THE EVOLUTION OF POLARITY RELATIONS IN GLOBINS

HELMUT VOGEL and EMILE ZUCKERKANDL
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE, MONTPELLIER

1. Introduction

The well known correlation between the hydrophobicity of amino acid residues and their position in the interior of protein molecules was predicted long ago (Kauzmann [9]) and was more recently verified by X-ray diffraction studies (Kendrew [10]).

At the Rutgers conference in 1964, the examination of substitution patterns in globins showed that "there are more sites that seem to specialize in carrying residues fit for apolar bonding than any other sites at which the residues found are limited to one given chemical category." Apolar bonding, it was said, "may be the most specifically determined business of molecular sites in globular proteins" and, further: "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" (Zuckerkandl and Pauling [19]).

This presumption was based in part on the already available knowledge that the majority of the polar amino acid residues are on the outside of the globins [10] and in part on the observation [19] that charged sites and other polar sites are more variable, on the average, than apolar sites, with the exception notably of glycine and of alanine sites.

In 1967, it was shown, by counting the minimal number of base substitutions on a deduced molecular phylogenetic tree, that maximum variability of sites coincided mostly with exteriority of the sites. This was established for a stretch representing two thirds of the globin chain (Derancourt, Lebor, and Zuckerkandl [5]).

On the basis of their study of the structural and functional implications of amino acid substitutions found in abnormal human hemoglobins, Perutz and Lehmann concluded in 1968 [12] that, whereas the conditions of amino acid substitutions are functionally restrictive in the interior of the hemoglobin molecule, they are liberal at its surface. The surface should be more variable than the interior, as Epstein [7] has also shown.

It was therefore unexpected to find (Zuckerkandl, Derancourt, and Vogel [18]), on the basis of an inventory of the different types of probable amino acid

This work was supported in part by grants from the Délégation Générale de la Recherche Scientifique et Technique and from the Fondation pour la Recherche Médicale Française.

substitutions during evolution, that in globins, and also in cytochrome *c*, apolar residues are lost and gained practically as frequently as polar residues. Alanine and glycine residues were not classified as apolar and were considered separately, as by Epstein [6], because they are found frequently at the surface as well as in the interior of globular proteins (Perutz, Kendrew, and Watson [11]). If these results are correct, then something must be incorrect about the assumption that a good correlation prevails between polarity, variability, and exteriority.

To reinvestigate this question, data on variability at each molecular site during globin evolution were obtained with the help of a Beckman 816 computer for a set of 39 chains. They include chains such as the frog β chain and the *Chironomus* (insect) chain that give to the comparisons and deductions a somewhat wider or better established evolutionary scope than was obtained heretofore.

Data on exteriority are readily available through the list of "internal" residues given by Perutz, Kendrew, and Watson in their 1965 paper (Table 4), and in the sinusoidal curve in Figure 1 of the same paper. Data on the sites involved in interchain contacts in oxy- and in deoxyhemoglobin are, in turn, available in recent papers by Perutz and Lehmann [13], and Bolton and Perutz [1]. These sites are exterior from the point of view of the tertiary globin structure, but interior (if the notion is used loosely) with respect to the quaternary structure of a tetrahemic hemoglobin molecule. Perutz's results are valid for the case of horse hemoglobin. It may be that during the evolution of tetrahemic hemoglobins the contacts between chains did not consistently involve the same sites. From the evolutionary point of view, whatever information is derived from the data on interchain contacts in horse hemoglobin is therefore only a probable approximation.

For an evaluation of polarity we adopted Woese's [15] polarity scale, based on his "polar requirement" index.

The present discussion is based largely on the results of Perutz and co-workers from which most advances in the field are derived and to whom we are so greatly indebted.

The list of globin chains used in the present computations is given in the legend to Table II.

In earlier work (Zuckerkandl and Pauling [19]), as in part of this paper, a concept of "variability" is used, that may be termed "comparative variability," in contrast to another concept, "evolutionary variability," that will be used mainly in the second part of this paper.

Comparative variability refers to a set of sites characterized as "sites for amino acid i" or as "sites for a subset A of amino acids." A site is called "site for amino acid i" if and only if amino acid i is found there in at least one known chain of the class of proteins under consideration. Evidently, a given site can thus be assigned to several amino acids. A site is called "site for the subset A" if and only if it is a site for at least one amino acid of the subset A. For instance, a site is called a "charged site" if it is a site for aspartic acid, glutamic acid, arginine, *or* lysine. (Histidine may or may not be counted among the charged amino acids). Comparative variability then refers to the number of different types of amino acids that occur at a site for amino acid i or at a site for a subset A, averaged over all those sites, and decremented by one.

Evolutionary variability refers only to a given site and is defined as the number of times that site was subject to amino acid substitution during evolution of the class of proteins under consideration. It has been customary to try to minimize the hypothetical element in this concept by using the minimum number of those substitutions that can be reconciled with the phylogenetic tree as determined from the topological analysis using the chains as a whole.

In the ideal case (no back mutations, only one step mutations resulting in amino acid substitutions), both variabilities should be numerically identical. Both back mutations and the attempt to account for two and three step mutations by counting them correspondingly often increase evolutionary variability. Since the tendency for back mutations and perhaps also for more-step mutations might vary from site to site, one cannot expect too close a correspondence between the two concepts of variability.

After presenting an overall picture of the variations of polarity in globins, we shall compare contemporary globin chains among themselves for numbers of types of substitutions (what and how many kinds of amino acids can be accommodated at different sites) and in a later section use the figures for the presumed evolutionary variability at each site as deduced from molecular phylogenetic trees (Zuckerkandl, Derancourt, and Vogel [18]) (number of times each residue has been substituted within the sector of evolution under consideration).

2. Variations of polarity in globins

The overall compositional polarity (the mean polarity per amino acid residue) seems to vary very little among globular proteins. For the sample given in Table I, this value is confined between 7.0 and 7.9, that is, between the values for alanine and glycine.

The constancy in the overall polarity of all hemoglobins (Table II) hides a considerable variation that becomes apparent as different corresponding stretches of different globin molecules are compared among themselves. Such variations relate to means obtained over stretches as long as, for instance, helix G (19 amino acids). When means characterizing stretches of significant length vary considerably, it would be surprising that the constancy of mean polarity of whole chains be a random effect and natural selection as the cause of such constancy is more probable. The existence of a tendency to preserve overall polarity in different groups of globin chains has been demonstrated (Epstein [7], H. Vogel [15]).

The modulation of the sectional mean polarities forms a "melody" charac-

TABLE I

MEAN POLARITY PER MOLECULAR SITE FOR DIFFERENT GLOBULAR PROTEINS

The polarity index used is Woese's [15] "polar requirement."

	$\bar{P}$
Hemoglobin	
human α chain	7.44
human β chain	7.53
Myoglobin, sperm whale	7.70
Ferredoxin	7.43
Ribonuclease, bovine	7.71
Tobacco mosaic virus, strain *Bulgare*	7.43
Chymotrypsinogen A, bovine	7.64
Glucagon, bovine	7.71
Cytochrome *c*, human	7.92

teristic of each type of chain. In Figure 1 the same values, as well as those for some individual chains, are represented in a notation in which the height of each note is proportional to the average "polar requirement" value and its duration is in rough relation to the length of the molecular section.

If the overall polarity of chains is considered, myoglobins are only slightly more polar than other globin chains (Tables I and II). Major differences appear however in individual helical and nonhelical regions. In comparison with α and non-α chains of hemoglobins, myoglobins are more polar along regions C, CD, D, EF, and H. Consideration of the "melodies" of Figure 1 shows that α and non-α chains differ from each other slightly, but significantly, in some sections. The patterns of the β chains and of the other non-α chains of mammalian hemoglobins are very similar. The lemur β chain is an exception. It is peculiar in many respects and in part quite different from both the human β and γ chains. It may well be the result of a distinct gene duplication (Zuckerkandl [17]). Carp α and frog β chains have a polarity line different from that of the corresponding mammalian chains, though the frog chain is recognizable as a β chain in terms of the mammalian pattern. For part of its polarity pattern, the dog α chain also is peculiar, whereas the dog β chain has a normal β chain pattern.

The stretches with highest polarity are seen to occur in the globin molecules that do not associate to form tetrahemic structures, namely in the myoglobins, in the lamprey chain, and in the *Chironomus* III chain. The *mean* polarity of the lamprey and *Chironomus* chains, however, does not differ from that of other hemoglobin chains.

The polarity of helices A and B is relatively stable, A being nearly always more polar than B. Helix C has a low polarity except in the monohemic globins. The evolution of tetrahemic globins may thus have required a lowering of the polarity of helix C. Helices E and F enclose the heme group. The polarity of E is not highly variable, that of F more so. Helix E is more polar than F, except

TABLE II

Variations in Polarity along the Globin Chains

Polarity indices as in Table I. Helical and interhelical sections as defined by Perutz, Kendrew, and Watson [11]. Except when indicated otherwise, the values represent the means of all chains of the type considered. The following globin chains were used in the present work: (1) α *chains*—human, *Rhesus*, dog (Jones [8]), mouse, rabbit, horse, pig, bovine, sheep A, goat A, llama, kangaroo, carp; (2) β *chains*—human, *Rhesus*, lemur, dog (Jones [8]), rabbit, horse, pig, llama, bovine, sheep A, sheep B, sheep C, barbary sheep, goat A, kangaroo, frog; (3) *others*—human γ, human δ, sheep fetal, bovine fetal, lamprey, *Chironomus* III (Buse, Braig, and Braunitzer [2]), myoglobins of sperm whale, horse, cattle, and kangaroo.

Sources: if not stated otherwise, Dayhoff [4].

Molecular section	NA	A	B	C	CD	D	E	EF	F	FG	G	GH	H	HC	
Number of residues per section in human β chain	3	16	16	7	8	7	20	8	9	5	19	5	24		Total
α chains	5.25	8.00	7.70	6.74	7.11	8.01	7.72	8.19	7.37	7.32	6.82	7.30	7.11		7.4 ± 0.1
β chains	6.00	7.90	7.46	6.54	7.75	7.45	7.91	8.29	6.98	8.15	7.29	7.69	7.31		7.5 ± 0.1
γ, δ, "fetal β"	5.90	7.85	7.43	6.58	7.73	7.13	7.97	8.45	6.95	8.27	7.63	7.70	7.30		7.5
Myoglobins	6.00	7.87	7.73	8.80	8.23	9.03	7.53	9.60	7.20	8.00	6.93	7.63	7.97	7.80	7.8 ± 0.2
Lamprey	7.0	7.4	7.3	7.6	7.0	9.3	7.9	8.9	8.2	7.3	6.4	7.4	7.1		7.5
Chironomus III	4.9	~8.3	7.0	7.8	7.2	8.7	~6.9	10.3	8.3	8.4	7.9	8.2	6.4		7.5

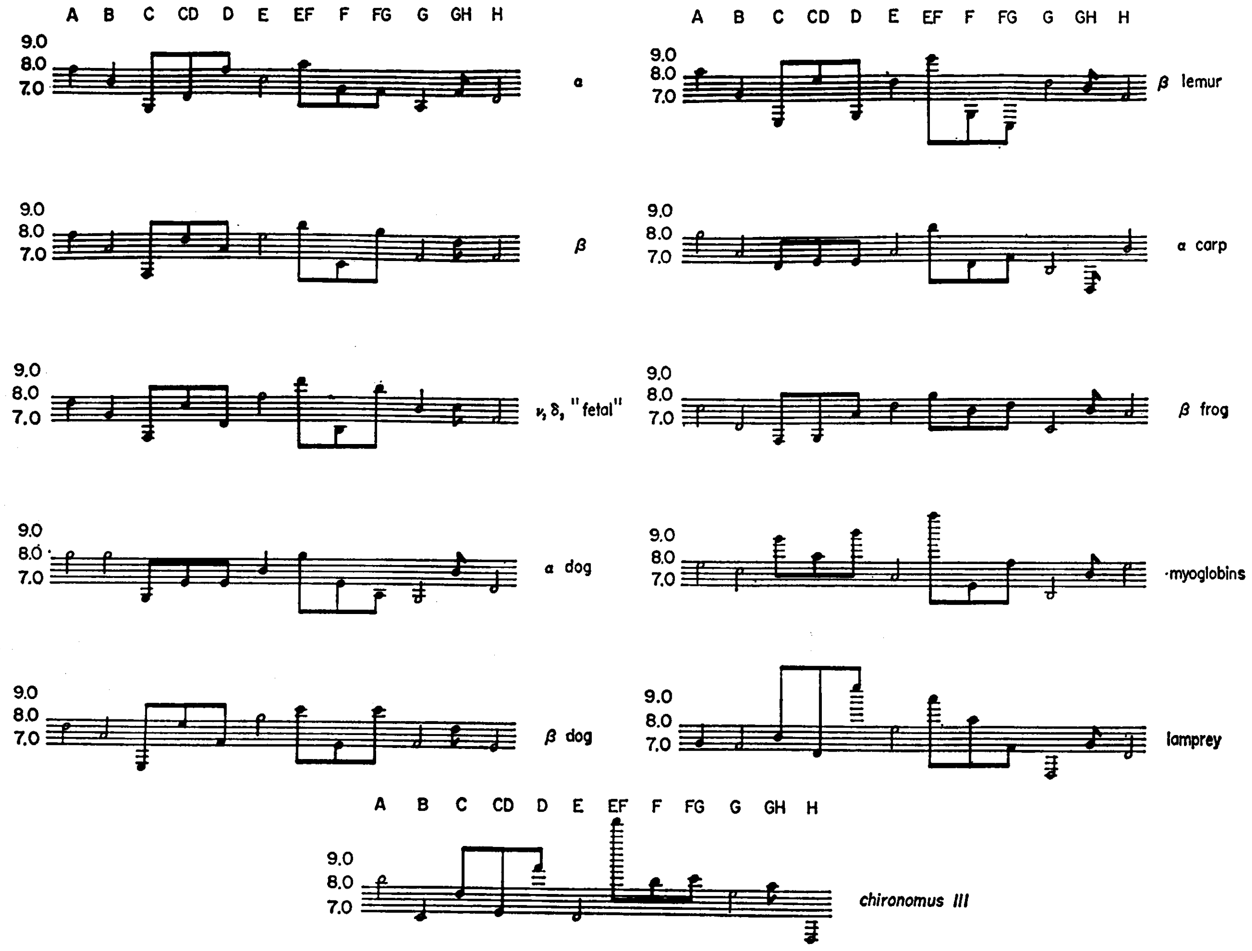

FIGURE 1

Change in "polar requirement" along globin chains (see Table II).

in lamprey and *Chironomus*, and G is nearly always less polar in α chains than in non-α chains, less also in the lamprey chain. Helix H displays a very stable polarity, that is, relatively high only in the myoglobins and low in the *Chironomus* chain. With the exception of the frog β chain, H is more polar than G in α chains, but not in non-α chains. The myoglobins and the lamprey chain behave in this respect like α chains.

As to interhelical sections, the polarity of the CD segment is lower in α chains, higher in β chains, the β chain of frog being an exception. The lamprey and *Chironomus* chains are in this respect like an α chain, the myoglobins like β chains. Section EF always has a relatively high polarity, but in the monohemic globins higher than in the others.

On the whole, the polarity line of the lamprey chain resembles that of an α chain more than that of a non-α chain. That of myoglobin is quite distinct from either.

Further data are needed for accurately picturing such apparent evolutionary trends and their functional significance.

3. Types of amino acids accommodated at different molecular sites

3.1. *"Internal" sites.* Perutz and his co-workers [11] listed 33 internal sites in globin, defined as sites cut off from the surrounding water. Of the 33, only three were then known to accommodate not only apolar, but also polar residues, namely serine or threonine. Internal sites able to accommodate polar residues now number fourteen, on the assumption that sites listed as internal by Perutz are so in all globin chains and that the homologies, taking into account gaps, have been correctly established. *Thus, 40 per cent or more of the "internal" sites in the globin molecule do not exclude polar residues at one time or another.* In nine of the cases, the polar amino acid found is serine or threonine. Other polar amino acids include glutamic acid or glutamine, aspartic acid, asparagine, and histidine (Table III).

A given chain usually accommodates no more than one polar residue in its interior. When an alignment of residues is such that a number of polar amino acids fall on molecular positions listed as interior, the alignment is likely to be faulty, such as that given in the *Handbook of Biochemistry* [14] for the lamprey chain, where at least nine polar residues coincide with "internal" sites. Apolarity at a maximum of internal sites in a given globin chain, and probably in globular proteins in general, may thus be used as a supplementary criterion for the correctness of amino acid alignments.

Internal polar sites are mostly those at which also either glycine, or alanine, or both have been found (9 cases out of 14; random expectation 5.1). Exceptions include *Chironomus* three times, which again raises the question of the strict applicability of the horse data. Conversely, at internal sites limited to apolar residues, glycine or alanine usually do not occur. (They do so only in 3 cases out of 19; random expectation 6.9).

TABLE III

"Comparative Variability" at Internal Globin Sites

Site		Val	Leu	Ile	Met	Phe	Tyr	Trp	Ala	Gly	Cys	Others	Organisms in which the other residues occur
(a) Sites not Known to Accommodate Other than Apolar Residues													
A	8	+		+									
	12					+		+					
	15	+		+						+			
B	10	+	+										
	14		+			+							
CD	1					+							
	4					+		+					
D	5	+	+	+	+								
E	4	+	+	+		+							
	8								+	+			
	11	+		+									
	15	+	+	+		+							
	18		+	+					+	+			
G	5		+			+							
	8		+	+									
	12		+	+		+							
H	15	+	+	+		+							
	19		+	+	+								
	23				+		+						
(b) Sites Accommodating also Polar Residues													
A	11	+	+	+		+			+			Ser, Thr	*Chironomus* III; α dog
B	6								+	+		Pro, Gln	*Chironomus* I, III
	9	+		+					+	+		Thr	β *Propithecus*, lemur, γ man
	13	+	+		+	+						Thr	α dog
C	4			+					+	+		Thr	α, β man; Mb sperm whale

TABLE III (Continued)

Site		Val	Leu	Ile	Met	Phe	Tyr	Trp	Ala	Gly	Cys	Others	Organisms in which the other residues occur
(b) Sites Accommodating also Polar Residues (continued)													
E	12	+	+	+					+			Ser	α rabbit
	19	+	+	+	+					+		Thr	α rabbit
F	1	+	+		+	+	+					Asn, Ser	*Chironomus* III, α rabbit
FG	5	+		+								Asx	*Chironomus* III
G	11	+	+	+					+		+	His	α carp
	16	+	+	+	+							Ser	γ *Lemur fulvus*
H	8	+	+		+	+	+	+				Thr	β kangaroo
	11	+	+		+	+			+			His	β frog
	12	+	+			+			+			Thr	α dog
												Ser	lamprey

Thus, it appears that when a site has a strictly apolar function in the interior, not only glycine, as expected, but also alanine is unlikely to fill this function. *The presence of alanine at a site probably is the general sign that a number of functionally different amino acids can be accommodated at that site.*

Most of the internal sites not open to alanine or glycine will accommodate most of the other apolar amino acids.

The series of functionally typical apolar residues comprises valine, leucine, isoleucine, methionine, phenylalanine, and tyrosine. To these tryptophan may be added (Table III). Among the residues that occur at a site, when the two most distant from each other in the preceding series, sizewise, are considered, those in between will no doubt also be found in globin chains, if they are not already known. For instance, at H 12, where valine, leucine, and phenylalanine are found, isoleucine and methionine should also be found.

Tyrosine hardly occurs at strictly apolar residue sites, as the series of 20 such sites shows. With the exception of the special case of H 23, tyrosine has not been found at such sites. The inclusion of tyrosine among the "apolar" amino acids is warranted by its position on the polarity index scale. Yet, in fact, tyrosine is more frequently excluded from purely apolar than polar sites.

Tryptophan remains very rare; only five tryptophan sites are known in globins at this writing.

Internal sites at which polar amino acids occur are distinctly more variable than the set of sites at which no polar amino acids have been found yet. The mean number of different amino acids is 5.3 for the first group as against 2.5 for the second. Taking all internal sites together, there are two at which seven different residues have been found to date, five with six different residues, three with five different residues, and five with four different residues. About one half of the internal sites can thus accommodate as many different amino acid residues as, or more such residues than, the average external noncontact site (see Table VII below).

Thus, *at a number of interior sites, interiority does not imply a specially rigorous specification of tolerable amino acid residues.*

3.2. *Contact sites.* From a consideration of the data provided by Perutz, Muirhead, Cox, and Goaman [13] and by Bolton and Perutz [1] on contacts between chains in horse oxyhemoglobin and deoxyhemoglobin, it appears that thirteen out of 37 interchain contact sites, that is, one third of their number, are restricted to one type of chain.

Table IV gives a list of the amino acids found in different globin chains at sites characterized as contact sites for horse hemoglobin.

Table V shows the number of sites at which a given amino acid residue is actually involved in contacts (judging from the situation in horse hemoglobin) and the number of times the same residue occurs at different "contact sites" in globins existing as free protomeres. There is no significant shift in types of amino acids used at "contact sites" between chains that do and that do not associate to oligomeric molecules. *All kinds of amino acids are used at contact* *sites, apolar ones, polar hydrogen bond forming ones, charged ones.* The proportion of sites is rather similar for each of these three groups and it is not significantly different when, for each group, associating and nonassociating chains are compared.

Thus, there is, with one possible exception, nothing remarkable about the types of amino acids that are used for establishing interchain contact. Neither charged, nor hydrogen bond forming, nor even apolar residues are significantly favored for the formation of quaternary structure. This remark leaves open the question as to *which part* of a residue is actually used for making a contact. Perutz, Muirhead, Cox, and Goaman [13] have shown that the contacts are mostly made between the apolar sections of residues. For contacts between chains, it looks as though the protein moieties are "satisfied" as long as they find apolar groups of atoms that are sterically fit and don't "mind" what else there may be around on the residue provided there is no steric or no charge exclusion.

In the above comparison, proline, tryptophan, cysteine, and alanine have not been counted in the typical "apolar series," but listed separately, because each of these amino acids has in its way a peculiar behavior. It is striking that cysteine occurs at four contact sites in chains that actually make contact with others, whereas no cysteine residues occur at these sites in monohemic globins. Remarkably, three of the four cysteine sites are on helix G, in the region engaged in contacts $\alpha_1\beta_1$ (sites 11, 14, 18). Also, the fourth cysteine site, D 6, is listed for a $\alpha_1\beta_1$ contact. A fifth cysteine site, G 15, not listed as a contact site, is also in the region of helix G engaged in the $\alpha_1\beta_1$ contact. *It could be that this rare residue is sometimes specially selected for interchain contact function.* The total number of cysteine sites for all globins known at this writing is eight, with none for myoglobins, none for *Chironomus* globin, and one for the lamprey chain. The three cysteine sites not yet listed are E 16, F 9, and H 13. With the exception of G 11, none of the eight sites is classified as internal. The same site G 11 is, however, also given as a contact site. It can be concluded that cysteine apparently is not used for *internal* contacts. The only occurrence so far known of cysteine in monohemic molecules is in the lamprey chain, at site H 13 which is not a "contact site" in α or non-α chains of tetrahemic hemoglobins. The occurrence of cysteine at four if not five *interchain* contact sites thus seems to be quite significant.

Of 37 distinct contact sites, only nine appear as "invariant" in tetrahemic hemoglobins at the present writing. (Among these nine, site C 7 is also counted, although it has either phe or tyr in the non-α chains, since only tyr has been found in α chains and since only α chains are reported to be involved in contacts at that site).

However, if it is considered that the two types of chains may be engaged at a given contact site common to them in two different specific ways, contact sites common to both chains with residues that differ in the α and in the non-α chains, but are in either case unique, may also tentatively be listed among

TABLE IV

Residues at Interchain Contact Sites

Contact sites as defined for horse hemoglobin (see text). For globin chains from tetrahemic hemoglobins invariance is indicated by a dot when general, by a circle when only the α chains or the non-α chains have so far been found invariant, or when the invariant residue is not the same in both α and non-α chains. Under "contact in oxy- and deoxyhemoglobin," 1 stands for an $\alpha_1\beta_1$ contact, 2 for an $\alpha_1\beta_2$ contact ([12], [13], [1]). When only one of the two types of chains participates in interchain contact at a given site ([12], [13], [1]), the residues occurring at that site in the chain that does not participate in contact are between parentheses.

		Globin chain from tetrahemic hemoglobins						Monohemic globins		
		Invariance in α and/or non-α		Contact in hemoglobin						
Site				Oxy-	Deoxy-	Residues in α chain	Residues in non-α chain	Myoglobins	Lamprey	*Chironomus*
B	11			1		Gly; Asp, Glu	(Gly, Ala)	Ile	Val	Tyr
	12	·		1		Arg	Arg	Arg	Lys	Ala
	15			1		Gly, Ala; Leu; Thr, Glm; His	Val, Leu, Ilu	Lys; Thr	Thr	Lys
	16		○	1		Gly; Val; Ser	Val	Gly, Ser, Thr	Ser	Ala
C	1		○	1		Phe, Tyr	Tyr	His	Thr	Asp
	2	·		2	2	Pro; Lys	Pro	Pro	Pro	Pro
	3		○	2	2	Thr, Glm	Trp	Glu	Ala	Ser
	5	○		2	2	Lys	Ser, Glm; Arg	Leu	Glm	Met
	6	○	○	2	2	Thr	Arg	Glu	Glu	Ala
	7	·		2	2	Tyr	(Phe, Tyr)	Lys	Phe	Lys
CD	2				2	Pro, Ala	(Ser, Thr; Asp, Glu)	Asp	Pro	Pro, Thr
D	2			1		(Gly)	Pro, Ala	Glu	Ala	Leu
	6			1		(His)	Leu, Met; Cys	Lys	Lys	Lys
FG	3	·		2		Leu	(Leu)	His	Phe	Thr
	4	○	○	2	2	Arg	His	Lys	Glm	His
	5	·		2	2	Val	Val	Ile, Val	Val	Asx
G	1	·		2	2	Asp	Asp	Pro	Asp	Glx
	2	·		2	2	Pro	Pro	Ile, Val	Pro	Leu
	3			2	2	Ala, Val	Ala, Glu	Lys	Glm	Asx
	4	·		2		(Asn)	Asn	Tyr, Phe	Tyr	Asx
G	10		○	1		His, Asn	Asn	Glu, Asp	0	0
	11			1		His; Cys	(Val, Ilu)	Ala	0	Gly

TABLE IV (Continued)

Site		Invariance in α and/or non-α	Globin chain from tetrahemic hemoglobins Contact in hemoglobin Oxy-	 Deoxy-	 Residues in α chain	 Residues in non-α chain	Monohemic globins Myoglobins	 Lamprey	 *Chironomus*
	13		1		Val, Leu	(Ala, Val, Ilu)	Ile	0	Val
	14		1		Val; Ser	Val, Glu; Thr; Cys	His, Glm	0	Ser
	17		1		(Ala, Met)	Gly, Ala, Ser	His, Glm	0	Lys
	18		1		Ala, Val, Phe; Ser, Asn; Cys	His, Arg, Glu	Ser, Ala	0	Ala
GH	2	◦	1		Gly, Pro	Gly, Glm	Pro, Ala	Ile	0
	5	•	1		Phe	Phe	Phe	Thr	Phe
H	1		1		(Thr; Pro)	Ser, Thr	Gly, Ala	Val	Gly
	2	◦	1		Pro	Pro; Ilu	Ala	Ala	Ala
	3		1		(Ala, Asp, Glu)	Ala, Val; Asp, Asx, Glu; Glm; Pro	Asp	Ala	Glu
	5	◦ ◦	1		His	Glm	Glm	Asp	0
	6		1		Ala, Asp; Met	Ala, Val; His	Gly, Ala	Ala	Ala
	9	◦	1		Asp	Glu; Glm	Asn, Thr, Ser, Lys	Glu	Gly
	10	◦	salt bridge		Lys	Ala, Lys	Lys	Lys	Thr
	23	•	2	2	Tyr	Tyr	Tyr	Tyr	Met
	24			2	(Arg)	His, Arg	Lys	0	0

TABLE V

NUMBER OF CONTACT SITES AT WHICH DIFFERENT AMINO ACID RESIDUES OCCUR
Compiled from data in Table IV. Only substitutions in the chain that is actually involved in a contact (in the case of horse hemoglobin) are recorded.

	α and non-α chains (Tetrahemic globins)		Monohemic globins	
	Number of sites	Proportion of total site count %	Number of sites	Proportion of total site count %
Gly	4		5	
Ala	8		11	
Val	10		4	
Leu	6		3	
Ile	2		6	
Met	2		2	
Phe	3		3	
Tyr	3		3	
Total "apolar series"	26	26.7	21	23.0
Pro	7		5	
Trp	1		0	
Cys	4		0	
Ser	6		5	
Thr	5		8	
Asn	3		1	
Glm	7		4	
Total "H bond forming residues"	21	21.7	18	19.8
Asp	4		6	
Glu	5		6	
Lys	2		9	
Arg	6		1	
His	8		5	
Total "charged residues"	25	25.7	27	29.6
Asx	1		3	
Glx	0		1	

"invariant" sites. They may indeed be invariant for one type of contact in the α chain, and invariant for another type of contact in the β chain. For instance, contact site H 5 with histidine in all α chains and glutamine in all non-α chains may be considered invariant. By virtue of this point of view, the number of invariant contact sites rises to 12. Thus, at the present writing, no more than one third of the contact sites, and perhaps less, are to be considered invariant.

Invariance is much larger for the $\alpha_1\beta_2$ than for the $\alpha_1\beta_1$ contacts, as already pointed out by Perutz, and Muirhead, Cox, and Goaman [13]. Of 15 sites involved in $\alpha_1\beta_2$ contacts (oxy- and deoxyhemoglobin), six are variable on a type of chain basis (Table IV). Of 21 sites involved in $\alpha_1\beta_1$ contacts, 18 are variable (salt linkage forming sites are not considered here). The more stable $\alpha_1\beta_2$ contacts are involved in the mechanism of oxygenation, during which "the contact $\alpha_1\beta_2$ undergoes drastic changes as the two subunits turn relative to each other by 13°" (Bolton and Perutz [1]). Thus, each contact residue on one chain may have to adapt to more than one group of atoms, or set of groups of atoms in the other chain, namely, to one set in the oxygenated and to another set in the deoxygenated state. It is plausible that such multiple stereochemical fitting reduces the evolutionary variability of the residues involved. The surprise, if any, is that it does not reduce it more drastically. On the other hand, Bolton and Perutz [1] report that the contacts $\alpha_1\beta_1$ undergo only slight changes upon deoxygenation. It is the $\alpha_1\beta_1$ dimer that is found in free solution upon spontaneous or electrolyte induced dissociation of tetrahemic hemoglobins [13]. The contact residues functioning between these two subunits, which have to adapt to essentially one situation and not to two sterically different ones, are free to vary.

Since the association of globin protomers to oligomers adds restrictions to the changeability of a certain number of molecular sites, one might expect globin chains, when engaged in tetrahemic hemoglobins, to "evolve" slightly more slowly than myoglobin chains. This is not verified by the only example that at present allows a precise checking of this point. In cattle and horse the α, β, and myoglobin chains have been analyzed for their amino acid sequences. The number of differences between cattle and horse are 24 for the α chains, about 31 for the β chains, and only 18 for the myoglobins.

The following conclusion can be drawn from the preceding discussion. *The formation of quaternary structure in hemoglobins has led to the effect of a relative invariance of the residues at contact sites, primarily at the $\alpha_1\beta_2$ contact sites, while not requiring a change in polarity of the residues involved in contact.*

Thereby, this set of contact sites should not fit well into any correlation between polarity and variability that might otherwise exist.

4. Frequency of amino acid substitutions during evolution

4.1. *Correlation between polarity and evolutionary variability.* We shall now consider variability in terms of the number of times amino acid substitution presumably has occurred at each molecular site during evolution ("evolutionary variability").

If all sites are considered for all chains, no significant correlation is found between polarity and variability in globins (Table VI).

Likewise (Table VI), no significant correlation between polarity and variability is found when different molecular sections, namely, the helical and the interhelical regions, are examined separately. There is no essential difference in the correlation between helical and nonhelical sections of the molecules. (The only mildly significant correlation of a section, that of region CD with

TABLE VI

THE CORRELATION BETWEEN POLARITY AND EVOLUTIONARY VARIABILITY

All globin chains (see legend to Table II) are considered. The correlation is calculated for the globin chain as a whole as well as for the individual sections of the chain.

Section	Sites	Length n	Correlation coefficient $r \pm$ standard deviation	Significance p (null hypothesis)
Total	1–165	165	0.09 ± 0.08	0.32
NA	1–12	12	-0.04 ± 0.35	
A	13–28	16	-0.11 ± 0.28	
B	30–45	16	0.17 ± 0.24	
C	46–52	7	$-0.14 \pm {}^{0.42}_{0.48}$	
CD	53–60	8	$0.57 \pm {}^{0.23}_{0.27}$	0.15
D	61–67	7	0.02 ± 0.46	
E	68–88	21	0.16 ± 0.24	
EF	89–97	9	$-0.19 \pm {}^{0.29}_{0.25}$	
F	98–106	9	$0.18 \pm {}^{0.25}_{0.29}$	
FG	107–111	5	$0.34 \pm {}^{0.39}_{0.57}$	0.54
G	112–130	19	-0.03 ± 0.24	
GH	131–136	6	$0.27 \pm {}^{0.48}_{0.56}$	
H	137–158	22	0.11 ± 0.22	0.62

$r = 0.57 \pm 0.27$ (significance level: 15 per cent) is to be expected as an extreme on a chance basis in a sample of 14 items).

However, a significant correlation between polarity and variability can be brought out if certain functional groups of sites are compared. This correlation is blurred out when the molecule is considered in its totality. The failure to find the correlation along linear sections of the molecule is no doubt due to the fact that the function with respect to which variability and polarity are both correlated, namely, exteriority, is not linearly distributed along the chain.

TABLE VII

MEAN SITE POLARITIES AND VARIABILITIES FOR ALL GLOBIN CHAINS

Two sites are common to the lists of internal and contact sites.

	Number of sites	Mean polarity per site	Mean variability per site
All sites	148	7.50 ± 0.1	4.2 ± 3.0
Internal sites	33	5.54 ± 0.81	3.1 ± 2.6
Contact sites	37	7.46 ± 1.78	3.4 ± 2.3
External noncontact sites	80	8.28 ± 1.88	5.0 ± 3.1

Table VII shows that the mean residue polarity and the mean variability both seem to increase by nearly the same factor as one goes from the set of internal sites to the set of external noncontact sites.

As expected (see the preceding section), contact sites do not participate in this numerical fit; their mean variability is not significantly higher than that of the internal sites, whereas their mean polarity is considerably higher. It must be pointed out that "contact sites" are considered here globally and defined as sites at which, in *some* chains, a contact with another chain is formed (on the basis of the situation in horse hemoglobin). There are of course chains that do not form these contacts. Therefore, the means in Table VII do not express the results of the contact function quantitatively, but only indicate a trend. If for every contact site, one considered only chains that are actually supposed to form a contact at that site, the mean variability for such a group should be lowered, and the contact sites should be further outside the polarity—variability correlation.

The same remark applies to the results obtained when groups of sites, defined by their exteriority, are investigated for individual helical sections (Table VIII).

TABLE VIII

MEAN SITE POLARITIES AND VARIABILITIES AND THEIR RELATIONSHIP FOR DIFFERENT HELICAL SECTIONS

Capital letters refer to helical sections. Data on helices B and C have been combined.

	Number of sites			Mean polarity per site			Mean variability per site		
	B + C	G	H	B + C	G	H	B + C	G	H
Internal sites	6	5	6	6.0	5.0	5.3	1.8	2.8	2.7
Contact sites	10	10	9	7.1	7.8	8.1	3.3	4.5	2.8
External non-contact sites	7	5	10	8.9	7.2	7.6	5.6	3.6	5.2

Mean polarity for this set of data: 7.0 ± 1.2
Mean variability for this set of data: 3.6 ± 1.2
Slope of regression line 0.68 ± 0.26. Intercept at −1.18.
Correlation coefficient $r = 0.71 \begin{smallmatrix} +0.15 \\ -0.26 \end{smallmatrix}$, $p = 0.007$.

Both polarity and variability at contact sites are always higher than at internal sites, but the relationship between contact sites and external noncontact sites is variable. The whole set of data shows a good correlation with a significance level at 0.007.

The mean variability at contact sites being much like that at internal sites, while the polarity at contact sites is much more like that found at external noncontact sites, it is seen again (Table VII) that contact sites must contribute to the blurring of a correlation between polarity and variability when all sites are considered together. It is to be noted that the contact sites do not have to be notably less polar, on the average, than the other external sites. Typical apolar residues, beginning with valine, have "polar requirement" values of 5.6 and less. The mean value for internal sites is 5.5. That for contact sites is 7.5, not far (taking into account the standard deviations) from the value of 8.3 that obtains for external noncontact sites.

4.2. *Correlation between polarity and exteriority and between variability and exteriority.* Because of the different possible orientations of residue side chains, the degree of exteriority of a residue can be established only by X-ray diffraction studies. But the position of residue sites along helical sections might allow one to define a degree of exteriority that is meaningful, if not in each individual case, at least on the average.

With this view in mind, exteriority $E(i)$ for site number i was characterized as follows. For each helix a site was chosen that seems to be exactly on the summit, or exactly in the valley, of a Perutz-wave ([11] Figure 1). Given this site and the number k, put $E(k) = \pm 1$, and define the $E(i)$ of the others by

$$E(i) = \cos \frac{2\pi}{3.6} (i - k) = \cos 100^\circ (i - k). \tag{1}$$

With P = polarity and V = evolutionary variability, the correlations $P(i)/E(i)$ and $V(i)/E(i)$ are then investigated with the help of a computer.

As Table IX shows, the correlation between exteriority thus defined and polarity is highly significant, as also are the differences between the correlation coefficients that are characteristic for certain groups of globins. Thus, the polarity-exteriority correlation coefficient seems to be a tool of some value for classifying globin chains. One is led to assume that the degree to which a globin protomer conforms to the correlation between polarity and exteriority is linked to function and, thus, determined by natural selection. Is a lowering of the correlation coefficient with respect to the ideal polarity-exteriority relationship due to the appearance of polar residues in the interior or due to apolar residues on the outside?

A comparison of the data of Tables II and IX shows that the higher the mean polarity of a type of globin chain, the better the correlation coefficient for polarity *versus* exteriority. Since myoglobins have a considerably larger number of charged groups on their surface than individual hemoglobin chains and since in spite of the occurrence of a number of polar groups in the interior of globin chains the polarity inside any one particular chain cannot be enhanced considerably, it is clear that the result of the above comparison of data implies that the lowering of the average polarity in the individual hemoglobin chains is brought about by placing apolar groups at the outside.

The relatively low polarity values for the two monohemic hemoglobin chains that have thus far been "sequenced," namely, the lamprey and *Chironomus* chains, are compatible with a hypothesis according to which low mean polarity is the more primitive condition in globins. If so, the passage from monohemic to tetrahemic hemoglobin did not require the lowering of surface polarity and

TABLE IX

Correlation Between Polarity and Exteriority for all Sites Belonging to Helical Sections of the Molecule

The correlation coefficient is

$$r = (\sum PiEi - \sum Pi \sum Ei/N)/[(\sum Pi^2 - (\sum Pi)^2/N)(\sum Ei^2 - (\sum Ei)^2/N)]^{1/2}$$

with P_i polarity of site i; E_i, defined for helices only, exteriority of site i; and N the number of sites. Thirteen deduced ancestral chains as given by Dayhoff [4] corresponding to six "α nodes" and seven "β nodes" were added to the contemporary chains. The standard deviations given after the individual values correspond to the approximately normal distribution of $\tanh^{-1} \rho$ (ρ: population correlation coefficient). The standard deviations of the group means (underlined) refer only to the averaging procedure of the individual values.

α chains					
Human	0.28 ± 0.09	Rabbit	0.31 ± 0.09	Sheep A	0.28 ± 0.09
Rhesus	0.25 ± 0.09	Horse	$0.23^{+0.09}_{-0.10}$	Llama	0.27 ± 0.09
Dog	0.29 ± 0.09	Pig	0.31 ± 0.09	Kangaroo	0.30 ± 0.09
Mouse	0.32 ± 0.09	Bovine	0.27 ± 0.09	Carp	0.28 ± 0.09
β chains					
Human	0.34 ± 0.09	Pig	0.26 ± 0.09	Goat A	0.33 ± 0.09
Rhesus	0.35 ± 0.09	Llama	0.35 ± 0.09	Barbary Sheep	0.31 ± 0.09
Lemur	0.34 ± 0.09	Bovine	0.37 ± 0.09	Kangaroo	0.33 ± 0.09
Dog	0.35 ± 0.09	Sheep B	0.35 ± 0.09	Frog	$0.38^{+0.08}_{-0.09}$
Rabbit	0.33 ± 0.09	Sheep C	0.34 ± 0.09		
Horse	0.29 ± 0.09	Sheep A	0.34 ± 0.09		
γ, δ, Fetal chains					
Human γ	0.35 ± 0.09	Sheep	0.36 ± 0.09		
Human δ	0.31 ± 0.09	Bovine	0.37 ± 0.09		
Monohemic globins					
Lamprey	0.31 ± 0.09	Mb Horse	$0.47^{+0.07}_{-0.08}$	Mb Kangaroo	0.46 ± 0.08
Mb sperm whale	$0.43^{+0.08}_{-0.09}$	Mb bovine	$0.50^{+0.07}_{-0.08}$	*Chironomus*	$0.20^{+0.09}_{-0.10}$

"Nodes"					
α	1	0.27 ± 0.09	β	7	0.35 ± 0.09
	2	0.27 ± 0.09		8	0.34 ± 0.09
	3	0.26 ± 0.09		9	0.33 ± 0.09
	4	0.25 ± 0.09		10	$0.48^{+0.07}_{-0.08}$
	5	0.27 ± 0.09		11	$0.38^{+0.08}_{-0.09}$
	6	0.30 ± 0.09		12	$0.38^{+0.08}_{-0.09}$
				13	0.31 ± 0.09

Mean values

α	β	γ, δ, fetal	Non-α	Monohemic
0.28 ± 0.03	0.34 ± 0.03	0.35 ± 0.05	0.34 ± 0.12	0.46 ± 0.05

one would be led to surmise that quaternary structures arise as a proper "lock and key" interaction between residues, irrespective not only of the polarity of the residues involved, but of the overall polarity at the surface.

If, on the other hand, high polarity was the primitive condition at the surface of the common ancestor of myoglobin and hemoglobin chains, then the tendency to form tetrahemic hemoglobins has implied a general lowering of average polarity at the surface, the lowering being slightly more pronounced for α chains than for non-α chains.

The only reason why a monohemic globin like that of lamprey is called hemoglobin and not myoglobin is that it is found in the blood. This is neither a structural nor an evolutionary reason. It is noteworthy that from the polarity-exteriority correlation the lamprey chain behaves like a real hemoglobin, with a value intermediary between that of the average α chain and of the average non-α chain, and not like a myoglobin.

The frog β chain has the highest value of the β chains, though it is almost identical with that of the bovine β chain. The carp α chain has a value identical with that of man and sheep. Evidently, at least one of the primitive vertebrates carp and frog has not preserved the correlation coefficient characteristic of the common ancestor of the α and β chains. The difference in the correlation coefficient that developed along the two lines presumably corresponds to some functional requirement linked to the interaction between the two types of chains. This functional requirement has been satisfied in lower as well as in higher vertebrates. It may have been satisfied relatively fast at the beginning of the evolution of tetrahemic hemoglobins.

The values for the "nodes" (see legend to Table IX), that is, for the deduced ancestral chains, are close to the average value for α chains for α chain "nodes," and close to the average value for β chains for β chain "nodes." This consistency encourages one to think that the deduced ancestral chains are not too far from reality. A severe drop in the correlation coefficient for the ancestral chains would have made one suspicious in this respect.

Finally, the correlation between exteriority and evolutionary variability (Table X) is also significantly positive ($p = 0.02$); that is, the greater the

TABLE X

OVERALL CORRELATION BETWEEN POLARITY, EXTERIORITY, AND EVOLUTIONARY VARIABILITY OF GLOBIN SITES

The computations are done on all sites and all chains, including the "nodes." In the case of "exteriority," the sites considered are, however, restricted to the helical sections.

	r	p
Polarity/exteriority	0.33 ± 0.08	0.005
Polarity/variability	0.09 ± 0.08	0.32
Exteriority/variability	$0.21 \begin{smallmatrix} + 0.10 \\ - 0.09 \end{smallmatrix}$	0.02

exteriority, the higher the variability. We thus note that polarity and exteriority, as well as exteriority and variability are significantly correlated, whereas, as mentioned before and as again shown in Table X, polarity and variability are not significantly correlated when all sites (exterior and interior) are taken into account simultaneously. Such a discrepancy is possible if the figures that fit badly in the two other correlations are combined in the third one.

Thus, all the expected correlations have been found, including that between polarity and variability, provided the latter is established for sets of sites grouped according to exteriority. At the same time, the present results verify that by looking at the whole population of molecular sites it was indeed correct to conclude (Zuckerkandl, Derancourt, and Vogel [18]) that the evolutionary variability of polar and nonpolar amino acids does not differ significantly.

In summary, we find (1) that a certain polarity modulation along the molecule is characteristic of a type of globin protomer and probably linked to function; (2) that a great constancy in mean polarity per residue as observed for total chains is compatible with great variability of polarity at individual sites, even at internal sites, and along various sections of the molecule; (3) that many internal sites do not exclude polar residues at one time or another; (4) that polarity of residues on the other hand is not significantly lowered at interchain contact sites; (5) that all types of amino acids—apolar, polar uncharged, and charged—are found at such sites; (6) that cysteine is a candidate for specializing in the formation of interchain contact; (7) that the formation of quaternary structure leads to a decreased rate of evolutionary change at contact sites, but in general neither to absolute invariance nor to decreased polarity at these sites; (8) that the correlation between polarity and evolutionary variability is upset by this behavior of contact sites; and (9) that, however, polarity and exteriority as well as exteriority and variability are significantly correlated.

The fact that interchain contacts between globin protomers do not lead to absolute invariance at contact sites suggests that absolute invariance, when found in other proteins and brought in relation to molecular contacts, is due not to an intrinsic necessity of "freezing" residues engaged in intermolecular contact, but to an absolute invariance of the partner chain based in turn on functional reasons *other* than those of establishing intermolecular contact.

REFERENCES

[1] W. Bolton and M. F. Perutz, "Three dimensional Fourier synthesis of horse deoxyhaemoglobin at 2.8 Å resolution," *Nature*, Vol. 228 (1970), pp. 551–552.

[2] G. Buse, S. Braig, and G. Braunitzer, "The constitution of the hemoglobin (erythrocruorin) of an insect (*Chironomus thummi thummi, diptera*)," *Hoppe-Seyler's Z. Physiol. Chem.*, Vol. 350 (1969), pp. 1686–1691.

[3] J. Chauvet and R. Acher, "Sequence of frog hemoglobin β," *FEBS-Letters*, Vol. 10, (1970), pp. 136–140.

[4] M. O. Dayhoff, *Atlas of protein sequence and structure*, Vol. 4, Silver Spring, Md., National Biomedical Research Foundation, 1969.

[5] J. Derancourt, A. S. Lebor, and E. Zuckerkandl, "Séquence des acides aminés, séquence des nucléotides et évolution," *Bull. Soc. Chim. Biol.*, Vol. 49 (1967), pp. 577–607.

[6] C. J. Epstein, "Relation of protein evolution to tertiary structure," *Nature*, Vol. 203 (1964), pp. 1350–1352.

[7] ———, "Non-randomness of amino acid changes in the evolution of homologous proteins," *Nature*, Vol. 215 (1967), pp. 355–359.

[8] R. T. Jones, Personal communication, 1970.

[9] W. Kauzmann, in *The Mechanism of Enzyme Action* (edited by W. McElroy and B. Glass), Baltimore, Johns Hopkins Press, 1954.

[10] J. C. Kendrew, "Side-chain interactions in myoglobin," *Brookhaven Symp. Biol.*, Vol. 15 (1962), pp. 216–228.

[11] M. F. Perutz, J. C. Kendrew and H. C. Watson, "Structure and function of haemoglobin, II," *J. Mol. Biol.*, Vol. 13 (1965), pp. 669–678.

[12] M. F. Perutz and H. Lehmann, "Molecular pathology of human haemoglobin," *Nature*, Vol. 219 (1968), pp. 902–909.

[13] M. F. Perutz, M. Muirhead, J. M. Cox and L. C. G. Goaman, "Three dimensional Fourier synthesis of horse oxyhaemoglobin at 2.8 Å resolution: the atomic model," *Nature*, Vol. 219 (1968), pp. 131–139.

[14] H. A. Sober (editor) and R. A. Harte (compiler), *Handbook of Biochemistry*, Cleveland, Chemical Rubber Co., 1968.

[15] H. Vogel, "Two dimensional analysis of polarity changes in globin and cytochrome *c*," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 5, pp. 177–191.

[16] C. R. Woese, in *The Genetic Code*, New York, Harper and Row, 1967, p. 172.

[17] E. Zuckerkandl, "Hemoglobins, Haeckel's "biogenetic law," and molecular aspects of development," *Structural Chemistry and Molecular Biology* (edited by A. Rich and N. Davidson), San Francisco, W. H. Freeman and Co., 1968.

[18] E. Zuckerkandl, J. Derancourt and H. Vogel, "Mutational trends and random processes in the evolution of informational macromolecules," *J. Mol. Biol.*, Vol. 59 (1971), pp. 473–490.

[19] E. Zuckerkandl and L. Pauling, "Evolutionary divergence and convergence in proteins," *Evolving Genes and Proteins* (edited by V. Bryson and H. J. Vogel), New York, Academic Press, 1965.