Int. J. Peptide Protein Res. 15, 1980, 211–224

RELATIONSHIP BETWEEN MUTABILITY, POLARITY AND EXTERIORITY OF AMINO ACID RESIDUES IN PROTEIN EVOLUTION

MITIKO GŌ and SANZO MIYAZAWA

Department of Biology, Faculty of Science, Kyushu University, Fukuoka, Japan

Received 17 April, accepted for publication 27 September 1979

A systematic study was carried out on mutability of amino acid residues in evolving proteins in relation to their polarity and location within three-dimensional structure of proteins. Exteriority of residue sites is quantitatively defined as accessibility based on their static accessible surface area to solvent water molecule. Residue sites are classified into interior and exterior depending on their accessibility. More frequent substitution on exterior sites is confirmed to be general in eight sets of homologous protein families regardless of their biological functions and of presence or absence of a prosthetic group. Virtually all types of amino acid residues are found to have higher mutabilities on the exterior than in the interior. No correlation between mutability and polarity was observed of amino acid residues in the interior and on the exterior, respectively. Amino acid residues are classified into three depending on their polarity, polar (Arg, Lys, His, Gln, Asn, Asp and Glu), weak polar (Ala, Pro, Gly, Thr and Ser) and nonpolar (Cys, Val, Met, Ile, Leu, Phe, Tyr and Trp). Amino acid replacements during protein evolution are very conservative; 88% and 76% of them in the interior and on the exterior, respectively, are within the same group of the three. Inter-group replacements are such that weak polar residues are replaced more often by nonpolar residues in the interior and more often by polar residues on the exterior.

Key words: accepted point mutation matrix; accessible surface area; evolutionary change; homologous proteins.

Frequency of amino acid substitutions in evolving protein molecules has been found to depend on the location of residue sites within their three-dimensional structures. X-ray crystallographers noticed that between a pair of homologous proteins there were more unvaried amino acid residues in the interior than on the surface of proteins (hen egg-white lysozyme and human lysozyme (Blake & Swan, 1971), elastase and α-chymotrypsin (Shotton & Watson, 1970), subtilisin BPN′ and subtilisin Carlsberg (Wright *et al.*, 1969) and lamprey hemoglobin and various globins (Hendrickson & Love, 1971)).

The variability of residues has also been recognized to correlate with their hydrophobicity or polarity. The tendency for hydrophobic residues in cytochrome c to be invariant residues was noticed and interpreted by Dickerson *et al.* (1971) to be due to the necessity of maintaining proper interactions with the heme. The tendency of the residues

0367–8377/80/030211–14 $02.00/0

interacting with the heme groups to be invariant was also noticed by Perutz *et al.* (1965) also in the case of hemoglobin. Zuckerkandl & Pauling (1965) also noted that the polar residues in globin chains were more variable than the non-polar ones.

It is well established that non-polar residues exist more frequently in the interior than on the exterior of a protein and vice versa for polar residues (Chothia, 1976). Thus the correlations between variability and exteriority, and between variability and polarity of residues, are not independent properties. Vogel & Zuckerkandl (1972) analyzed the correlations between polarity, variability and exteriority of residues in globins. They found that polarity and exteriority as well as exteriority and variability were significantly correlated. But polarity and variability were unexpectedly not correlated. They argued it was because the correlation between polarity and variability was upset by the high polarity and the low variability at the contact sites among the globin monomers.

In these situations we are naturally led to the following questions: Is absence of correlation between variability and polarity a phenomenon peculiar to globins, or is it more general? Is the existence of prosthetic group essential for hydrophobic residues to be less variant as observed in cytochrome c?

The purpose of this paper is to conduct a systematic study of correlations between variability, polarity and exteriority in several sets of homologous proteins in order to clarify such questions as stated above, i.e. the presence or absence of the correlations and their generality as to the variety of proteins. For this purpose a quantitative definition of variability, defined here as mutability, is given, a quantitative scale of polarity of residues is described, and a quantitative definition of exteriority, defined here as accessibility based on the accessible surface area of residues, is given.

Special care must be paid to see the correlation between mutability and polarity. Because of the correlation between mutability and accessibility (which will be shown to be quite general in this paper) and the well-established correlation between polarity and accessibility, it is natural to expect that polar residues are more mutable if we consider all residues in a protein at a time. What we are more interested in is whether or not polar residues are intrinsically more mutable. In order to see this, the residue sites are classified into two, interior and exterior, based on the accessibilities of residues. Then the correlations are examined in each of the interior and exterior reisdues.

METHODS

A. Quantitative definition of exteriority of amino acid residues by their accessibility to water molecule

Exposure of protein atoms to solvent is obtained by computing static accessible surface area of the atom to water molecule (Lee & Richards, 1971; Shrake & Rupley, 1973; Chothia, 1975). The accessible surface area of an atom is calculated by the method of Shrake & Rupley (1973). Accessible surface area of a residue is obtained by summing the accessible surface area of its component atoms. *The accessibility of a residue* is defined by the accessible surface area of the residue in native conformation divided by that in an extended state. This quantity measures quantitatively the degree of exposure of a residue to solvent. The accessible surface area of residue X in an extended state is calculated for a tripeptide Gly-X-Gly in a β conformation ($\phi = -140°$, $\psi = 135°$) with dihedral angles of side chain in *trans* positions (χ's = 180°), except when X is a proline residue, in which case the conformation of endo positions of the c^{γ} atom ($\phi = -75.0°$ and $\chi_1 = 18.67°$) (Momany *et al.*, 1975) and ψ-value of poly-L-proline II ($\psi = 146.0°$) are adopted. The conformation of peptide bond is fixed at *trans* position ($\omega = 180°$) in all residues. The atomic coordinate of the tripeptide is generated by using ECEPP program (submitted to Quantum Chemistry Program Exchange, Indiana Univ., by H.A. Scheraga) and employing the geometrical parameters described by Momany *et al.* (1975). The van der Waals' radii used in the computations of the surface areas of the solvated van der Waals' spheres are given in Table 1. We do not explicitly consider hydrogen atoms, which are incorporated into the van der Waals' radii for groups. Though

TABLE 1

Assumed van der Waals' radii[a]

		Radius in Å
Non-aromatic carbon	$-CH_3$, $-CH_2-$, >CH–	2.0
	$-CH=CH_2$, >C=, ⩾C–	1.74
Aromatic carbon	His ring, Trp fused ring (ring 5)	1.77
	Phe ring, Tyr ring, Trp fused ring (ring 6) tosyl, porphin	1.86
Nitrogen	main chain amide, His ring, Trp ring, porphin	1.65
	Gln side chain, Asn side chain	1.75
	Arg guanidino group	1.8
	Lys side chain	1.9
All oxygen		1.5
All sulfur		1.8
Fe^{3+}		0.64
Water molecule		1.4

[a] Taken from Bondi (1968), Ch. 14, and Pauling (1960), Ch. 13.

most of the values in Table 1 are those of Bondi (1968) and Pauling (1960), the rests are assumed from the values of the similar groups.

Shrake & Rupley (1973) calculated the accessible surface area of the residue X in conformation of the Gly-X-Gly tripeptide from the corresponding atoms in the native protein and they averaged it over its location in the folded molecule. These values for 20 types of amino acid residues are found to be similar to the accessible surface area of X in Gly-X-Gly in β conformation, i.e. the difference is smaller than 10%.

The coordinates of atoms determined by the X-ray analysis of the following proteins are used for the computation of the accessible surface area of residues in native conformations. They are tuna cytochrome c, oxidized (Swanson *et al.*, 1977), bovine pancreatic ribonuclease S (Fletterick & Wyckoff, 1975), hen egg white lysozyme (Diamond, 1974), sperm whale myoglobin (Takano, 1977), tosyl α-chymotrypsin (Birktoft & Blow, 1972), subtilisin BPN′ (Alden *et al.*, 1971) and horse deoxyhemoglobin (Bolton & Perutz, 1970). For hemoglobin chains α and β, the classification of location of a residue site is done for a tetrameric state consisting of two α and β chains. For most cases the X-ray coordinates are not available for homologous proteins to the proteins mentioned above. In this situation we assume that the folding structures of all homologous proteins are the same as the one for which the X-ray structural analysis has been done.

B. Classification of residues into exterior and interior

We define residues with accessibility less than 0.27 as interior, otherwise as exterior. The choice of this number 0.27 is somewhat arbitrary. For this value, the numbers of interior and exterior residues become roughly the same in such small proteins as cytochrome c and myoglobin. As proteins become larger, there are more interior residues. We also used 0.20, instead of 0.27, to classify residues into interior and exterior. We obtained results essentially unaltered from those described in this paper. In this paper we give results obtained for the choice of 0.27.

C. Invariant sites

Dayhoff *et al.* (1972*a*) constructed phylogenetic trees of several protein families from the available protein sequences. Ancestral sequences were determined at the same time. They counted the number of replacements occurring at each residue site along the branches of the phylogenetic trees. A residue site is defined as invariant if no replacement is counted at that site. This definition of the invariant sites depends on the number and diversity of species from which the homologous proteins are sampled. Apparently the number of the invariant sites shrinks as new sequences of homologous proteins are added. Zuckerkandl & Pauling (1965) anticipated that the final number of the invariant sites in globins may be 1 or 2. The invariant sites defined in this paper should be understood as those with very low variability.

D. Mutabilities of each type of the amino acid residues in the interior and exterior

We first construct the accepted point mutation matrix A introduced by Dayhoff *et al.* (1972*b*). When a phylogenetic tree is given for a protein family, we assign a pair of types of amino acid residues (i, j) (i, j = 1, 2, . . . , 20, corresponding to the 20 types of amino acid residues) for each residue site and along each branch of the phylogenetic tree, in such a way that an amino acid of type i is substituted by that of type j at the residue site and along the branch in question. If no substitution occurs, then i = j. The element A_{ij} of the accepted point mutation matrix is the number of (i, j) pairs counted along all branches and at all residue sites. Following Dayhoff *et al.* (1972*b*) we symmetrize the obtained matrix by redefining A_{ij} and A_{ji} by $(A_{ij} + A_{ji})/2$. Dayhoff's reasoning for this symmetrization is that the likelihood of amino acid i replacing j would depend on the product of the frequencies of occurrence of the two amino acids and on their chemical and physical similarity. Besides the above reason, the symmetrized matrix A ensures the detailed balance of substitutions in the stationary state. Therefore, the process of symmetrization is equivalent to assuming that the amino acid composition in the protein families studied is already in the stationary state.

We calculate the accepted point mutation matrices in the interior (A_{in}) and exterior (A_{ex}) in order to obtain the mutabilities of each type of the amino acid residue in the interior and exterior. The homologous proteins used are cytochrome c, hemoglobin α and β chains, and myoglobin whose contemporary and ancestral sequences are taken from Dayhoff *et al.* (1972*a*, *c*) (first three proteins) and from Romero-Herrera *et al.* (1973, 1976) (Myoglobin). The matrices for the four protein families are added to obtain common interior (A_{in}) and exterior (A_{ex}) matrices. In the construction of matrices long branches of the phylogenetic trees whose lengths are over 15 PAM's (accepted point mutation per 100 amino acid sites) are omitted. The sequences which contain more than 10% unknowns (including either asparagine or aspartic acid and either glutamine or glutamic acid) and gaps are not used in the construction of the matrices.

It is plausible to assume that amino acid replacement in proteins is caused by single base change at a time in the structural genes coding the proteins. In fact, all point mutations, over 250, observed in abnormal hemoglobins are explainable by considering a single base change in the structural genes. We use in this paper the terms one and two base changes in the sense that at least such a number of base changes is required for an amino acid to be replaced by the other amino acids. Because long branches are omitted in the construction of the accepted point mutation matrices, only 16% and 8% of the replacements on the exterior and in the interior, respectively, are caused by two base changes and the remaining 84% and 92%, respectively, are by single base changes. In order to examine the amino acid replacements in short time intervals, we will use in the following part of this paper the accepted point mutation matrices where the elements corresponding to more than one base change are reduced to zero. It must be pointed out, however, that the possibility is not zero that the replacements by multiple mutations are still included in the present data. For instance, the observed replacement of Gly by Glu is considered one base change but it might happen in the actual process as two base change, i.e. the two continuous one step mutations of Gly to Asp and of Asp to Glu.

The mutability of an amino acid of type i in the interior $m_{in,i}$ (or on the exterior $m_{ex,i}$) is defined as the probability that an amino acid of type i in the interior (or on the exterior) will be replaced by an amino acid of other type along an average branch of the phylogenetic trees considered. They are given by

$$m_{a,i} = 1 - \left(A_{a,ii} / \sum_{j=1}^{20} A_{a,ji} \right)$$

$$\begin{pmatrix} a = \text{in, ex} \\ i = 1, 2, \ldots, 20 \end{pmatrix} \qquad (1)$$

The frequency of amino acid of type i in the interior, $f_{in,i}$ (or on the exterior $f_{ex,i}$), is given by

$$f_{a,i} = \sum_{j=1}^{20} A_{a,ji} / \sum_{i,j=1}^{20} A_{a,ji}$$

$$\begin{pmatrix} a = \text{in, ex} \\ i = 1, 2, \ldots, 20 \end{pmatrix} \qquad (2)$$

Evolutionary change in the interior, μ_{in} (or on the exterior, μ_{ex}), is defined as the probability that an amino acid of any type in the interior (or on the exterior) will be replaced by an amino acid of other type along an average branch considered. They are given by

$$\mu_a = 1 - \sum_{i=1}^{20} A_{a,ii} / \sum_{i,j=1}^{20} A_{a,ji}$$

$$(a = \text{in, ex}) \qquad (3)$$

or

$$\mu_a = \sum_{i=1}^{20} m_{a,i} f_{a,i}$$

$$(a = \text{in, ex}). \qquad (4)$$

It should be remarked again that the mutabilities and the evolutionary change as defined above are the probabilities of amino acid replacement along an average branch of phylogenetic trees considered. One of the reasons for omitting long branches from the construction of accepted point mutation matrices is to restrict the distribution of time lengths of branches as narrowly as possible. This narrowness of the distribution of the branch lengths allows the four protein families to be added to obtain the commonly accepted point mutation matrices A_{in} and A_{ex}. The average time lengths taken in interpretation of the mutabilities and the evolutionary change in the interior and on the exterior are, of course, the same because the identical branches of phylogenetic trees are used in the construction of the two matrices.

E. Polarity scale and classification of 20 amino acid residues into three groups according to their polarity

We make use of the polarity scale given by Grantham (1974). This dimensionless scale which shows only relative ordering was determined from the published data of Woese (1973) and Aboderin (1971). In Table 2 the polarity

TABLE 2

Grouping of 20 amino acids according to their polarity[a]

Group	No.	Amino acid	Polarity	Volume (Å^3)
Polar	1	Arg	10.5	124.0
	2	Lys	11.3	119.0
	3	His	10.4	96.0
	4	Gln	10.5	85.0
	5	Asn	11.6	56.0
	6	Asp	13.0	54.0
	7	Glu	12.3	83.0
Weak polar	8	Ala	8.1	31.0
	9	Pro	8.0	32.5
	10	Gly	9.0	3.0
	11	Thr	8.6	61.0
	12	Ser	9.2	32.0
Nonpolar	13	Cys	5.5	55.0
	14	Val	5.9	84.0
	15	Met	5.7	105.0
	16	Ile	5.2	111.0
	17	Leu	4.9	111.0
	18	Phe	5.2	132.0
	19	Tyr	6.2	136.0
	20	Trp	5.4	170.0

[a] Polarity and volume of each amino acid are taken from Grantham (1974). Polarity is a dimensionless quantity to show only relative ordering. Volume is for side chain only.

scale as well as volumes of side chains, which are also estimated by Grantham (1974), are given. In this Table, the amino acids are arranged so that nearby amino acids have similar values of both polarity and volume. Twenty amino acid residues are grouped into three groups: polar, weak polar and nonpolar, as in Table 2 according to the polarity scale. It should be noted here that the weak polar residues possess incidentally small side chain volumes. It is possible to group the residues in different ways. For an example, only the charged residues at physiological *ph* might be put into the polar group and Gln and Asn into weak polar group. As will be shown in a later section (in Table 4), however, Gln and Asn are more similar to the charged residues than the other weak polar residues in their more frequent existence on the exterior than in the interior and also in the more frequently accepted mutations between those two residues and the charged residues than between those two residues and the weak polar residues.

RESULTS

A. Distribution of invariant sites in the three-dimensional structure of proteins

An example of calculated accessibilities of residues is given in Fig. 1 for the case of cytochrome c. In Fig. 2 the value of accessibility is divided into eight ranges, and the number of sites with accessibility in each range is plotted by dotted lines for each of the eight proteins. The fraction of the invariant sites in those sites is shown by the columns. The tendency of residual sites to be less invariant as they are more exposed is clearly seen in all proteins examined.

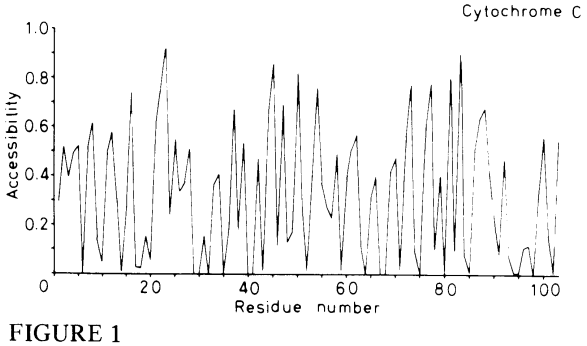


FIGURE 1

The accessibility of residue is plotted against the residue number for cytochrome c.

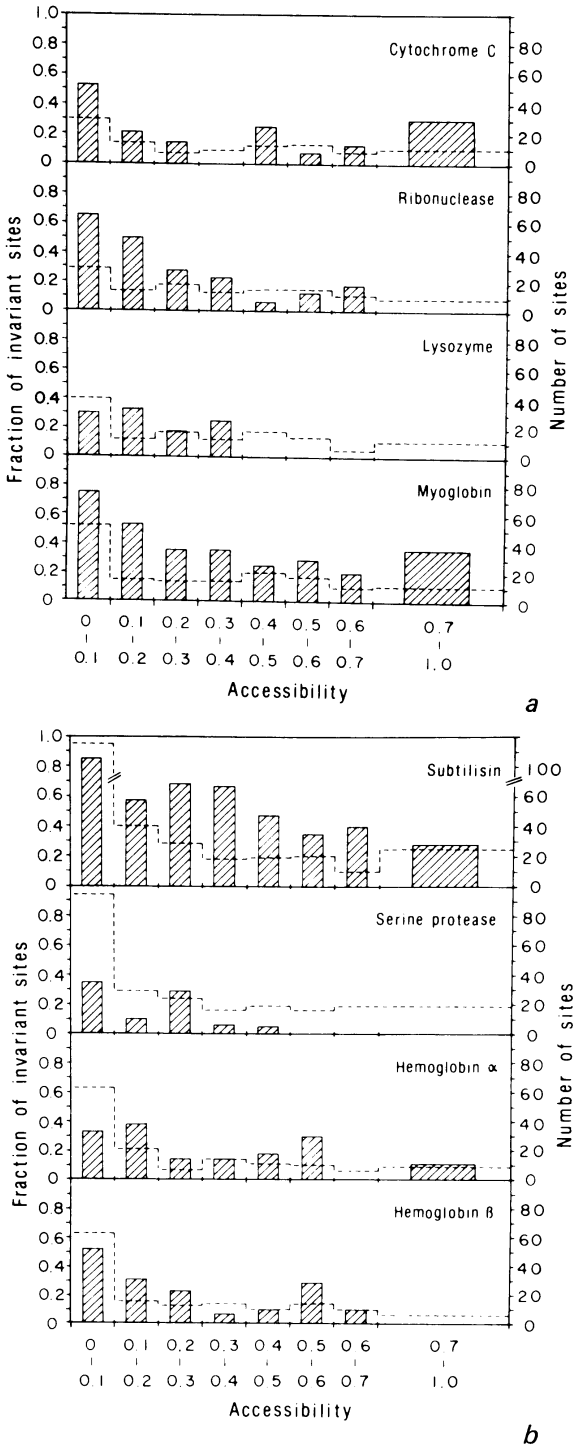


FIGURE 2

(a) The number of the residues is plotted against their accessibilities (dotted line) and the fraction of the invariant residues in them (column) for cytochrome c, ribonuclease, lysozyme and myoglobin. (b) The same to (a) but for subtilisin, serine protease, hemoglobin α and β chains.

TABLE 3

Ratio (f_{in}/f_{ex}) of the frequency of the invariant residues in the interior (f_{in}) to that on the exterior (f_{ex})[a]

Protein	f_{in} (invariant/interior total res.)	f_{ex} (invariant/exterior total res.)	f_{in}/f_{ex}
cytochrome c	19/47	9/56	2.5
ribonuclease	30/54	9/70	4.3
lysozyme[b]	19/65	3/64	6.2
myoglobin	50/76	23/77	2.2
subtilisin	136/177	44/98	1.7
serine protease[c]	40/139	5/97	5.6
hemoglobin α[d]	30/87	8/54	2.3
hemoglobin β[d]	41/90	7/56	3.6

[a] Sequences used to identify the invariant residues are from Dayhoff *et al.* (1972*d*), Boulter *et al.* (1972), Croft (1973, 1974, 1976), Romero-Herrera *et al.* (1973, 1976) and Beintema *et al.* (1977).
[b] Lactalbumin from three species are added to lysozyme from eight species.
[c] This consists of three trypsin, three chymotrypsin, one elastase and one thrombin.
[d] The tetrameric state consisting of two α chains and two β chains is used to classify the residues into the interior and exterior.

In order to express the above observation slightly differently, the frequencies of the invariant sites in the interior (accessibility $\leqslant$ 0.27) and exterior (accessibility $>$ 0.27) sites are compared. The numbers of the invariant sites in the interior and exterior and the numbers of the interior and exterior sites are shown in Table 3 as well as the ratio of the frequency of the invariant sites in the interior to that on the exterior. The ratios are much larger than unity for every protein set as expected from the distribution of the invariant sites shown in Fig. 2. (Comparison of the ratios given in Table 3 among the proteins is meaningless because of the different sampling space of the protein sequences employed in determination of the invariant sites. The particularly high ratios shown by lysozyme and serine protease seems due to the functional divergence; i.e. the former set of homologous proteins consists of lactalbumin and lysozyme and the latter consists of trypsin, chymotrypsin, elastase and thrombin.)

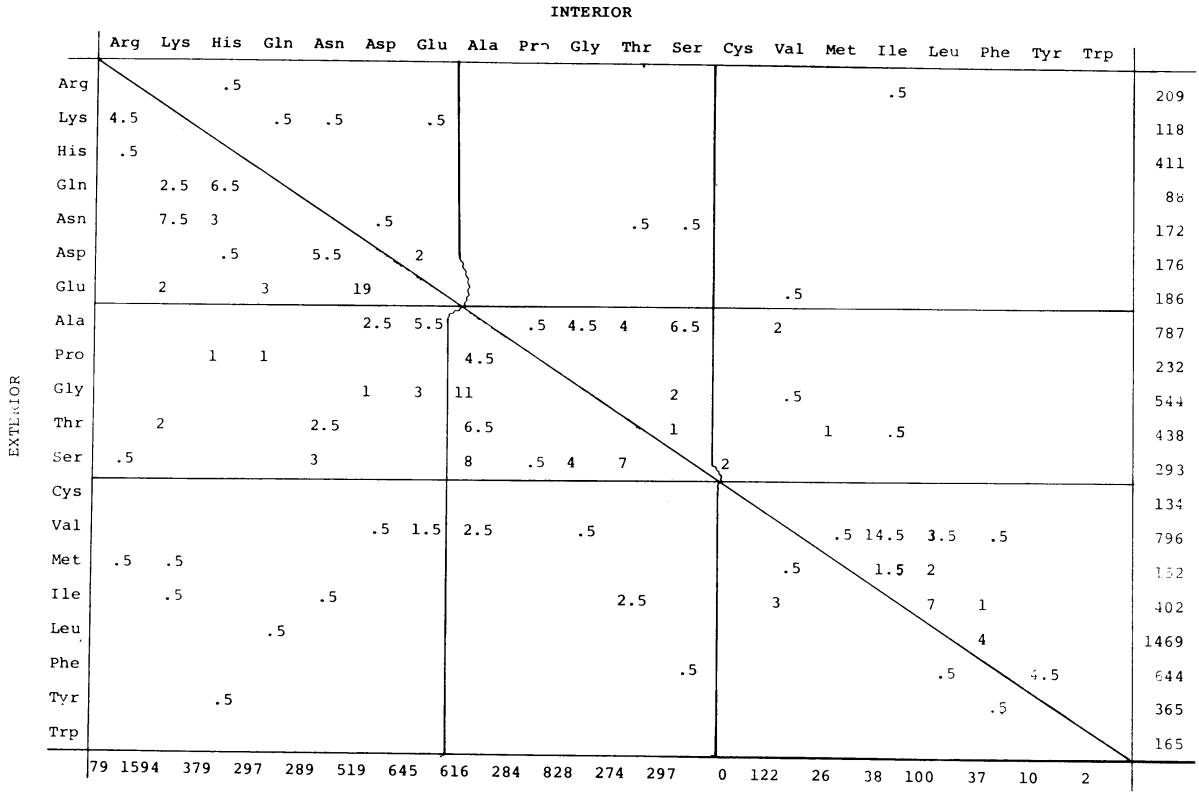
The proteins examined include those without a prosthetic group. Yet, we observed in all cases the tendency that a residue site is more probably invariant, as it is less accessible. This indicates that the reason for this feature should be looked for in the globularity of conformations rather than in properties specific to each protein family.

B. Mutabilities of 20 amino acid residues and their dependence on the accessibility

The accepted point mutation matrices in the interior (A_{in}) and on the exterior (A_{ex}) are shown in Table 4. The matrix elements are obtained by summing the elements of the corresponding accepted point mutation matrices for cytochrome c, myoglobin and hemoglobin α and β chains. The mutabilities of each amino acid residue in the interior and exterior are calculated by eqn. (1) from the accepted point mutation matrices A_{in} and A_{ex} and are plotted in Fig. 3. It is seen in Fig. 3 that almost all the amino acid residues exist in the upper left region except cysteine. This amino acid residue was never observed on the exterior of proteins along the phylogenetic trees studied (see Table 4). Except for this case the amino acid residues always possess higher mutabilities when they exist on the exterior than in the interior. In the preceding subsection the tendency of less accessible residues to be more invariant was observed to be common to all protein families studied. Here the same tendency is found to be common virtually to all types of amino acid residues.

TABLE 4

The matrices of the accepted point mutations in the interior and exterior[a]



[a] The upper right half is the matrix for the interior, and the lower left half for the exterior. The diagonal elements are shown in the last column and in the last row for the interior and for the exterior, respectively. The empty shows zero elements. These matrices are evaluated, using cytochrome c, myoglobin and hemoglobin chains α, β. The matrix elements corresponding to more than one base change are reduced to zero.

C. Correlation between mutability and polarity

To see whether there is any correlation between the mutabilities and polarities of amino acid residues, the mutabilities in the interior and exterior are plotted against the polarity of 20 types of amino acid residues in Fig. 4(a) and (b), respectively. The mutability of amino acid residues shows no clear correlation with its polarity both in the interior and on the exterior. This result is consistent with the finding of no correlation between polarity and variability in globins (Vogel & Zuckerkandl, 1972). The mutabilities of tryptophan, leucine and lysine are fairly small but the ones of isoleucine, serine, valine, methionine and alanine are fairly large both in the interior and on the exterior.

D. Evolutionary change

The product of the mutability and frequency of each amino acid is calculated, to see the contribution from each type of amino acid to the evolutionary changes defined by eqn. (3) or (4). The difference in the product of the mutability and frequency between the exterior and interior sites, $m_{ex,i}f_{ex,i} - m_{in,i}f_{in,i}$, is plotted in Fig. 5 for each type of amino acid residue against its polarity. The exterior-interior difference in this product shows clear correlation (the correlation coefficient is 0.87) with the polarity of the amino acid residue. The positive contribution from the polar residues and the negative contributions from the nonpolar residues to the difference in evolutionary changes, $\mu_{ex} - \mu_{in}$, which are seen in Fig. 5, are mainly due to the higher frequency of the polar residues and the lower frequency of the nonpolar residues in the exterior than interior, because the correlation between polarity and mutability is not dis-

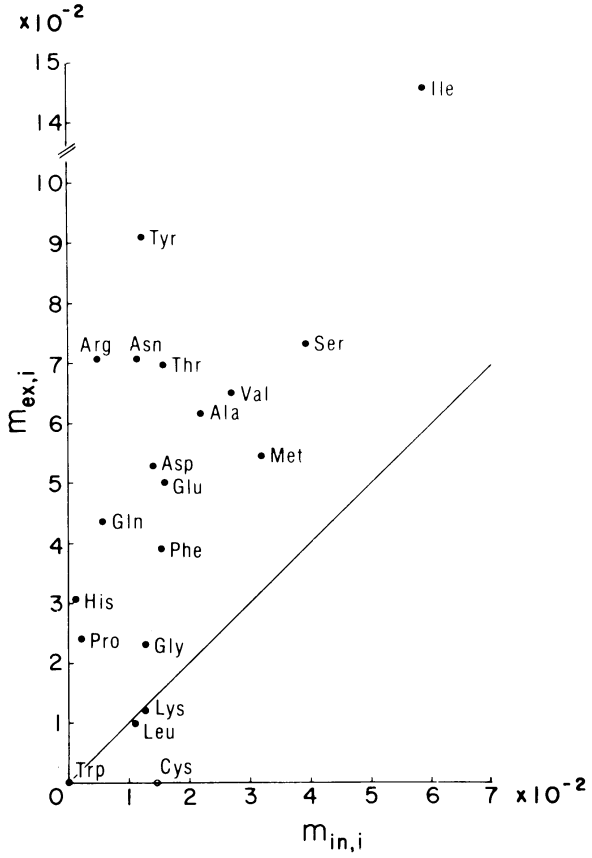


FIGURE 3

Mutabilities of 20 amino acid residues in the interior (abscissa) and on the exterior (ordinate) are plotted. Solid line shows a hypothetical case where the mutabilities in the interior and those on the exterior have the same value. Open circle for cysteine shows that this residue does not appear on the exterior of the proteins along the phylogenetic trees studied.

tinctive (the correlation coefficients are -0.30 and -0.09 in the interior and exterior, respectively) and the exterior residues possess higher mutabilities than the interior residues (Fig. 3) irrespective of their polarities. In other words, the correlation between the contribution to the difference in the evolutionary changes ($\mu_{ex} - \mu_{in}$) and the polarity are in fact due to the higher frequency of the polar residues on the exterior than interior but not due to higher mutabilities of polar residues than other residues.

We now want to see if the above observations made on "an average protein" hold in each individual protein family. For this purpose we have to lump amino acid residues into three

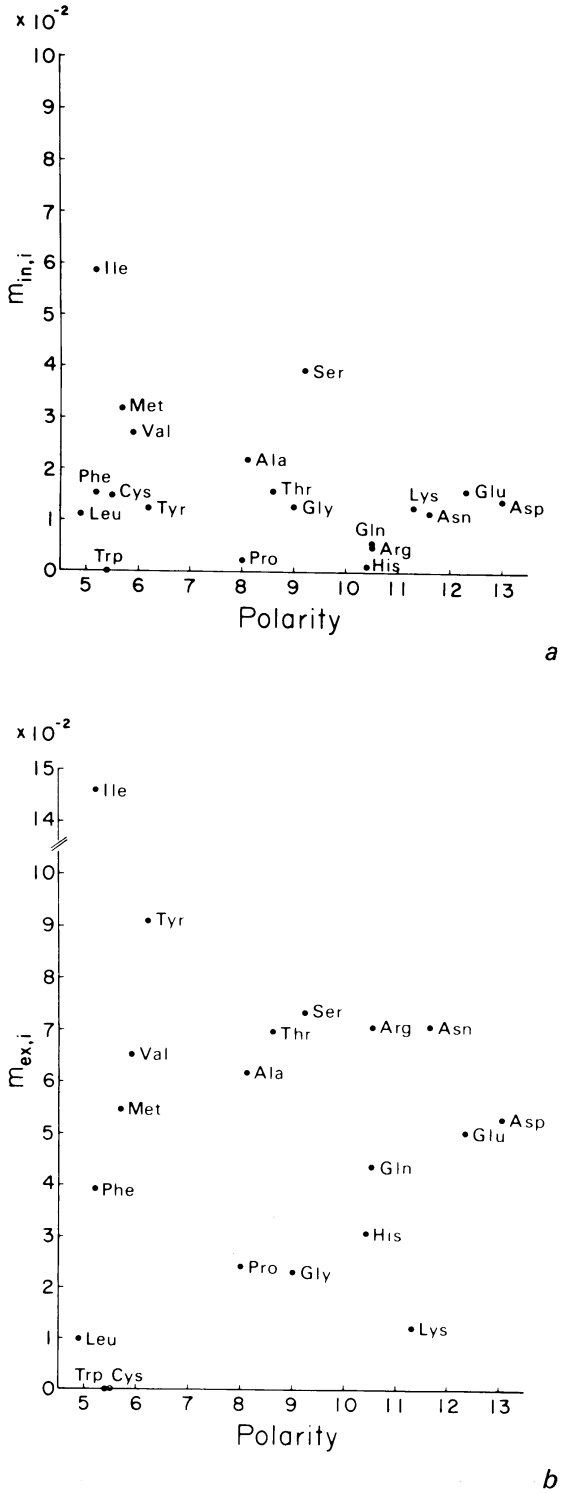


FIGURE 4

(a) Mutabilities of 20 amino acid residues in the interior are plotted against their polarity. (b) The same to (a) but for the residues on the exterior.

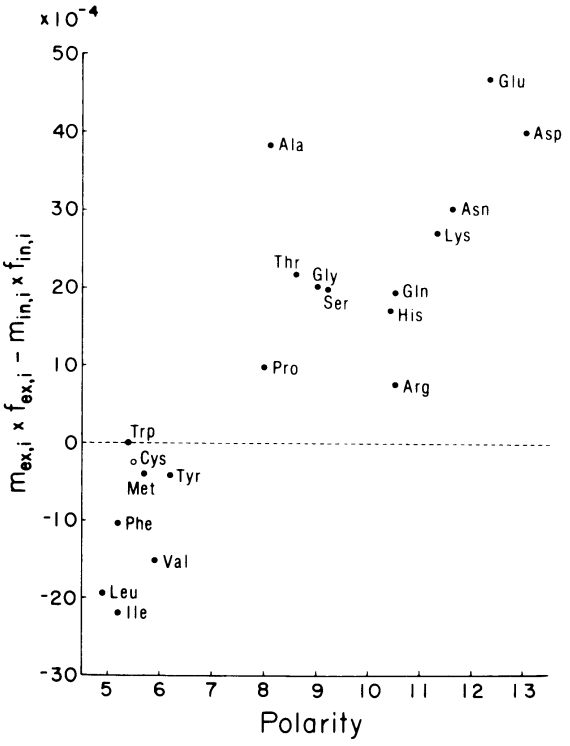


FIGURE 5

Contribution to the evolutionary change on the exterior relative to the interior from each amino acid residue, $m_{ex,i}f_{ex,i} - m_{in,i}f_{in,i}$, is plotted against its polarity.

groups as in Table 2 in order to avoid statistical fluctuations due to paucity of data. The contribution to the evolutionary changes μ_{in} and μ_{ex} from each of three amino acid groups, the mutability $m^*_{a,i}$ and the frequency $f^*_{a,i}$ are calculated for each protein family and shown in Table 5(a–d). The mutability of the amino acid group in the whole molecule is not correlated with its polarity. The highest mutability of the weak polar group is common in four proteins studied. In cytochrome c and myoglobin, the mutabilities of amino acid groups show weak negative correlation with their polarity both in the interior and exterior. However, such correlation is not observed in hemoglobin α and β chains; the weak polar group shows the highest mutability both in the interior and exterior. Thus, the higher mutabilities described by Zuckerkandl & Pauling (1965) of polar residues are not a common feature in protein evolution. However, it is a common character in the protein families studied that each amino acid group always shows higher mutability on the exterior than in the interior, irrespective of its polarity. The ratios of the evolutionary change on the exterior μ_{ex} to that in the interior μ_{in} are shown in Table 6 for cytochrome c, myoglobin and hemoglobin α and β chains. The ratio ranges from 1.8 to 3.1 and the averaged value over the proteins is 2.2. In other words, the probability of the amino acid substitution on the exterior is 2.2 times larger than that in the interior for the hypothetical average protein.

E. Replacements within and between polar, weak polar and nonpolar amino acid groups

So far we have discussed mutabilities of amino acid residues regardless of the type of substituting amino acid residues. We now study the frequency of replacements between various types of amino acid residues. Because of the paucity of the data we can discuss only "an average protein family" and only by lumping amino acid residues into the three groups of Table 2. The accepted point mutation matrices in Table 4 are now reduced into 3 x 3 matrices corresponding to the classification of amino acid residues into the three groups. The results are shown in Fig. 6. As pointed out already, the replacements are very conservative, i.e. the most probable replacement is that within the same group both in the interior and exterior. The averaged probabilities of the replacements within the same group are 0.88 and 0.76 in the interior and exterior, respectively, i.e. the substitutions in the interior tend to be slightly more conservative than those on the exterior. The weak polar group shows a unique pattern of replacements; i.e. in the interior it has higher probability to be substituted by nonpolar group than by polar group, but on the exterior it has higher probability to be substituted by polar group than by nonpolar group. This characteristic pattern of the replacements indicates that the weak polar group has a bifacial character, i.e. it behaves like polar group on the exterior and like nonpolar group in the interior.

When a nonpolar residue at a certain site in the interior is replaced by a weak polar residue, the weak polar residue has higher probability to be substituted by weak polar or nonpolar residues than by a polar residue. Similarly, when a

TABLE 5

Mutability m, frequency f* and their product (evolutionary change) of three amino acid groups in:*

Group	Interior			Exterior			Whole molecule		
	m*	f*	m* · f*	m*	f*	m* · f*	m*	f*	m* · f*
(a) cytochrome c									
Polar	0.0116	0.151	0.0018	0.0276	0.582	0.0161	0.025	0.38	0.0093
Weak polar	0.0139	0.341	0.0047	0.0372	0.342	0.0127	0.026	0.34	0.0090
Nonpolar	0.0206	0.508	0.0105	0.0442	0.076	0.0034	0.024	0.28	0.0067
Total			0.0170			0.0322			0.0250
(b) myoglobin									
Polar	0.0086	0.184	0.0016	0.0391	0.637	0.0249	0.0323	0.412	0.0133
Weak polar	0.0145	0.250	0.0036	0.0469	0.316	0.0148	0.0327	0.283	0.0093
Nonpolar	0.0229	0.565	0.0129	0.1005	0.047	0.0047	0.0289	0.305	0.0088
Total			0.0181			0.0444			0.0314
(c) hemoglobin α									
Polar	0.0082	0.157	0.0013	0.0257	0.522	0.0134	0.0199	0.295	0.0059
Weak polar	0.0253	0.343	0.0087	0.0446	0.442	0.0197	0.0338	0.381	0.0129
Nonpolar	0.0169	0.500	0.0084	0.0118	0.0357	0.0004	0.0167	0.324	0.0054
Total			0.0184			0.0335			0.0242
(d) hemoglobin β									
Polar	0.0042	0.203	0.0009	0.0480	0.568	0.0273	0.0315	0.339	0.0107
Weak polar	0.0220	0.245	0.0054	0.0630	0.395	0.0249	0.0420	0.301	0.0126
Nonpolar	0.0195	0.552	0.0108	0.0390	0.037	0.0014	0.0202	0.360	0.0073
Total			0.0171			0.0536			0.0306

TABLE 6

Ratio of the evolutionary change on the exterior to that in the interior

Protein	Ratio of evolutionary change (μ_{ex}/μ_{in})
Cytochrome c	1.9
Myoglobin	2.4
Hemoglobin α[a]	1.8
Hemoglobin β[a]	3.1
Total proteins	2.2

[a] The tetrameric state consisting of two α chains and two β chains is used to classify the residues into the exterior and interior.

polar residue at a certain site on the exterior is replaced by a weak polar residue, the weak polar residue has higher probability to be substituted by weak polar or polar residues than by a nonpolar residue. Thus, the majority of replacements between different groups occurs in such a way as to restore the polarity through the "buffer" role of the weak polar group.

The bifacial character disclosed and discussed above can be understood as a consequence of the conservatism of the amino acid replacements. The conservatism recognized strongly by the markedly larger amino acid replacements

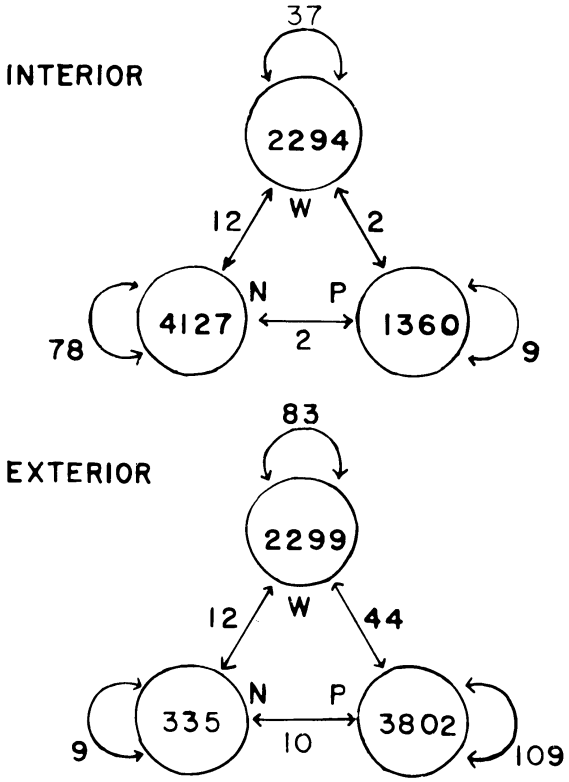


FIGURE 6

Illustration of reduced 3 × 3 accepted point mutation matrices. The number of non-substituting pairs of amino acid residues in each group is given in a corresponding circle. Numbers of substitutions within a group are shown on arrows flowing within a group. Numbers of inter-group substitutions are given on arrows connecting two circles. Off-diagonal elements of the reduced 3 × 3 matrix are half the number on arrows connecting two circles.

within the three groups than inter-group replacements. Slightly less stringent conservatism is recognized by the greater number of inter-group replacements between closer groups than distant groups. This fact together with the fact that mutability of 20 amino acid residues does not have clear correlation with their polarity, and that there are more nonpolar residues in the interior and polar residues on the exterior, explains the bifacial character of weak polar residues.

CONCLUSIONS AND DISCUSSION

In the pioneering work in the field of the molecular evolution Zuckerkandl & Pauling (1965) said, "Since in any one 'edition' of the globin chain the majority of the charged sites, along with other polar sites, may be expected to be at the outside of the molecule, in conformity with Kendrew's finding on myoglobin, and since charged sites and other polar sites are more variable, on the average, than apolar sites (with the exception of glycine, alanine, and proline, . . .), we may venture the generalization that the outside of the globin molecule, and perhaps of globular proteins in general, is more variable than the inside. . ." The only X-ray data available for them when they published the above paper was that of myoglobin. Now, with more data available, we conducted in this paper a systematic and quantitative study on correlations between variability, polarity and exteriority of residues in several sets of homologous proteins.

The extent of exposure of an amino acid residue to the surface of the molecule is defined quantitatively based on the static accessible surface area. This measure of exposure of amino acid residues is more refined than the simple method employed by us earlier (Gō & Miyazawa, 1978*a*), i.e. the measure of exposure based on the number of c^{α} atoms existing within a certain distance of the c^{α} atom of the amino acid residue in question.

Zuckerkandl & Pauling's ventured generalization that the exterior residues are more variable than the interior residues was indeed confirmed in this paper in eight sets of protein families. These families include those with and without a prosthetic group. Zuckerkandl (1976) discussed the concept of contact functions and classified them into two: 1) specific interactions either with other polypeptide chains or with other parts of the same polypeptide chain and 2) interactions with invariable molecular partners, such as prosthetic groups and cofactors. The fact confirmed here indicates that both types of contacts do indeed reduce the variability of residues in the molecular evolution. One may expect that the reason for the low variability of the residues in the interior of the proteins is more restriction on the volume changes accompanied by substitutions in the interior than that on the exterior. However, we reported that volume (and polarity) changes accompanied by the substitutions are not signi-

ficantly different in the interior and on the exterior (Gō & Miyazawa, 1978*b*).

The greater variability of exterior residues than interior residues is general not only in different protein families but also in different types of amino acid residues. Virtually all 20 types of amino acid residues are found to be more variable on the exterior than in the interior.

It is well established that there are more polar and nonpolar residues on exterior and interior of globular proteins, respectively. When combined with the correlation between mutability and exteriority confirmed here, this leads naturally to an anticipation that polar residues are more variable than nonpolar residues. In a sentence cited at the top of this section Zuckerkandl & Pauling noted that in globin polar sites are more variable. The question is whether or not polar residues are intrinsically more variable. In order to study this problem residues sites are classified into exterior and interior, and mutability of each type of amino acid residues is calculated in each of the interior and exterior. No correlation was observed between polarity and mutability in each of the exterior and interior sites. This means that polar residues are not *intrinsically* any more variable than nonpolar residues.

ACKNOWLEDGMENTS

One of the authors (M.G.) would like to dedicate this contribution to Professor N. Saito on the occasion of his sixtieth birthday. The authors are indebted to Protein Data Bank in Cambridge and Brookhaven National Laboratory for sending them the atomic coordinates of the proteins. They acknowledge Mr. H. Mizuno and Dr. N. Gō for allowing them to use the computer program for the accessible surface area and Professor H. Matsuda for his encouragement during the work. Calculation in the present work was carried out by FACŌM 230-75 at Computer Center, Kyushu University. This work was supported partially by grants-in-aid from the Ministry of Education, Japan.

REFERENCES

Aboderin, A.A. (1971) *Intern. J. Biochem.* **2**, 537–544

Alden, R.A., Birktoft, J.J., Kraut, J., Roberts, J.D. & Wright, C.S. (1971) *Biochem. Biophys. Res. Commun.* **45**, 337–344

Beintema, J.J., Gaastra, W., Lenstra, J.A., Welling, G.W. & Fitch, W.M. (1977) *J. Mol. Evol.* **10**, 49–71

Birktoft, J.J. & Blow, D.M. (1972) *J. Mol. Biol.* **68**, 187–240

Blake, C.C.F. & Swan, I.D.A. (1971) *Nature New Biol.* **232**, 12–15

Bolton, W. & Perutz, M.F. (1970) *Nature* **228**, 551–552

Bondi, A. (1968) *Molecular Crystals, Liquids and Glasses*, John Wiley & Sons, New York

Boulter, D., Ramshaw, J.A.M., Thompson, E.W., Richardson, M. & Brown, R.H. (1972) *Proc. R. Soc. Lond.* **B181**, 441–455

Chothia, C. (1975) *Nature* **254**, 304–308

Chothia, C. (1976) *J. Mol. Biol.* **105**, 1–14

Croft, L.R. (1973) *Handbook of Protein Sequences*, Joynson-Bruvvers, Oxford

Croft, L.R. (1974) *Handbook of Protein Sequences*, Suppl. A, Joynson-Bruvvers, Oxford

Croft, L.R. (1976) *Handbook of Protein Sequences*, Suppl. B, Joynson-Bruvvers, Oxford

Dayhoff, M.O., Park, C.M. & McLaughlin, P.J. (1972*a*) Building a phylogenetic tree: cytochrome c. In: *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.), p. 7, The National Biomedical Research Foundation, Maryland

Dayhoff, M.O., Eck, R.V. & Park, C.M. (1972*b*) A model of evolutionary changes in proteins. In: *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.), p. 89, The National Biomedical Research Foundation, Maryland

Dayhoff, M.O., Hunt, L.T., McLaughlin, P.J. & Jones, D.D. (1972*c*) Gene duplications in evolution: the globins. In: *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.), p. 17, The National Biomedical Research Foundation, Maryland

Dayhoff, M.O., Hunt, L.T., McLaughlin, P.J. & Barker, W.C. (1972*d*) Data section. In: *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.), p. D–1. The National Biomedical Research Foundation, Maryland

Diamond, R. (1974) *J. Mol. Biol.* **82**, 371–391

Dickerson, R.E., Takano, T., Eisenberg, D., Kallai, O.B., Samson, L., Cooper, A. & Margoliash, E. (1971) *J. Biol. Chem.* **246**, 1511–1535

Fletterick, R.J. & Wyckoff, H.W. (1975) *Acta Crystallogr.*, Sect. A **31**, 698–700

Gō, M. & Miyazawa, S. (1978*a*) Relationship between mutability and location of amino acid residues in the three-dimensional structure of proteins. In: *Evolution of Protein Molecules* (Matsubara, H. & Yamanaka, T., eds.), p. 33, Japan Scientific Societies Press, Tokyo

Gō, M. & Miyazawa, S. (1978*b*) *Int. J. Peptide Protein Res.* **12**, 237–241

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

Hendrickson, W.A. & Love, W.E. (1971) *Nature New Biology* **232,** 197–203

Lee, B. & Richards, F.M. (1971) *J. Mol. Biol.* **55,** 379–400

Momany, F.A., McGuire, R.F., Burgess, A.W. & Scheraga, H.A. (1975) *J. Phys. Chem.* **79,** 2361–2381

Pauling, L.C. (1960) *The Nature of Chemical Bond,* 3rd edn., Cornell University Press, New York

Perutz, M.F., Kendrew, J.C. & Watson, H.C. (1965) *J. Mol. Biol.* **13,** 669–678

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

Romero-Herrera, A.E., Lehmann, H., Castillo, O., Joysey, K.A. & Friday, A.E. (1976) *Nature* **261,** 162–164

Shotton, D.M. & Watson, H.C. (1970) *Nature* **225,** 811–816

Shrake, A. & Rupley, J.A. (1973) *J. Mol. Biol.* **79,** 351–371

Swanson, R., Trus, B.L., Mandel, N., Mandel, G., Kallai, O.B. & Dickerson, R.E. (1977) *J. Biol. Chem.* **252,** 759–775

Takano, T. (1977) *J. Mol. Biol.* **110,** 569–584

Vogel, H. & Zuckerkandl, E. (1972) The evolution of polarity relations in globins. In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability,* (Lecam, L.M., Neyman, J. & Scott, E., eds.), p. 155, University of California Press, Berkeley and Los Angeles

Woese, C.R. (1973) *Naturwissenschaften* **60,** 447–459

Wright, C.S., Alden, R.A. & Kraut, J. (1969) *Nature* **221,** 235–242

Zuckerkandl, E. (1976) *J. Mol. Evol.* **7,** 167–183

Zuckerkandl, E. & Pauling, L. (1965) Evolutionary divergence and convergence in proteins. In: *Evolving Genes and Proteins* (Bryson, V. & Vogel, H.J., eds.), p. 97, Academic Press, New York

Address:

Dr. *Mitiko Gō*
Department of Biology
Faculty of Science
Kyushu University
Fukuoka 812
Japan