.42	3.23	3.18	4.23	ALA
				0.85	1.70	2.48	2.78	1.96	2.37	2.78	3.54	3.58	2.76	3.67	3.60	3.34	4.14	4.08	5.13	GLY
					0.89	1.65	2.06	1.31	1.87	1.94	2.71	2.74	2.15	3.04	2.95	2.67	3.45	3.33	4.38	SER
						1.12	1.83	1.40	2.05	1.32	2.10	2.03	1.42	2.25	2.14	1.86	2.60	2.45	3.50	THR
							0.84	0.99	1.47	0.32	1.06	1.13	2.13	2.70	2.57	2.30	2.81	2.48	3.42	GLN
								0.85	0.90	0.96	1.14	1.45	2.97	3.53	3.39	3.13	3.59	3.22	4.08	GLU
									0.65	1.29	1.84	2.04	2.76	3.49	3.37	3.08	3.70	3.42	4.39	ASN
										1.72	2.05	2.34	3.40	4.10	3.98	3.69	4.27	3.95	4.88	ASP
											0.79	0.82	2.11	2.59	2.45	2.19	2.63	2.27	3.16	HIS
												0.40	2.70	2.98	2.84	2.63	2.85	2.42	3.11	LYS
													2.43	2.62	2.49	2.29	2.47	2.02	2.72	ARG
														0.91	0.85	0.62	1.43	1.52	2.51	VAL
															0.14	0.41	0.63	0.94	1.73	LEU
																0.29	0.61	0.86	1.72	ILE
																	0.82	0.93	1.89	MET
																		0.48	1.11	PHE
																			1.06	TYR
																				TRP

Table 2. Average pair distance within and between groups. Amino acid pairs more than two step are excluded in this calculation

			1	2	3A	3B	4A	4B
Group 1	Special	Cys		2.03		3.06		2.65
Group 2	Neutral, Small	Pro Ala Gly Ser Thr	2.03	0.69	2.09	2.59	2.47	4.07
Group 3A	Hydrophilic, Relatively small	Gln Glu Asn Asp		2.09	0.80	1.21	3.11	3.68
Group 3B	Hydrophilic, Relatively large	His Lys Arg	3.06	2.59	1.21	0.61	2.58	2.49
Group 4A	Hydrophobic, Relatively small	Val Leu Ile Met		2.47	3.11	2.58	0.54	1.10
Group 4B	Hydrophobic, Relatively large	Phe Tyr Trp	2.65	4.07	3.68	2.49	1.10	0.48

Coarse Grained Relative Rate of Substitution

In order to investigate how observed substitution frequencies correlate with physico - chemical differences between changing residues, we consider the rate of substitution between amino acids a_i and a_j relative to the rate expected by chance from the frequency of occurrence of the amino acids. The expression for relative rate of substitution between a_i and a_j is

$$(2)\quad R_{ij} = A_{ij}/N^{obs}(f_i f_j/\Sigma_{[I]} f_i f_j)$$

Where A_{ij} is an accepted point mutation matrix element defined by Dayhoff et al. (1972 a) and f_i is the amino acid frequency of a_i. N^{obs} is the total number of substituted amino acids, i.e.,

$$\Sigma_{[I]} A_{ij}$$

and [I] means a set of all the one-step pairs. Following Dayhoff et al. (1972 a), the matrix A is derived from the comparison between closely related sequences, including contemporary and inferred ancestral sequences.

In the present analysis, amino acid pairs involving more than two-step changes are ignored. Almost all the amino acid changes arise as the result of a one-step change in a relatively short time span as we can see in the point mutations of abnormal hemoglobin variants. Two-step changes result from the consecutive one-step changes. Let us

consider a case in which an amino acid a_i changes to a_j through another amino acid a_α (i.e., $a_i \rightarrow a_\alpha \rightarrow a_j$, where both the pairs (a_i, a_α), (a_α, a_j) are one-step pairs), and that the physico-chemical properties of the pair (a_i, a_j) are similar, though those of (a_i, a_α) and (a_α, a_j) are disimilar. In this case, the high frequency of the substitutions between a_i and a_j may not be expected to be observed, even if they have similar physico-chemical properties. Then, in order to obtain the exact correlation between the observed substitution frequency and the physico-chemical difference between changing residues, the accepted point mutation matrix must be constructed using closely related sequences, only, in which most of the substitutions correspond to one-step changes.

Instead of the direct use of R_{ij} in equation (2), we use R(d) the coarse grained version of R_{ij}, that is, the average value of R_{ij} in the interval of distances d and $d+\triangle$:

$$(3)\quad R(d) = \sum_{d \leq d_{ij} < d+\triangle} R_{ij} \,/\, N(d)$$

where N(d) is the number of one-step pairs expected from the genetic code in the interval of distances d and $d+\triangle$. Hereafter we will call this R(d) as 'coarse grained relative rate of substitution' (RRS).

When the substitution patters of various protein families are compared by means of RRS defined above, the values of R(d) may show appreciable deviation around the mean from protein to protein because of the small sample sizes. To avoid this fluctuation, it may be more suitable to leave out the direct comparison by the values of R(d) and to compare the number of pairs in the interval of distances d and $d+\triangle$, whose relative rate of substitutions exceeds some threshold value. This procedure may correspond to a more coarse analysis of the substitution pattern than the analysis by RRS.

There is another reason to analyse the substitution pattern of each protein family, not by the use of R(d), but by means of the number of pairs for which the substitutions are observed frequently. Even in the substitution data which were accumulated from inferred ancestral sequences and closely related sequences, there are still non-zero matrix elements corresponding to two-step pairs. This indicates that, even in one-step pairs, substitutions would occur in an indirect way in which one amino acid is exchanged for the other through successive one-step replacements (i.e., $a_i \rightarrow a_\alpha \rightarrow a_j$, where the amino acid pair (a_i, a_j) is a one-step pair and a_α is an intermediate amino acid). These indirect substitutions would occur with low frequencies compared with direct one-step substitutions (i.e., an amino acid a_i changes to a_j in a direct way without going through an intermediate amino acid a_α). Both the direct and indirect substitutions contribute to the relative rate of substitution R_{ij} (i.e., $R_{ij} = R_{ij}$ (direct) + R_{ij} (indirect)). For a pair (a_i, a_j) with small d_{ij}, R_{ij} (indirect) may be negligibly small compared with R_{ij} (direct) in a short time span. A serious case may arise when we consider a dissimilar amino acid pair (a_k, a_l) with a distance d_{kl} so large that no direct substitution occurs between a_k and a_l. Nevertheless, the substitutions between them may be possible through successive one-step changes between similar amino acid pairs. In this case, the main contributor to R_{kl} must be the indirect substitutions. In any case, as the indirect substitutions may be observed with low frequencies so long as we use the data accumulated from closely related sequences, the indirect effect may be excluded by introducing an appropriate threshold value for R_{ij}. The elimination of the indirect effect is particularly important

when we compare the substitution patterns of different protein families whose substitution data are accumulated from sequences in different time intervals, or when we analyse the substitution pattern by using distantly related sequences.

Therefore, for the comparison of the substitution patterns in various protein families, it is more appropriate to use the ratio, F(d), of the number of pairs whose substitution frequencies exceed some threshold value relative to the number of one-step pairs expected from the genetic code in the interval of distances d and d+$\triangle$. Here, we tentatively use the mean value $\bar{R}$ of all the R_{ij} of one-step pairs as a threshold value (i.e.,

$$\bar{R} = \Sigma R_{ij} / N \quad [I]$$

where N is the total number of one-step pairs expected from the genetic code and [I] means all one-step pairs). Then, F(d) is expressed as follows:

$$(4)\quad F(d) = N[(i, j); R_{ij} > \bar{R},\ d \leq d_{ij} < d+\triangle] / N(d)$$

where $N[(i, j); R_{ij} > \bar{R},\ d \leq d_{ij} < d+\triangle]$ is the number of pairs whose R_{ij} exceeds $\bar{R}$ and also d_{ij} is in the interval d and d+$\triangle$. The value of 0.5 is always fixed for the interval of distance, $\triangle$, throughout this analysis. For example, there are 16 one-step pairs in the interval of distances 0.5 and 1.0, of which 12 pairs have the relative rate of substitutions [R_{ij}] over the mean $\bar{R}$ (=0.97) in the data of Dayhoff et al. (1972 a), then, we have F(0.5) = 12/16 = 0.75. The use of F(d) seems to be suitable particularly for the analysis in which distantly related protein sequences are used (see Fig. 6). Hereafter we will call this F(d) as 'relative number of frequent substitution' (RNFS).

Results

Dayhoff et al. (1972 a) have calculated the matrix, A, of accepted point mutations ac; cumulated from closely related sequences of protein families such as cytochrome c, hemoglobin, myoglobin, virus coat potein, chymotrypsinogen, glyceraldehyde 3-phosphate dehydrogenase, clupeine, insulin and ferredoxin. This matrix is suitable for the present purpose, but it includes non-zero matrix elements which correspond to two- and three-step mutations. These matrix elements are omitted in this analysis. Dayhoff et. al. (1972 a) have pointed out that the matrix of accepted point mutations computed separately from each family reflects the same pattern as the accumulated matrix, within the variation expected from the smaller sample size. This accumulated matrix may therefore be regarded as the matrix of an 'average protein'.

Figure 1 shows results of RRS and RNFS, calculated from equation (3) and equation (4) using this matrix. There is an obvious negative correlation both between R(d) and d and between F(d) and d, indicating that, as has already been pointed out by several workers, the substitutions which involve relatively small physico-chemical changes are much more frequent than those involving relatively large changes. The negative correlations between R(d) and d and between F(d) and d are highly significant. For R(d), the correlation coefficient is equal to -0.966 and the regression of R(d) on d is R(d) = -0.8d + 2.44 for $d < 3.5$. For F(d), the correlation coefficient is -0.990 and the regression is F(d) = -0.39d + 1.07 for $d < 3.0$.

From these results, we may say that the substitutions preferentially occur between members of pairs having more similar physico-chemical properties. In order to confirm

that this assures that the conformation of the protein departs as little as possible from the existing structure and is a result of natural selection, it is appropriate to analyse the pattern of amino acid changes observed in unstable abnormal and abnormal human hemoglobins.

Unstable abnormal hemoglobins offer a good opportunity for investigating how physico-chemical differences between changing residues affect the tertiary structure of the protein. Unstable abnormal hemoglobins are a group of abnormal hemoglobins that pro-

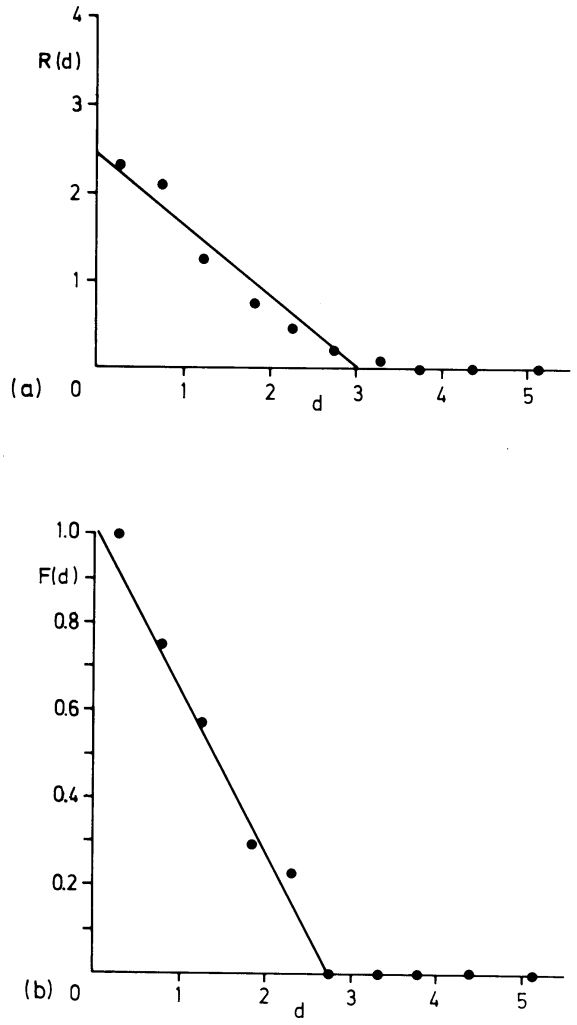

Fig. 1. **(a)** Relation between distance (d) and RRS (R(d)). The distance d stands for the physico-chemical difference between changing residues (see eq. 1), and R(d) stands for the coarse grained relative rate of substitutions in the closely related sequences accumulated from homologous protein families (see eq. 3). The value of d for each set of plots is the average distance of pairs in d and d+0.5. Straight line represents the regression of R(d) and d, i.e., $R(d) = -0.80d + 2.44$ for $0 < d < 3.5$. Correlation coefficient is -0.966. Most of the observed substitutions were for pairs of similar amino acids with a small value of d, and no substitutions were found where d was greater than 3.5 for this data. **(b)** Relation between distance and RNFS (F(d)), where F(d) stands for the ratio of the number of one-step pairs for which substitutions are observed frequently to the number of one-step pairs expected from the genetic code in the interval of d and d+0.5 (see eq. 4). Straight line is the regression of F(d) on d, i.e., $F(d) = -0.39d + 1.07$ for $0 < d < 3.0$. Correlation coefficient is -0.990. Data for (a) and (b) are from Dayhoff et al. (1972a)

duce a characteristic hemolytic anemia. The single amino acid substitution in unstable abnormal hemoglobins results in an instability of the hemoglobin molecules. In unstable abnormal hemoglobins, the change of one particular amino acid has a well-defined pathological consequence at the molecular level, and the clinical and hematological features are proportional to the severity of the structural lesion of the molecule (Lehmann and Huntsman, 1974). Hayashi (1975) has classified unstable abnormal hemoglobins into four groups according to the degree of severity of the hemolytic anemia. Following Hayashi's classification, group 1 and group 2 are groups of unstable abnormal hemoglobins where hemoglobinopathy is mild (mild group) and group 3 and group 4 are those where hemoglobinopathy is severe (severe group). Figure 2 shows the distributions of the physico-chemical differences, i.e., distance, between changing residues in the unstable abnormal hemoglobins, separately according to the severity of the hemolytic anemia. The broken line represents the distribution expected from the genetic code if a single amino acid change occurs randomly in human hemoglobins. The average amino acid contents of α- and β-chain of normal human hemoglobin are used for obtaining the random distribution.

The mean distances of the distributions for the mild group and the severe group and of the random distribution expected from the genetic code are 1.85, 2.67 and 1.80 respectively. As a measure of the extent of deviation from the random distribution, χ^2 values have been calculated for the mild group and the severe group, which are 11.3 and 19.7 respectively. The deviation from the random distribution is significant in the severe group ($\chi^2=19.7$, df=9, $p<0.05$). Judging from these results for the mean distance and χ^2 value, the more the extent of the severity increases, the more the distribution deviates towards a greater distance. This means that the extent of hematologic severity is positively correlated with the physico-chemical difference induced by the change of one particular amino acid. Since the hematologic features are proportional to the severity of structural disruption of the molecule in unstable abnormal hemoglobins, we may infer that the extent of conformational change caused by single amino acid change is positively correlated with physico-chemical difference between changing residues.

Abnormal human hemoglobins seem to be a direct reflection of the base mutations in the hemoglobin genes, except a few of them, such as HbS, which are selected for (Allison, 1954). If we exclude the latter from consideration, then the comparison between the pattern of amino acid exchanges in abnormal hemoglobins and that of substitutions in 'normal' proteins makes it possible to detect any influence of natural selection on the type of substitution. Comparisons of this kind have already been made by several workers concerned with the non-randomness of base substitutions (Vogel, 1955; Fitch, 1967; Vogel, 1972; Vogel and Kopun, 1977).

Up to the present day, examples of about 250 abnormal hemoglobin variants, including the variants of α-, β-, γ- and δ-chain are known (International hemoglobin information center, 1976). For the application of these data to the present analysis, we must take note of the fact that the variants are usually detected by electrophoresis. The sampling is therefore biased in favour of mutations that cause different charges. In a few instances, the amino acid mutations, although not involving a charge change, may disrupt the tertiary structure of the molecule so that the overall charge is affected and the change detected. Because of this non-uniformity for the detection, it is necessary to correct the expression for the denominator of equation (2) such that a different weight factor,

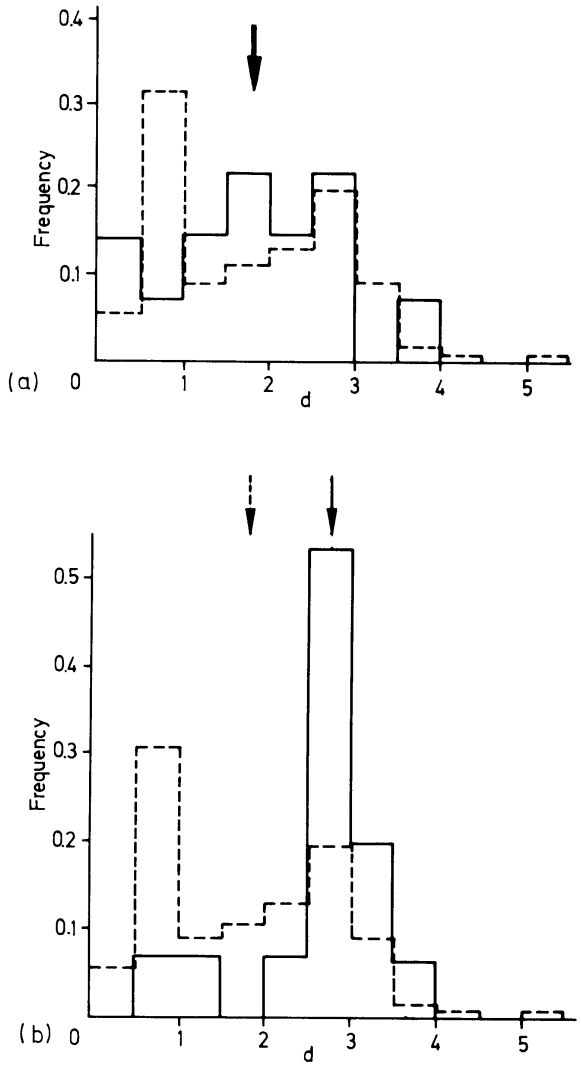

Fig. 2. Frequency distributions of d, the physico-chemical difference, in unstable abnormal hemoglobins classified by the extent of their haemolytic anemia. **(a)** Mild haemoglobinopathy group. **(b)** Severe haemoglobinopathy group. Broken line represents the frequency distribution (random distribution) expected from the genetic code in normal human hemoglobin. Dotted arrow represents the mean distance of the random distribution and solid arrow represents that of the distribution in unstable abnormal hemoglobins

w(i,j), is used for the amino acid pair (a_i,a_j), according to whether it involves a charge change or not. Of the 250 point mutations, about 80% involve charge changes. Thus in this analysis, we tentatively put w equal to 0.8 for a pair involving charge change and w equal to 0.2 for a pair not involving charge change.

Figure 3 shows the substitution pattern of abnormal hemoglobins. There is a positive correlation between R(d) and d and between F(d) and d. That is, the regression of R(d) on d is $R(d) = 0.30d + 0.99$ (the correlation coefficient $\gamma = 0.49$), and that of F(d) on d is $F(d) = 0.17d + 0.08$, ($\gamma = 0.75$). Particularly, the significant frequencies of amino acid replacements are observed at rather larger distances (i.e., $d > 3.5$), where no substitutions are observed in the conservative type. This may well be explained that the abnormal hemoglobin involving a large change in d would have a bigger physiological effect so that it is more likely to be found. This explanation is consistent with the argument mentioned in the analysis of unstable abnormal hemoglobins that the more dissimilar the amino acids are, the larger effect the change between them has on the structure and function of the protein. This positive correlation may be due to a sampling

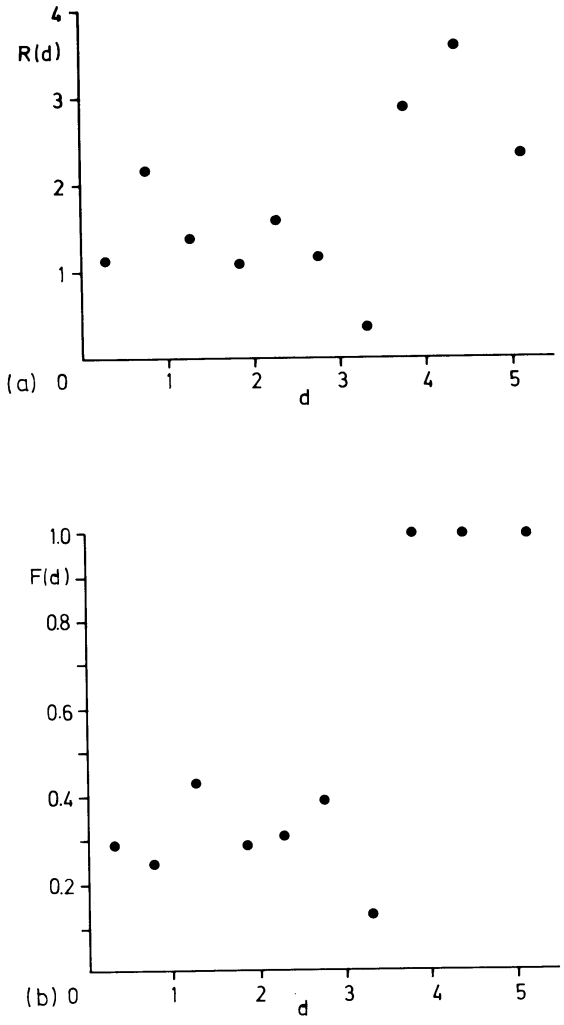

Fig. 3. Relation between distance (d) and RRS (R(d)); **(a)**, and relation between distance and RNFS (F(d)); **(b)**, in abnormal hemoglobin variants. Distance d stands for the physico-chemical difference between changing residues (see eq. 1), and R(d) the coarse grained relative rate of substitutions defined by equation 3. The RNFS (F(d)) stands for the ratio of the number of one-step pairs for which substitutions are observed frequently to the number of one-step pairs expected from the genetic code in the interval of d and d+0.5 (see eq. 4). The value of d for each set of plots is the average distance of pairs in d and d+0.5

effect in the detection. It seems likely that R(d) and F(d) depend less on d in abnormal hemoglobin variants, if they are detected uniformly from population. As the amino acid changes observed in abnormal hemoglobins are not the substitutions that have spread to fixation in populations during evolution, but are only the result of mutations which have not yet been eliminated from populations by natural selection, we may therefore conclude that the negative correlation of substitution frequencies of the amino acids with physico-chemical differences between changing residues in the conservative substitution is the result of natural selection which acts to maintain the original conformation of the protein.

The accepted point mutation matrix for each protein family is also obtained by comparing observed sequences with inferred ancestral sequences of that protein family. Figure 4 shows the calculated results of F(d) for various protein families such as globin (hemoglobin and myglobin), cytochrome c and variable domains of immunoglobulin (hypervariable regions are omitted), including the 'average protein'. The data of myoglobin and immunoglobulin are from Romero-Herrera et al. (1973) and Barker et al. (1972) respectively and the others are from Dayhoff et al. (1972 a, b, c, d). Although the plots are somewhat scattered, the negative correlation between F(d) and d reappears for $d < 3.0$ and F(d) vanishes for $d \geqslant 2.5$. This is the typical substitution pattern of the conservative type, indicating that these protein families have been forced to conserve their existing structures and functions during their evolution since they had assumed their original functions.

It has been confirmed by the above analysis that the negative dependency of the substitution frequency on the difference of the physico-chemical properties between changing residues is a result of natural selection acting on the proteins so as to eliminate the disruptive mutations for the preservation of their conformation. According to this argument, it is expected that, in a protein or a protein moiety in which less selective constraint is imposed on the structure, the substitution frequency depends less on the difference of the physico-chemical properties of changing residues. That is, the amino acid substitutions observed in such polypeptides or protein moieties as hypervariable regions of immunoglobulins, fibrinopeptides and insulin c-peptides are expected to show a pattern distinct from that of conservative substitution. Here, the fibrinopeptides and

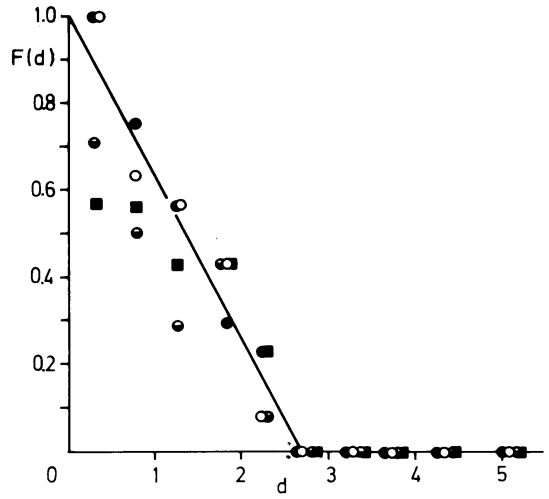

Fig. 4. Relation between distance (d) and RNFS (F(d)) in 'average protein': ●, globin: ○, cytochrome c: ■, and variable domains of immunoglobulin (hypervariable regions are excluded): ⊖. The RNFS (F(d)) stands for the ratio of the number of one-step pairs for which substitutions are observed frequently to the number of one-step pairs expected from the genetic code in the interval of d and d+0.5 (see eq. 4). The value of d for each set of plots is the average distance of pairs in d and d+0.5. Plots are somewhat shifted right and left to distinguish each other when they overlap. Straight line is the regression line of 'average protein'. Amino acid replacements of globin, cytochrome c and variable domain of immunoglobulin were recorded from phylogenetic trees

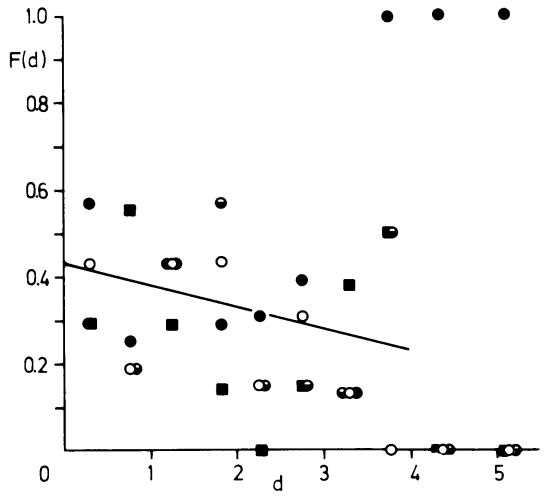

Fig. 5. Relation between distance (d) and RNFS (F(d)) in abnormal hemoglobins: ●, hypervariable regions of immunoglobulin variable domains: ○, fibrino-peptide: ■ and average of hypervariable regions of immunoglobulin variable domains, fibrino-peptide and c-peptide of insulin: ⊖. The RNFS (F(d)) stands for the ratio of the number of one-step pairs for which substitutions are observed frequently to the number of one-step pairs expected from the genetic code in the interval of d and d+0.5 (see eq. 4). The value of d for each set of plots is the average distance of pairs in d and d+0.5. Plots are somewhat shifted right and left to distinguish each other when they overlap. Straight line is the regression line on the plots, i.e., $F(d) = -0.05d + -.43$ for $0 < d < 4.0$. Amino acid replacements of hypervariable regions of immunoglobulin variable domain, fibrino-peptide and c-peptide of insulin were recorded from phylogenetic trees

insulin c-peptides are protein moieties which are removed from the precursors. The hypervariable regions of immunoglobulins are parts of the variable domains (V_L and V_H). A high frequency of the amino acid substitutions and gross structural changes are observed in these regions (Poljak et al. 1976; Padlan and Davies, 1975; Padlan, 1977). It has been confirmed by structural analysis that the hypervariable regions represent the complementarity-determining parts of antibodies and that these regions are close together in space to form the antigen binding sites (Padlan, 1977). The structural divergence of hypervariable regions thus results in the functional (i.e., specificity for antigen) divergence of immunoglobulins.

Figure 5 shows the relations between F(d) and d for the polypeptides that are removed from the original chains and for protein moieties in which gross structural divergence is observed. The result for abnormal hemoglobins is also shown in this figure for comparison. Although plots are somewhat scattered, the strong negative dependency of F(d) on d, the characteristic of conservative substitution, is not seen in these polypeptides or protein moieties, but the substitution frequencies seem rather not to depend on physico-chemical factors. In fact, the regression of F(d) on d is $F(d) = -0.5d + 0.43$ for $d < 4.0$ for the data accumulated from fibrinopeptides, hypervariable regions of immunoglobulins and insulin c-peptides. Another different point between the substitution pattern in these polypeptides and that of conservative substitution is that, in these polypeptides, substitutions are observed with appreciable frequency over d=3.0, where almost all the changes are excluded in conservative substitution. Although the substitution patterns in these polypeptides are rather similar to the pattern in abnormal hemo-

globins in the point that F(d) depends less on physico-chemical factors, there is a difference above d=4.0. That is, no substitution is found there during evolution of the polypeptide regions examined, in striking contrast to the substitution pattern in abnormal hemoglobins. Presumably it means that the most extreme changes are not acceptable even in peptides that are excised or in hypervariable regions. We may therefore infer that most of the substitutions that occur in these parts of proteins are almost free from those selective constraints that tend to maintain the precise original conformation in whole proteins. If some selection intervenes here also, it bears on different parameters (Zuckerkandl, 1975).

Discussion

The typical pattern of conservative substitution (or 'high-constraint substitution', contrasting with 'low-constraint substitution') is that the observed frequency of amino acid substitutions decreases as the physico-chemical difference between changing residues increases and vanishes when the extent of this difference exceeds some critical value. As Figure 4 shows, there is no appreciable difference between substitution patterns in various protein families, within the variation expected from the small sample size. This fact shows that amino acid substitutions always abide by the same substitution pattern, irrespective of protein families, whenever amino acid substitutions occur. Indeed, as we can see in Table 3, the mean distances of substituted amino acids are nearly constant for various proteins such as cytochrome c, myoglobin, hemoglobin α- and β-chain, and immunoglobulin variable domains. As the characteristic pattern in conservative substitution results from the natural selection that tends to maintain the existing three dimensional conformation, we may venture the generalization that the relative intensity of selection pressure against the amino acid substitutions is almost the same in different globular proteins. It should be noted that, although the present analysis has shown that mutations that are more likely to disrupt the structure are eliminated by natural selection, this does not tell us anything about how the substitutions that are observed do occur and by what mechanism accepted mutations are fixed in populations. Whether they are selectively neutral or not remains to be determined (Kimura and Ohta, 1974).

The negative correlation between the observed frequency of amino acid substitutions and the extent of physico-chemical difference may be interpreted as follows: The degree to which the tertiary structure of proteins is affected by substitutions, even though the same physico-chemical differences between changing residues are involved, may differ from site to site in the protein. The substitutions between amino acids having similar physico-chemical properties, i.e., $d \leqslant 1.0$, may not result in any significant change of tertiary structure over almost all the variable sites in the protein within a short time interval of evolution. There may be some sites where an amino acid can be substituted for another having distinct physico-chemical properties (for example, $d \sim 3.0$) without causing the structural disruption of the protein. But the number of these sites may be very limited in the protein molecule. In other words, there may be more sites available for substitutions between similar amino acids that for substitutions between dissimilar amino acids in a short time interval of the protein evolution. This

Table 3. Mean distance, standard deviation and number of substituted amino acid pairs for various protein families, in which substitution patterns obey the conservative type

	Mean of distance	Standard deviation	Number of substitutions
Cytochrome	1.16	0.81	121
Myoglobin	1.07	0.70	140
Hemoglobin α and β	1.05	0.66	201
Immunoglobulin variable domains (Hypervariable regions are excluded)	1.15	0.87	123

may be the reason why the substitution frequency is reduced as the distance d increases in the interval $0 < d < 3.5$. The changes between more distant amino acids, i.e., $d > 3.5$, are so disruptive for the original conformation of the protein that there are not sites which accept them.

The above characteristic patterns in amino acid substitutions make it possible to quantify the intensity of selective constraint acting against a codon change occurring on variable sites in a cistron. As a codon specifies an amino acid except three termination codons, the 'distance between codons' can be defined by using the same physico-chemical parameters as for amino acids. A point mutation occurring on any one of the codons leads to a change of the codon to another whose physico chemical properties differ from those of the original one by d. (Here, we define $d = 0$ for a synonymous change and $d = \infty$ for a change to a stop codon.) According to the present analysis, it seems reasonable to assume that, as an average, the rate of acceptance, s(d), of the mutation decreases linearly on d and vanishes over some critical distance d_c, irrespective of protein families, whenever the change occurs at any one among variable sites. We therefore have an expression for s(d) as

$$s(d) = \begin{cases} 1 & ; \text{for } d = 0 \text{ (synonymous change)} \\ f_c(1 - d/d_c) & ; \text{for } 0 < d < d_c \\ 0 & ; \text{for } d \geqslant d_c \end{cases}$$

where the value of s(d) is measured relative to the value for synonymous changes. The parameter f_c is the rate of acceptance of a mutation by which any one of the codons changes to another with the distance between them being infinitesimal. According to our conjecture that amino acid substitutions always abide by the same substitution pattern whenever they take place, it seems likely that, though f_c is not necessarily equal to unity, it has a constant value regardless of protein families. The critical distance d_c takes a distinct value, depending on whether the substitution pattern is of a high-constraint or a low-constraint type: From our present analysis, d_c takes nearly constant value 3.0 for high-constraint substitution and is sufficiently large for low-constraint substitution. We may therefore say that the inverse of d_c reflects the extent of selective constraint that tends to maintain the original conformation of the protein molecule. This model may be useful to predict to what extent a contemporary nucleotide se-

quence or amino acid sequence will undergo substitutions during the course of its evolution from now on. An application of this model will be given in a separate paper.

So far, we have been analysing the pattern of amino acid substitutions by using the closely related sequences only. McLachlan (1971) has counted the large number of amino acid substitutions which have occcurred in 17 homologous protein families in which distantly related sequences are also included and has constructed the matrix of accepted point mutations. All the matrix elements corresponding to two- and three-

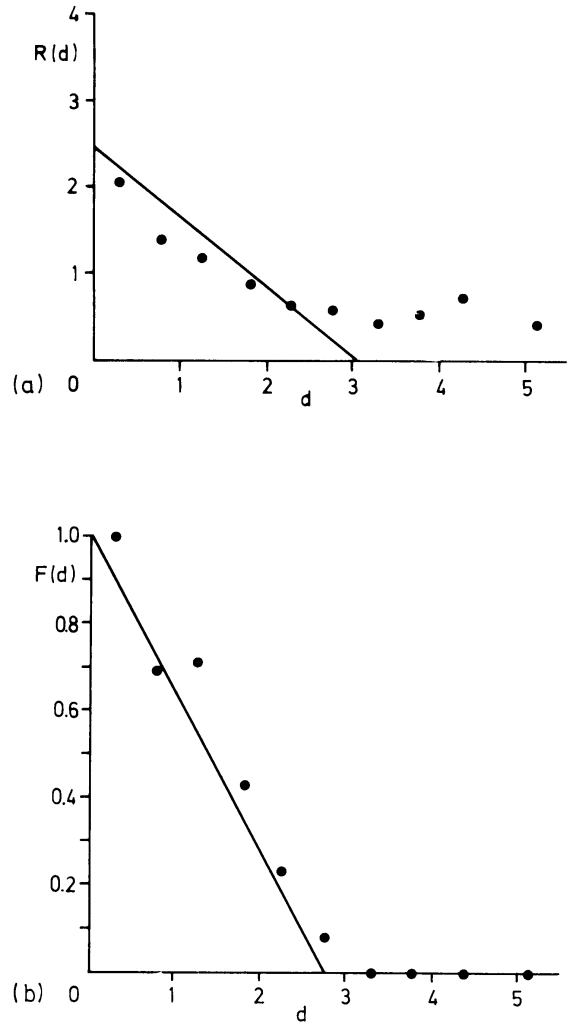

Fig. 6. **(a)** Relation between distance (d) and RRS (R(d)) in distantly related sequences accumulated from homologous protein families compiled by McLachlan (1970). The RRS stands for the coarse grained relative rate of substitutions (see eq. 3), and d the physico-chemical difference between changing residues (see eq. 1). The value of d for each set of plots is the average distance of pairs in d and d+0.5. Straight line is the regression line of 'average protein'. **(b)** Relation between distance and RNFS (F(d)), the ratio of the number of one-step pairs for which substitutions are observed frequently to the number of one-step pairs expected from the genetic code in the interval of d and d+0.5 (see eq. 4). The value of d for each set of plots is the average distance of pairs in d and d+0.5. Straight line is the regression line of 'average protein'

step changes have non-zero values since distantly related sequences are also compared with each other and accumulate a large number of substitutions. Even the matrix elements corresponding to one-step pairs may contain many changes produced by successive one-step changes.

Figure 6 shows the result of the caluculation for RRS and RNFS, by applying these data and by neglecting the matrix elements corresponding to two- and three-step change. The regression lines in the case of closely related sequences, i.e., 'average protein' are also shown for comparison. For relatively small distances, i.e., $d < 3$, the plots of R(d) fit well with the regression line of the 'average protein', though they are slightly lower than this line. But appreciable frequencies of substitution are observed even for large distances, i.e., $d > 3$. On the contrary, the plots of F(d) fit well with the regression line of 'average protein' over all the ranges of distance.

These features are interpreted as the result of conservative substitution: In a relatively short time span, almost all the substitutions are produced by one-step changes. A substitution from amino acid a_i to a_j may be eliminated, if the physico-chemical difference between a_i and a_j is large enough, i.e., $d_{ij} \geqslant 3.5$. But in a relatively long time span, there may be many paths from a_i to a_j by successive one-step substitutions which are conservative. The substitution between a_i and a_j through these paths is also conservative for the tertiary structure of a protein. Because the frequency of substitution between a_i and a_j may be proportional to the products of the frequencies of substitution at each step in that path, F(d) goes to zero for relatively large distances. Recently, Goodman and Moore (1977) have analysed the similar amino acid substitution process by using their conformational parameter distance and have suggested that by way of intermediate amino acids almost any amino acid can ultimately be substituted for another through successive one-step changes without damage to an evolving protein's conformation during the process.

Acknowledgements. We are greatly indebted to Prof. H. Matsuda, Dr. M. Go and Dr. K. Ishii for many stimulating discussions and for valuable suggestions. We also wish to express our thanks to Prof. E. Zuckerkandl and two reviewers of this journal for stimulating and important comments.

References

Allison, A.C. (1954), British Medical Journal **1**, 290

Barker, W.C., McLanghlin, P.L., Dayhoff, M.O. (1972), Evolution of a complex system: The immunoglobulins, In: Atlas of protein sequence and structure, M.O. Dayhoff ed., pp. 31-40, Maryland: National Biomedical Research Foundation

Clarke, B. (1970), Nature (Lond.), **228**, 159-160

Dayhoff, M.O., Eck, R.V., Park, C.M., (1972 a), A model of evolutionary change in proteins, In: Atlas of protein sequence and structure, M.O. Dayhoff ed., pp. 89-100, Maryland: National Biomedical Research Foundation

Dayhoff, M.O., Park, C.M., McLaughlin, P.J., (1972 b), Building a phylogenetic tree: Cytochrome c., In: Atlas of protein sequence and structure, M.O. Dayhoff ed., pp. 7-16, Maryland: National Biomedical Research Foundation

Dayhoff, M.O. Hunt, L.T., McLaughlin, P.J., Jones, D.D. (1972 c). Gene duplication in evolution: The globin. In: Atlas of protein sequence and structure. M.O. Dayhoff ed., pp. 17-30. Maryland: National Biomedical Research Foundation.

Dayhoff, M.O., Hunt, L.T., McLaughlin, P.J., Barker, W.C. (1972 d). Data section. In: Atlas of protein sequence and structure. M.O. Dayhoff ed., pp. D-87-D-98. pp. D-173-D-228. Maryland: National Biomedical Research Foundation

Epstein, C.J. (1967). Nature (Lond.) **215**, 335-359

Fitch, W.H. (1966). J. Mol. Biol. **10**, 9-16

Fitch, W.H. (1967). J. Mol. Biol. **26**, 499-507

Fitch, W.M., Markowitz, E. (1970). Biochemical Genetics **4**, 579-593

Goodman, M., Moore, G.W. (1977). J. Mol. Evol. **10**, 7-47

Grantham, R. (1974). Science **185**, 862-864

Hasegawa, M., Yano, T. (1975). Viva Origino **4**, 11-18 (in Japanese)

Hayashi, A. (1975). Abnormal hemoglobin. In: Gendai Seibutu Kagaku, Vol. 16, Taisha Izyo, U. Yamamura ed., pp. 1-41 (in Japanese), Tokyo: Iwanami publishing company

International Hemoglobin Information Center (1976). R.N. Wtightstone, director, Medical College of Georgia

Kimura, M., Ohta, T. (1974). Proc. Nat. Acad. Sci. USA **71**, 2848-2852

Lehmann, H., Huntsman, R.G. (1974). Unstable haemoglobins and haemoglobins with altered oxygen affinity. In: Man's Haemoglobins. pp. 217-235, Amsterdam, Oxford: North-Holland publishing company

McLachlan, A.D. (1971). J. Mol. Biol. **61**, 409-417

Padlan, E.A. (1977). Quant. Rev. Biophys. **10**, 35-65

Padlan, E.A., Davies, D.R. (1975). Proc. Nat. Acad. Sci. USA **72**, 819-823

Perutz, M.F., Lehmann, H. (1968). Nature **219**, 902-909

Poljak, R.J., Amzel, L.M., Phizackerley, R.P. (1976). Prog. Biophys. Molec. Biol. **31**, 67-93

Romero-Herrera, A.E., Lehmann, H., Joysey, K.A., Friday, A.E. (1973). Nature **246**, 389-395

Sneath, P.H.A. (1966). J. Theor. Biol. **12**, 157-195

Vogel, F., Kopun, F. (1977). J. Mol. Evol. **9**, 159-180

Vogel, F. (1972). J. Mol. Evol. **1**, 334-367

Vogel. F., Rohrborn, G. (1966). Nature **210**, 116-117

Zuckerkandl, E. (1975). J. Mol. Evol. **7**, 1-57

Zuckerkandl, E., Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In: Evolving genes and proteins. V. Bryson and H.J. Vogel eds., pp. 97-116, New York: Academic Press

Received October 7, 1977/Revised May 28, 1978