Evolutionary Divergence and Convergence in Proteins

EMILE ZUCKERKANDL

Laboratoire de Physico-Chimie Colloidale du C. N. R. S., Montpellier, France

AND

LINUS PAULING

California Institute of Technology, Pasadena, California

I. THE MOLECULAR APPROACH TO THE ANALYSIS OF THE EVOLUTIONARY PROCESS

Exponents of chemical paleogenetics have been faced at the present meeting by two disapproving scientific communities, the organismal evolutionists and taxonomists on the one hand, and some pure (very unorganismal) biochemists on the other hand. Some of the biochemists point out, or imply, that the interest in the biochemical foundation of evolutionary relationships between organisms is a second-rate interest. According to them (and to us), what most counts in the life sciences today is the uncovering of the molecular mechanisms that underly the observations of classical biology.

The concept of mechanism should, however, not be applied exclusively to short-timed processes. The type of molecules that have been called informational macromolecules (68) or semantides (75) (DNA, RNA, proteins) has a unique role in determining the properties of living matter in each of three perspectives that differ by the magnitude of time required for the processes involved. These processes are the short-timed biochemical reaction, the medium-timed ontogenetic event, and the long-timed evolutionary event. Although the slower processes must be broken down into linked faster processes, if one loses sight of the slower processes one also loses the links between the component faster processes.

Why are semantides to play a privileged role in the understanding

of living matter? For Simpson (60), "the most truly causative element in the adaptive system" is to be located at the organismal level. There may be no such thing as the most truly causative element. Yet there is no greater concentration of causal factors than in informational macromolecules. There is indeed no greater concentration of information. Here the concept of information replaces to advantage the concept of cause, since it is a better tool in the analysis of reality.

An organism is, by virtue of its genome, what one might call an informostat, by analogy with a chemostat or a thermostat. It keeps nearly constant the information that it contains and that it passes on. Its main memory banks are those polynucleotides that are capable of self-duplication. In order to accept the special importance of the analysis of informational macromolecules, it is sufficient to subscribe to the following propositions: (a) The level of biological integration that contains the greatest concentration of "causal factors" will further our understanding of life more than any other. (b) A concentration of information is a concentration of "causal factors." (c) The largest concentration of information present in an organism, and perhaps also the largest amount of information, and the only organically transmissible information, is in its semantides.

This last proposition has been discussed elsewhere (75) and will not be re-examined here. It suggests the decisive importance of studying all processes of life at the level of their macromolecular foundation, including long-timed processes (evolution). It also suggests that semantides are potentially the most informative taxonomic characters and not, as has been contended at this meeting, just one type of characters among other, equivalent types.

No organism would be apprehended as such if it were not for the nontransient character (through survival or reproduction) of a certain, complex constellation of traits whereby it is defined. The factors of constancy that define the organism as a nontransient organization are all in the informational macromolecules, and in that sense the essence of the organism is there and not at any level of the environment of these molecules.

Since taxonomy tends, ideally, not toward just any type of convenient classification of living forms (in spite of a statement to the contrary made at this meeting), but toward a phyletic classification, and since the comparison of the structure of homologous informational macromolecules allows the establishment of phylogenetic relationships, studies of chemical paleogenetics have a bearing on taxonomy. The taxonomic competence of this nascent discipline is, in fact, only a by-product in relation to the main concern, which is not a classification of organisms, but is the modes of macromolecular transformations retained by evolution,

types of changes in their information content, consequences of these changes for molecular function and for the organism as a whole, and the history of evolution as seen from these points of view. The evidence with a specifically taxonomic bearing, in the phyletic sense, that is to be derived from studies in chemical paleogenetics has been considered by some biochemists as uninteresting. It has the measure of interest one is willing to grant to progress in knowledge of the biochemical nature of extinct organisms, progress in knowledge of the course evolution has actually taken at a particularly important level of biological integration, and progress in the as yet nonexistent knowledge about phases of evolution that have left no trace among fossils. Certainly we cannot subscribe to the statement made at this meeting by a renowned biochemist that comparative structural studies of polypeptides can teach us nothing about evolution that we don't already know.

It is more important to understand the general than the particular, but the first is achieved only through the second. By what precise channels of molecular transformations evolution probably proceeded on earth is indeed particular (and at the same time only one of the possible achievements of chemical paleogenetics). The full value of this type of knowledge will be apparent only if and when terrestrial evolution can be compared with evolution on other planets and organic evolutions become a class of phenomena to which general laws will apply. Only then, presumably, will it become known whether certain general trends of anagenesis (progressive evolution), such as the appearance of central nervous systems and of their higher stages of development, occasionally culminating in hominoids, are in their essential features the necessary result of a relatively small number of definable factors (as we believe on the basis of the numerous cases of independent parallel evolution on earth), or the result of such a large number of accidents that, for all practical purposes, the trend is not expected to be reproducible (Simpson, 61; Dobzhansky, 15).

The relative importance of the contributions to evolution of changes in functional properties of polypeptides through their structural modification on the one hand, and of changes in the timing and the rate of synthesis of these polypeptides on the other hand, constitutes a further important problem that in itself would justify the study of evolution at the level of informational macromolecules. The evaluation of the amount of differences between two organisms as derived from sequences in structural genes or in their polypeptide translation is likely to lead to quantities different from those obtained on the basis of observations made at any other, higher level of biological integration. On the one hand some differences in the structural genes will not be reflected else-

where in the organism, and on the other hand some differences noted by the organismal biologist may not be reflected in structural genes. The first propositon should hold on account of the degeneracy of the genetic code (Zuckerkandl and Pauling, 75; see also Itano, 34, and, especially, T. Sonneborn, this volume); the second proposition—concerning the existence of phenotypic differences not reflected in the base sequence of structural genes—should hold if all regulator genes are not at the same time structural genes active in putting out a metabolically functional polypeptide product. Many phenotypic differences may be the result of changes in the patterns of timing and rate of activity of structural genes rather than of changes in functional properties of the polypeptides as a result of changes in amino acid sequence (69). Consequently, two organisms may be phenotypically more different than they are on the basis of the amino acid sequences of their polypeptide chains, and they may be phenotypically less different than they are on the basis of the base sequence of their structural genes. Quantitatively, prominent enzymes and structural proteins that are likely to be investigated from the point of view of their amino acid sequences may not, as a rule, be controlled by structural genes that are at the same time regulator genes in relation to other structural genes. If this supposition is correct, the compounded results of sequence studies in polypeptides will give a measure of phyletic distance that is unique in kind. The type of measure of differences at the polypeptide level will take into account basic constituents of the organism only, whereas other types of measure, at higher levels of biological integration, will include modifications in the interaction patterns between these constituents. By subtraction, an evaluation of the contribution of the latter processes to the over-all difference between two organisms may be possible. If no significant difference is found, the implication is that most structural genes are also controller genes, and that amino acid substitution in one type of polypeptide chain is in general reflected by changes in rate or period of synthesis of other polypeptide chains.

Another justification for putting time and effort into comparative structural studies of homologous polypeptide chains was quoted at this meeting as though in opposition to the perspective we are developing here, namely the interest to correlate differences in amino acid sequence with differences in physicochemical, functional properties of a given type of protein. Of the primary importance of such studies we have, of course, been fully aware (49, 75); efforts in that direction have led to results of which a few have already been published (72), and others are included in this article. By furnishing probable structures of ancestral proteins, chemical paleogenetics will in the future lead to deductions concerning molecular functions as they were presumably carried out in the

distant evolutionary past. Progress in laboratory methods of polypeptide synthesis, already partly accomplished, will permit the study of the functional properties of ancestral polypeptide chains directly.

There is yet an ultimate reason, of a more philosophical nature, for interest in the paleogenetic approach. Whereas the time dependence of evolutionary transformations at the molecular level (see Section III) can be established only by reference to extraneous sources, the topology of branching of molecular phylogenetic trees should in principle be definable in terms of molecular information alone. It will be determined to what extent the phylogenetic tree, as derived from molecular data in complete independence from the results of organismal biology, coincides with the phylogenetic tree constructed on the basis of organismal biology. If the two phylogenetic trees are mostly in agreement with respect to the topology of branching, the best available single proof of the reality of macro-evolution would be furnished. Indeed, only the theory of evolution, combined with the realization that events at any supramolecular level are consistent with molecular events, could reasonably account for such a congruence between lines of evidence obtained independently, namely amino acid sequences of homologous polypeptide chains on the one hand, and the findings of organismal taxonomy and paleontology on the other hand. Besides offering an intellectual satisfaction to some, the advertising of such evidence would of course amount to beating a dead horse. Some beating of dead horses may be ethical, when here and there they display unexpected twitches that look like life.

Simpson (60) believes that, for studying affinities between organisms, characters that are far removed from the genes are better than characters of the genes themselves or of the closely related polypeptide chains. By a character that is "far from the genes" one is apparently to understand a character that is determined by the interaction of numerous genes, with the implicit specification that the consequences of this interaction are apprehended at a high level of biological integration. It seems to us, however, that this specification is not really relevant to establishing a difference between the effect of the action of numerous genes and the effect of the action of a single gene. The cellular, tissular, systemic, organismal consequences of a single gene mutation are also "far removed" from the gene. The "length of the functional chain from the genes to the character selected for or against" (Simpson, 60) is always considerable, for instance from the structural change in the single gene responsible for phenylketonuria to the feeble-mindedness of the affected individual. Furthermore, although characters observed by the organismal biologist are determined by many genes at a time, any *change* in such characters is presumably determined by discrete single gene mutations, and it is on

such *changes* that natural selection is expected to act. The effect of natural selection should therefore be correlated, not exceptionally, as Simpson proposes, but generally, with single mutations in single genes. Evidently, the further we are from the gene, the better we usually understand why selection occurred, since the function of the gene is definable only in terms of the interaction of direct and indirect products of the gene with direct and indirect products of other genes and with factors from the environment of the organism. At the level of the individual gene itself we find no basis for selection, since we find no basis for defining function. Selection occurs wherever function occurs. Although it is in the nature of function to be understood in terms of events at a lower level of integration in relation to events at a higher level, this fact seems to have little bearing on the question whether, in living matter, there is any level of organization as informative for tracing phyletic affinities as that of the informational macromolecules.

To the extent to which differences in gene regulation are not reflected in amino acid sequences of polypeptides, the analysis of levels "far from the gene" provides information that is not to be obtained at a level as close to the gene as that of the polypeptides. If so, the reason, in this case, is that the polypeptide level is not quite close enough to the gene. It is almost unavoidable to postulate that differences in gene regulation must correspond to differences in nucleotide sequences in self-duplicating polynucleotides, or in the order along the genetic chromosomal or extrachromosal units of the functional subunits—again a matter of sequence.

Although we disagree with some of the views put forward by Simpson in his recent article (60), the ability of this author to state important issues in clear terms offers rewarding opportunities for thought and discussion. A further objection of organismal "evolutionists" to chemical paleogenetics will be dealt with in Section III.

II. Patterns of Amino Acid Substitution

Various aspects of this topic have been examined by others, notably by Šorm and his group (64), by Lanni (39), and by Pattee (48). In the present treatment we shall examine some effects of divergent evolution on the amino acid sequence of a number of homologous globin chains. We make the likely assumption that, except in regions of the molecule where the comparison between the chains reveals a deletion or an addition, the differences noted are attributable to successive single amino acid substitutions. The globin chains whose amino acid sequence has been available to us in whole or in part are the sperm whale myoglobin chain (58),[1] the human myoglobin chain (23), the human hemoglobin

[1] The sequence of sperm whale myoglobin used is one that was kindly com-

α, β, γ, and δ chains (58, 22), the horse α and β chains (12, 62), the cattle α and β chains (53, 3), the cattle fetal chain (3), the pig α and β chains (10), various Primate β chains (24), the gorilla α and β chains (76, 77), the carp "α" chain (11), and a lamprey hemoglobin chain (52).

The different residues are so numbered that homologous residues in different chains carry the same numbers. The numbering system used is that introduced by Kendrew *et al.* (cf. Cullis *et al.*, 13). Capital letters refer to helical regions, and pairs of capital letters refer to regions between the corresponding helices. The general use of this numbering system, although perhaps not strictly rational, since the system applies specifically to the sperm whale myoglobin chain, allows one to visualize easily the region of the molecule in which a given residue is located. Abbreviations are used as by Cullis *et al.* (13), except that asn stands for asparagine, glm for glutamine, and ilu for isoleucine.

1. Invariant Molecular Sites

Until very recently it seemed that 16 homologous sites of hemoglobins and myoglobins were strictly invariant. This figure, which represents about 11% of the total number of residues per polypeptide chain of this type, had already been decreased in the light of new information on amino acid sequences in comparison to the figure of 13% given earlier (69). The sequence analysis of the cattle fetal chain by Drs. Donald Babin and W. A. Schroeder (3) has removed one further residue from the list of the invariant ones (phe in the cattle chain instead of try at position A12), and the partial sequence analysis of a lamprey hemoglobin chain by Drs. V. Rudloff and G. Braunitzer (52) has removed four further residues from this list (in the lamprey chain, lys instead of arg at B12; met instead of leu at CD7; arg instead of lys at E5; ilu instead of val at E11). At this writing we are left with 11 invariant sites in hemoglobin and myoglobin polypeptide chains, representing about 8% of the total number of sites.

One may wonder how far this progressive shrinkage of the number of apparent invariant sites will eventually go. It is unlikely that this number will decrease to zero, because it is unlikely that the function of hemoglobin and myoglobin can be maintained if the iron atom of the heme is linked to a residue other than histidyl. But the "final" number of invariant residues may indeed turn out to be 1 or 2. Not even in the immediate environment of the heme group is invariance assured any

---

municated to us by Dr. L. Stryer in March 1964 as including that latest corrections made by Drs. Edmundson and Kendrew. This sequence is not always identical as to the nature of the residues and the exact location of helical sections with the sequence published by Dr. Perutz in his book in 1962 (50).

longer. At the writing of Kendrew's 1963 article (38), of 11 residues of globin that interact with the heme group, 8 appeared invariant on the basis of the information available to him. At present the count is reduced to 4, one of which is the so-called proximal histidine, the one to which the iron atom is linked. Another of these residues that still appears invariant evolutionarily, the so-called distal histidine, situated at the opposite site of the heme group, has been shown to be substituted in a functional, although somewhat unstable, abnormal human hemoglobin, Hb Zurich (44). It is not very probable, but possible, that arginine will be found at this site in some normal globin, i.e., one that has become widespread in a species by natural selection. The two remaining invariant residues in contact with the heme group are C4 (threonine) and CD1 (phenylalanine). The threonine is shown by Kendrew to interact with one of the vinyl groups of heme, and the phenylalanine with one of the pyrrole rings. The residue that interacts with the other vinyl side chain (H14) and the two residues (FG5 and E11) that interact each with one other pyrrole ring (the fourth pyrrole ring is not reported to be in contact with any amino acid residue) all have been shown to be variant residues. Consequently, the two remaining invariant residues may also eventually be found to be substituted in some hemoglobin or myoglobin chain that is distantly related to the one whose sequence is presently known.

The other remaining invariant sites in the polypeptide chains are A14 (lys), B6 (gly), B10 (leu), C2 (pro), GH5 (phe), H9 (lys), and H22 (tyr). At A14, valine is substituted by aspartic acid in the α chain of the abnormal human hemoglobin I (45). Even though the substitution is probably an unfavorable one in man (a HbI homozygote has not yet been found, and therefore the "seriousness" of the condition in man cannot be evaluated), HbI is recognizable as a functional hemoglobin. It is therefore not unlikely that a substitution at site A14 will be found to occur in some as yet unknown normal hemoglobin or myoglobin. The invariant glycyl residue at position B6 has been shown by Kendrew and his collaborators to interact with another glycyl at position E8. The two glycyls meet at a point of crossing between two helices, the helices B and E. It has been surmised by some that the invariance of these glycyl residues is critical for the structure of hemoglobins and myoglobins, since any other residue at these sites would not permit an equally short contact between the two helical segments. It has, however, turned out that the glycine at E8 is replaced by alanine in the cattle α chain and in a lamprey hemoglobin chain. Consequently it is also possible that the glycine at position B6 will be found to be replaced by alanine in some forms.

It is worth noting that the majority of the remaining "invariant"

residues in globin chains—namely those that appear evolutionarily invariant—are neutral. This is especially the case if we do not consider the two histidines at E7 and F8, which have special functional roles, with respect not to the general structure of the globin, but to the property of reversible oxygenation of the heme iron. In that case we count 8 neutral residues out of 9, and of these 8, 6 are apolar. As will be mentioned again later, at sites where charged residues occur replacements are found more frequently than at sites where apparently only apolar residues occur. Over and above the sites that appear to be invariant in both hemoglobins and myoglobins, we count at present 15 sites that are invariant in hemoglobin chains only. This brings the total number of invariant sites, for the hemoglobins alone, to 26, a figure that will no doubt decrease further. Of the 15 sites mentioned, 12 are apolar (E4, FG5, G2, G5, G7, G8, G12, G17, H1, H5, H14, H18), 1 is polar and noncharged (F5), and 2 are charged (EF3, G1). Of the nonpolar residues only one, the residue at G17, is found to be replaced by a different type of residue, namely a charged residue, in a myoglobin chain.

Even when a residue is seen to remain strictly invariant in forms very far removed from each other on the evolutionary scale, it cannot be deduced from such an observation that none of the other nineteen amino acids would be compatible with the maintenance of the functional properties of the molecule. It is possible, in particular, that some favorable substituent cannot be reached through a single mutational step, nor through two steps one of which represents an isosemantic substitution (74). Any of the possible intermediate substituents may be unfavorable. Furthermore, some residues may remain invariant, not because they are intrinsically necessary, but because any change would have to be coordinated immediately with one or more changes in other parts of the molecule. For example, residue H22 (tyrosine) is known to interact (13) with residue FG5 (valine or isoleucine). This interaction may be critical for keeping helix H in position and for assuring the stability of the whole molecule. Conceivably, two or more simultaneous changes would be required here to reach a different, functionally satisfactory state. The time allotted to evolution on earth may not be long enough for such an event to occur. This consideration is of some consequence in relation to the problem of molecular convergence (cf. Section V). It is remarkable that most globin residues can presumably be changed—sometimes, as we shall see, only within narrow limits—without any simultaneous adjustments in other parts of the molecule. The question is often asked why only twenty different amino acids are provided for proteins by the genetic code, since sixty-four different triplet code words are available, and many more than twenty different amino acids are produced metabol-

ically. Without going into a discussion of this question here, we may note that, in the case of most globin sites, more different amino acids appear to be coded for than are necessary for achieving a certain molecular function.

The residues common to all known globin chains are unlikely to be engaged in specific interactions between polypeptide chains, since the vertebrate myoglobins exist as single chains. These residues should rather be instrumental in intrachain stabilization, i.e., in the stabilization of the tertiary rather than of the quaternary structure, or in allowing the heme to carry out its basic function. We saw that only one or two residues may play this latter role, and that the number of the other invariant residues may eventually decrease to zero or to a very small number. This means that no, or almost no, amino acid residue is specifically needed for stabilizing a given tertiary structure. Conversely, if absolute invariance is observed, on the basis of what is becoming apparent in the case of the globin chains it would seem that this invariance, in most instances, is probably not necessary for the stabilization of the tertiary structure of the polypeptide. This statement may be valid also for proteins with little or no helical content, since in hemoglobins and myoglobins the amino acid sequence in interhelical regions is also highly variable. The progressive shrinkage of the number of invariant sites in globin chains brings into relief the fact, first strongly suggested by a comparison of myoglobin and hemoglobin polypeptide chains (50), that what function is associated with and what nature selects for is a tertiary more than a primary structure. A restriction to this statement will be discussed in the next subsection. The fact is that most interactions between helices and most sequential devices for stabilizing nonhelical regions are not strictly linked to the presence of one unique amino acid residue at one unique site.

These observations on globin structure are spectacularly at variance with observations on the primary structure of cytochromes *c*. In these cytochromes about 50% of the residues seem to have remained evolutionarily immutable since the time of the common ancestor of yeast and man (41). The question arises whether this difference with the globins is due to the presence in cytochrome *c* of a large proportion of residues of intrinsic absolute evolutionary stability, while the remaining residues change evolutionarily at a rate comparable with the rate of evolutionary change in hemoglobins; or whether in cytochromes *c* evolutionary changes are so slow throughout the molecule that the time elapsed between the epoch of the common ancestor of man and yeast and the present, although enormous (presumably 1 to 2 billion years), is insufficient for a high percentage of the residues to be substituted.

If we compare human polypeptide chains with the corresponding chains in horse, cattle, pig, and rabbit we find, on the average, 22 differences between the adult major component hemoglobin chains (cf. Section III.4) and 10 differences between the cytochromes *c*. These figures represent, respectively, 15% and 10% of the total numbers of residues in the chains. Thus in mammalian evolutionary history the difference in rate of change of the two types of proteins seems to have been small. This difference can be totally eliminated if, on the basis of available information, certain plausible assumptions are made about the number of strictly invariant residues in tetrahemic hemoglobins and in cytochromes *c*, and if the number of changed residues is related to the probable number of changeable residues rather than to the total number of residues.

If no such assumptions are made, the application of the quantitative relation to be described in Section III of this article to the case of cytochrome *c* leads to an unacceptable value for yeast. The common ancestor of yeast and man appears at an epoch that is implausibly recent. In other words, during a plausible span of time too few residues have been changed in cytochrome *c*. This incongruence encourages one, in answer to the question formulated above, to assume the absolute evolutionary immutability of an important proportion of residues in cytochrome *c*.

It has been proposed that the observed invariance could be due to the presence of mutational "cold spots," to a local lack of mutability of the genic DNA. According to this tentative interpretation the mutation rate would be the limiting factor in the rate of evolutionarily effective mutations. This is very unlikely. If a reasonable value of total mutation rate and of interbreeding population size is assumed, one finds that the probable number of mutations per amino acid site and per line of descent is far larger (by 100 to 1000 times) than the average number of evolutionarily effective mutations that show up in the hemoglobin α chain–non-α chain comparison. This should apply to proteins generally, at least in organisms with a generation time up to an order of magnitude of one year. Furthermore, from what is now known about globin chains, it is certain that 94% of the coding triplets are mutable. This percentage is slightly higher than would correspond to the 8% evolutionarily unchangeable polypeptide sites referred to earlier. Indeed, although the replacement of the "proximal" and "distal" histidines (E7 and F8) has never been evolutionarily effective, so far as we know, these residues have been found to be replaced through point mutation in abnormal human hemoglobins: Hb M_{Boston}, Hb M_{Emory} (20), Hb Zurich (44), Hb $M_{Kankakee}$ (35). Another evolutionarily "unchangeable" residue has

been replaced in the abnormal human hemoglobin I, as already mentioned. This reduces the possibly immutable sites to 8, i.e., to 6% of the molecule. There is reason to think that the apparent absolute stability of 6% of the coding triplets is caused not by the absence of mutation but by the action of natural selection. It is not plausible that a small number of base triplets, known to make sense in terms of an amino acid, and scattered over the molecule, should radically differ in mutational behavior from the overwhelming majority of triplets. Moreover, a very slow rate of evolutionarily effective substitutions at certain molecular sites does not imply a slow mutation rate. In hemoglobins and myoglobins, at molecular sites that are substituted very rarely during evolutionary history, substituents are nearly exclusively confined to functionally closely similar residues. It is true that according to Eck's proposal for a complete genetic code nearly all transitions between functionally closely related amino acids can be brought about by one single mutational step. This fact could be quoted as evidence in favor of the plausible hypothesis that the genetic code, as it stands today, was evolved at early times of the history of life through the action of natural selection. The evidence cannot at present be so construed, because Eck's code was based in part on considerations of functional similarity between amino acids, so that there would be an element of circularity in this reasoning. One must await the establishment of a definitive complete code before settling this matter. However, even if the present impression should then be confirmed, namely that indeed the code provides one-step transitions between functionally related amino acids by a statistically significant bias, this would not mean that other one-step transitions do not also occur. It would be unreasonable to assume that, at highly but not absolutely invariant molecular sites, only those transitions occur that are evolutionarily effective and actually observed, namely between functionally closely related amino acids. We conclude that there is no reason to suppose that, in hemoglobins and myoglobins the residues that are only rarely substituted successfully during evolution undergo, on the average, fewer mutational changes, including the evolutionarily noneffective ones, than the residues that are changed during evolution with a relatively high frequency. Mutational hot spots and cold spots may exist, but the relative frequency of evolutionarily effective substitutions at different molecular sites gives no clue as to the distribution of such hot spots and cold spots. From the evidence that has become available, it seems highly unlikely that any cold spot in globin genes is cold enough to prevent mutational changes altogether during the time allotted to evolution.

There is no reason to assume that mutation rates in cytochromes should differ radically from those that obtain in globins, in that one half of the residues represent cold spots so icy that no mutations occur at all, whereas the other half of the residues mutate at a normal rate. The improbability of such a view is obvious.

How is the high resistance to evolutionarily effective mutation of part of the cytochrome *c* molecule to be explained? A decrease in the rate of evolutionarily effective substitutions may be brought about by the aging of the molecule: the phase of the most active molecular transformation is expected to be the incipient phase, just after a polypeptide chain has become established in a new function (49). Cytochrome *c* may have gone through this phase even before the epoch of the common ancestor of yeast and man. However, the type of stability obtained through the aging of a molecule should not be absolute. Conservative (41) (= isogenetic, ref. 72) evolutionarily effective mutations are expected to occur occasionally within the limits of relative evolutionary stability (cf. Section III). Such conservative substitutions are found in cytochrome *c* at most of the changeable molecular sites (see Section II.3). Their occurrence does not indicate that, at these changeable sites, cytochrome *c* has not yet reached its optimal state in relation to any given set of circumstances. If the substitution is strictly indifferent, a shift between closely related residues may nevertheless spread in the population through genetic drift. The assumed indifference of the substitution implies, in that case, a satisfactory adaptation of the original residue to molecular function, because an identical degree of inadaptation of two residues is unlikely to occur. Alternatively, one may assume that the substitution of one residue by another even closely related one never has a selective value of exactly zero. In that case the occurrence of either one of two closely related residues in different lines of descent suggests that each of the residues is better adapted than the other to certain conditions of the external and internal environment of different organisms. In each form the best-adapted residue presumably occurs, and the observed variation is again compatible with satisfactory molecular adaptation in any given set of circumstances. The apparent absence of substitutions from nearly one-half of the molecular sites of cytochrome *c* cannot be taken as evidence for stability obtained through progressively improved adaptation in the course of evolution, if the presence of conservative substitutions at other sites is, at any given time, compatible with a fully adapted state. The measure of stability in amino acid sequence that one may expect to be reached by virtue of a relatively rapid phase of initial molecular adaptation should not be absolute.

Absolute stability can hardly be attributed to evolutionary old age of the molecule. It should more probably be due to very stringent functional requirements.

What type of stringent functional requirement in relation to numerous molecular sites can one conceive for cytochrome *c* that does not also obtain for hemoglobins and myoglobins? This question is being discussed by Margoliash and Smith in their contribution to this volume. As these authors point out, by analogy with globins, one would expect that few, if any, residues are absolutely required for the maintenance of a given tertiary structure. To be sure, a larger number of residues than in globins, where it is perhaps 1 or 2, may be required for preserving the basic functional properties of the cytochromes. On the other hand, from what is known today about the structure–function relationship in proteins it is not to be expected that the absolute stability of a large number of residues is required for maintaining the basic functional properties of a prosthetic group or of an active site. The most plausible explanation of absolute stability over and above that of the residues associated with the prosthetic group or the active site might be restrictions imposed on structural variation by specific interactions between molecules or molecular subunits. This effect, which may be called the Ingram effect (32; cf. Section III), is expected to lead to complete structural invariance of a section of a polypeptide chain only if the interacting molecule is itself invariant, or if the number of variable interacting molecules is sufficiently large. The stable sites of cytochrome that are not concerned in preserving the basic function of the prosthetic group may be required to interact with several other macromolecules or (Margoliash and Smith, this volume) with some evolutionarily invariant organic molecule such as a coenzyme. Alternatively, the stable regions not directly associated with the prosthetic group may be required to interact with a single type of macromolecule, which, however, is in turn required to interact with a number of others. This latter situation might obtain for the cytochromes because of their association with a cellular organelle, the mitochondrion. As was also mentioned independently by a participant at the present meeting, requirements for the stability of the primary structure of a polypeptide may not be the same for proteins that are normally in free solution and proteins associated with cellular organelles. Surface structures that reversibly interact with soluble proteins in such organelles may display a high structural constancy. If so, the interacting regions on the soluble proteins should be equally invariant. If this consideration applies to cytochrome *c*, the stable sequences should be found at the outside of the molecule, or in parts that move to the outside if the molecule becomes reversibly bound and perhaps reversibly

denatured within the mitochondrion. It will be of interest to see whether some of the stable sequences in cytochrome *c* are found again in other mitochondrial enzymes.

One further suggestion may be made tentatively. The reduced form of cytochrome *c* is completely resistant to proteolysis, whereas the oxidized form is susceptible to attack by proteolytic enzymes (40a, 63). This observation suggests that oxidation and reduction are accompanied by an intramolecular rearrangement. Since cytochrome *c* contains one polypeptide chain per molecule, the rearrangement could not be compared to the one that is observed upon oxygenation and deoxygenation in hemoglobins (43), which seems to be primarily one in quaternary structure. For cytochrome *c* a probable change in tertiary structure may be inferred. The cytochrome polypeptide chain should therefore have well-defined kinetic properties for the conformational rearrangement to occur properly, and some severe restrictions may thus be imposed on sequence. This hypothetical feature of cytochrome *c* might in particular explain the evolutionary stability of some of the glycine residues, if reversible bending movements have to occur in some nonhelical regions of the molecule. At certain sites any residue other than glycine might be in the way of such movements of the main chain.

Whatever the reason for the stability of one-half of the cytochrome *c* molecule, it helps one realize how exceedingly old proteins, and therefore structural genes, may be. It is possible that this will be found to apply to a large proportion of the proteins, including the globins, and that living matter has changed less in the last 1 to 2 billion years than morphological comparisons suggest. No doubt biochemical evolution has added new proteins to old ones and has, in the process, dispensed with some of the latter. But a significant proportion of the oldest proteins and genes may have been preserved, if we consider as oldest, in relation to life as we know it today, an epoch not much further removed in time than that of the common ancestor of yeast and man.

2. *Variable Molecular Sites*

What is called the evolution of a given type of protein molecule amounts in general to the introduction of the greatest structural changes compatible with the smallest functional changes. It seems that the greatest functionally tolerable changes in tertiary structure are small, whereas the greatest tolerable changes in primary structure will affect, according to the type of protein, part of the molecule or its near totality.

The different types of mutational changes that may occur in proteins have often been enumerated. Deletions or additions of one to several amino acid residues are expected to be eliminated by natural selection in

a high proportion of cases. Those that are preserved should be mostly found at either end of a chain, at the end of helices, in short helices, or in nonhelical regions, notably in loops that may be shortened or lengthened without affecting the steric relationships in the rest of the molecule. A deletion or addition in the middle of a long helix would result in so many simultaneous alterations in side-chain interactions that it is highly unlikely that the tertiary structure and the function of the molecule could survive such an event. The deletions or additions found in hemoglobin and myoglobin chains are compatible with these generalities.

The outlook for survival of the molecule should be likewise poor if two or more adjacent amino acid residues are replaced by an equal number of residues. No evidence of such an event in globins has as yet been forthcoming, and it would not be easy to ascertain.

A major conformational rearrangement of the molecule, while leading to the disappearance of one molecular function, might of course be accompanied by the appearance of a new molecular function. In that case, however, the molecule is no longer of the type under consideration, and it may not be recognized or isolated readily.

Perhaps the most frequently occurring type of mutational event and, at any rate, the most frequently observed and preserved type is the replacement of one single amino acid residue by one other amino acid residue. As we saw, such substitutions can occur mutationally anywhere along the globin chain. In general one may expect natural selection to act against the following effects of such substitutions:

1. At the inside of the molecule, against a notable increase in bulk (there would mostly be no space for such an increase) and against an increase in polarity (this would destabilize the structure; cf. Schachman, 54).

2. At the outside of the molecule, against a notable change in polarity (because the solubility characteristics of the protein would be altered).

3. At bends, against substitutions that destabilize them and tend to turn the regional sequence into a straight helix.

4. At helices, against substitutions that destabilize *them* (cf. Schellman and Schellman, 55).

5. At sites where conformational specificity requirements are particularly high, namely at "active sites," or at sites of binding of prosthetic groups, or at sites of binding of coenzymes and of molecules that act through allosteric effects (42), or at sites of binding of associated polypeptide chains, or at sites remote from these, but whose modification directly affects one or the other of them.

If any of these restrictions are ignored by an amino acid substitution, the functional properties of the molecule should be partly or totally impaired, while its tertiary structure, in many cases, may not be, or may be only slightly affected. Single amino acid substitutions, precisely because the tertiary structure of the molecule has a fair chance not to be greatly changed, are expected to lead to the appearance of a new molecular function in a smaller proportion of cases than the more radical mutational changes considered above. Typically, they lead to impairment of function, not to the evolvement of a new function.

The restrictions to functional amino acid substitutions enumerated on a priori grounds seem stringent. Yet the degrees of liberty, for amino acid substitutions, on the basis of the limited number of globin chains that are at present available to us for comparison, are already as high as 6 or 7 at some sites (7 or 8 different residues found at these sites), and as high as 2.0 per site, on the average, for the whole molecule (3.0 different residues found on the average per site). This situation is expressed in the well-known fact that the differences in amino acid sequence between mammalian hemoglobin chains and myoglobin chains are very numerous indeed. For instance, between the human hemoglobin α chain and the sperm whale myoglobin chain—two chains that are very similar in tertiary structure and function—there seem to be 107 differences in amino acid sequence, over the 141 molecular sites that are common to the two chains. Thus the two chains differ over 76% of their sites.

How is one to reconcile the considerable plasticity of the primary structure within the limits of a given type of tertiary structure and of function with what appear to be good reasons for expecting stringent restrictions with regard to evolutionarily effective amino acid substitutions? These two features are compatible principally on the following two grounds.

Firstly, from the evidence derived from the comparison of different globin chains it appears that the presence of a certain residue at a certain site is only exceptionally required for "absolute" functional reasons, as apparently in the case of the heme-linked histidine in globins. Most of the time residues are needed at some sites in relation to other residues at other sites. What seems to count, most often, from the functional standpoint, is a system of interactions between two or more residues and not the chemical nature of any single residue per se. As more extensive data will be forthcoming about the interactions between the different side chains in hemoglobins and myoglobins, evolutionary changes will best be analyzed in terms of changes in these interactions, rather than simply in terms of changes in amino acid sequence. Since the changes in primary structure are introduced most of the time residue

by residue, groups of interacting residues may be transformed progressively. When more than two residues are involved in such a group, the alteration of one does not necessarily compromise the local stability of the molecule, and this first alteration may make possible subsequent ones, at other sites, that heretofore were not permissible.

Secondly, numerous amino acid substitutions are possible because many of the amino acid residues found in proteins are so similar in structure and chemical properties to at least one other type of residue that transitions from one amino acid to the most closely related ones will usually represent a very small modification indeed. These are the "conservative" transitions. We shall see that their list is extensive and, in effect, includes couples that have not been considered, on a priori grounds, as candidates for conservatism.

However, even in globins, whose amino acid sequence proves to be so variable, the plasticity of the primary structure in relation to the tertiary structure is by no means unlimited. Although we saw that the number of strictly invariant sites is dwindling, a number of sites are nevertheless of high evolutionary stability in that the only substituents tolerated at these sites are residues most closely related to those usually found. Even this very restricted type of transition can be extremely rare. Thus, at residue B12, arginine, found in all known globin chains, is replaced by lysine in lamprey hemoglobin; and, similarly, at E11, valine by isoleucine, and at CD7, leucine by methionine. (The transition leucine–methionine is a very common one, as the data in Tables I and II show, and their functional kinship is thereby strongly suggested, in spite of the fact that methionine is somewhat polar, whereas leucine is not.) There are 41 sites in normal globins at which so far only one substituent amino acid has been found. This figure represents 28% of the 148 globin sites under consideration here, each of which is shared by at least some type of hemoglobin chain with the mammalian myoglobin chain. The most frequent conversions at these sites occur between leucine and phenylalanine (5 times), between lysine and arginine (3 times), and between valine and leucine (3 times). At 20, i.e., one half of these sites, the transitions are of the type defined here as "very conservative" (see Section II.3). At the 21 remaining relatively stable sites, the expectancy of finding some other substituent in the future is highest. However, the sites at which only one single highly conservative conversion has so far been found are probably mostly sites at which substitutions are greatly restricted. If we add to these sites the number of strictly invariant globin sites, we find that at about 31 globin sites, i.e., at 21% of the total number of sites, possible substitutions, if any occur, may be restricted, in each case, to one most conservative one. Thus, although the number of ab-

solutely invariant sites is very small, one-fifth of the globin molecule will tolerate substitutions, if any, of maximum conservatism only.

In the present treatment we do not examine primarily substitutions at particular molecular sites. Rather, by tabulating the evidence for all individual globin sites we attempt to diagnose some general features of amino acid conversions and of properties of residues.

In contrast to the stable molecular sites, the most variable sites are F3, with 8 different residues recorded to date, and B4, E13, and G14, with 7 different residues at each site. At bends and in interhelical regions the variability is about the same as elsewhere in the molecule. For the 28 interhelical sites found in myoglobin as mentioned (the list of these sites may not be exactly coincident with the one applicable to hemoglobins), the mean number of different residues found at a site is 3.0, identical with the mean that holds for the whole molecule.

The number of different residues that may replace a given residue is very large, if we consider conversions in general rather than conversions at any particular site. At lysine, histidine, alanine, leucine, serine, and threonine sites all amino acids have already been found in globins, or the few that are missing occur in such small numbers in the molecule that their nonparticipation in certain conversions is at the moment nonsignificant. If we consider only conversions whose absence is of possible general significance because of the relatively frequent occurrence of the residues in the molecule, the only conversions not so far found in globins are the following: Neither aspartic nor glutamic acid has been replaced by phenylalanine. Proline has not been replaced by the large apolar or related amino acids valine, leucine, phenylalanine, or methionine. Phenylalanine has been replaced neither, as mentioned, by proline, aspartic, or glutamic acid, nor by arginine or glycine. Glycine has not been replaced by arginine or phenylalanine. None of these conversions is allowed to proceed in one mutational step according to Eck's proposal for a genetic code (16), except the conversion glycine–arginine. None of them is allowed according to the set of coding triplets of Wahba *et al.* (cf. Jukes, 37). Perhaps the reason why none of these conversions occurs evolutionarily resides in the requirement of intermediate mutational steps, although two-step conversions might be circumvented if the two residues are arrived at through single steps from a common ancestral residue. More probably the causal relationship is in the opposite direction: Conversions that most of the time are nonfunctional may have become two- or three-step conversions by virtue of a natural selection that modified the genetic code before it became fixed.

We shall examine conversions between amino acids in relation to

TABLE I

RESIDUE SITES AND THEIR CONVERSION CHARACTERISTICS IN GLOBINS

Residue site	Number of residue sites	Number of different substituents found (maximum = 19)	Substituents occurring at above 20% of sites (in parentheses, percentage substituent sites in relation to total number of sites)	Average number of substituents per site	Number and, in parentheses, percentage of residue sites with only zero or one substituent. Nature of unique substituents.
asp	32	14	glu(47), ala(44), gly(34), asn(22), lys(28), ser(25), his(22),	2.9	7 (22) *glu*, lys, his, gly, ser, pro
glu	34	15	asp(44), lys(41), ala(50), gly(38), pro(24)	3.1	1 (3) ala
lys	39	19	glu(31), asp(28), gly(23), ala(23), arg(21), ser(21), leu(21)	2.7	7 (18) *asp*, *arg*, leu
arg	12	12	lys(58), his(50), ser(25)	2.3	3 (25) *lys*
his	27	18	lys(30), asp(26), leu(26), ser(22), thr(22)	2.8	4 (15) leu, asp
pro	14	10	glu(57), asp(43), ala(50), gly(43), lys(21)	2.6	4 (29) asp, ala, (ilu?)
gly	32 + 1[a]	14	ala(51), ser(36), asp(33), glu(30), lys(27)	2.7	5 (15) *ala*, asp, ser
ala	51	19	gly(33), ser(31), glu(31), asp(28)	2.7	8 (16) *gly*, *leu*, val, glu, ser, pro

[a] Because of an uncertainty concerning homology relations, N-terminal residues are not included in the evaluation of conversion characteristics.

TABLE I (*Continued*)

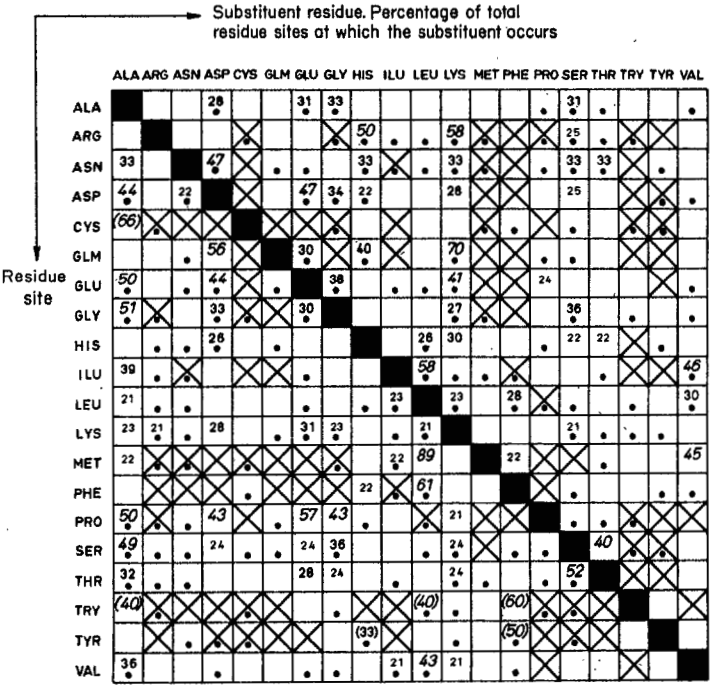
Residue site	Number of residue sites	Number of different substituents found (maximum = 19)	Substituents occurring at above 20% of sites (in parentheses, percentage substituent sites in relation to total number of sites)	Average number of substituents per site	Number and, in parentheses, percentage of residue sites with only zero or one substituent. Nature of unique substituents.
val	28 + 1[a]	17	leu(43), ala(36), ilu(21), lys(21)	2.4	9 (31) *leu, phe, thr*, ilu, gly, ala
leu	40	18	val(30), phe(28), ilu(23), lys(23), ala(21)	2.1	15 (40) *phe, val, met*, ilu, lys
ilu	13	13	leu(58), val(46), ala(39)	2.5	4 (33) *leu*, val, (pro?)
phe	18	11	leu(61), his(22)	1.7	11 (61) *leu*, val, his, try
met	9 + 1[a]	10	leu(89), val(45), ilu(22), phe(22), ala(22)	2.6	2 (22) *leu*
ser	33	16	ala(49), thr(40), gly(36), lys(24), asp(24), glu(24)	2.8	5 (15) thr, asn, gly, ala, leu
thr	25	17	ser(52), ala(32), glu(28), gly(24), lys(24)	2.9	3 (12) ser, val
asn	15	14	asp(47), lys(33), his(33), ala(33), ser(33), thr(33)	3.3	1? (7?) (tyr?)
glm	10	12	lys(70), asp(56), his(40), glu(30)	3.1	1 (10) asp
tyr	6	10	phe(50), his(33)	2.1	2 (30) asn
try	5	9	phe(60), leu(40), ala(40)	2.8	1 (20) phe
cys	3	8			

general functional types of residues before considering individual kinds of residues.

The types of residues that appear at the largest number of globin sites when the evidence from the different known globin chains is pooled (cf. Table I) are the charged residues except arginine, the smallest, functionally most neutral residues glycyl and alanyl, the bulky apolar

TABLE II

SUBSTITUTION FREQUENCIES IN GLOBINS[a]



[a] Along the column dimension, the alphabetical list of amino acids represents residue sites. Figures at intersections between rows and columns represent percentages of these residue sites at which a given substituent has been found. (Example: glutamyl occurs at 47% of aspartyl sites.)

Large figures in italics: "very conservative" substitutions (see Table V), at or above 40%. Medium-sized figures: "fairly conservative" substitutions (see Table VI), above 25%. Small figures: substitutions that occur at 21 to 25% of any given residue site. Frequencies of rarer substitutions are not listed. Blank spaces therefore represent substitutions that so far have been found to occur rarely or very rarely in globins. Spaces occupied by a cross indicate substitutions so far not found at all. In parentheses, figures of particularly low level of significance on account of the small number of residue sites involved. Solid dots: one-step conversions allowed according to Eck's (16) genetic code.

residue leucyl, and the small hydrogen-bond-forming residue seryl. If we group related residues, we find that several major categories of residues occur at one-third to one-half of the globin sites (Table III). On the other hand, at one time or another, nearly all molecular sites have been uncharged. The number of sites at which only one or the other of these types of residues occur is restricted: alanyl plus glycyl alone occur at only 3 sites; acidic residues alone at 2 sites; basic residues alone at 7 sites; charged residues alone at 15 sites; seryl, threonyl plus the amides,

TABLE III

NUMBER OF RESIDUE SITES OF DIFFERENT TYPES IN GLOBIN CHAINS PRESENTLY AVAILABLE FOR COMPARISON

Type of sites	Number of sites	Per cent of total sites
Total	148	
Charged	90–91	61
Acidic	51–52	35
Basic	57	39
Hydrogen-bond-forming (ser + thr + asn + glm)	61	41
ser + thr alone	45	30
ala + gly	67	46
Apolar with bulky side chains (val, leu, ilu, phe) + met	66	45
Uncharged	133	90

alone, at 3 sites. On the other hand the apolar bulky residues, valyl, leucyl, isoleucyl, phenylalanyl plus methionyl, occur to the exclusion of any other residue at 26 sites. These figures give an indication of the surprising frequency of conversions between residues with different chemical properties. They also show 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 may be the most specifically determined business of molecular sites in globular proteins.

The correctness of this last statement in relation to globins is confirmed by the figures that represent the average number of different substituents found for any particular amino acid residue (Table I). For most amino acids one finds an average number of substituents per site that is close to 3. Exceptions are valine, leucine, and foremost phenylalanine, the latter with 1.7 substituents per site on the average. The number of sites with no substituents or with only one substituent is highest for leucine and phenylalanine, the two apolar amino acids with the bulkiest side chains.

It may turn out that the larger the side chains in aliphatic residues, the smaller, in general, is the number of substituents that occur. At the moment this correlation is upset by the case of isoleucine (Table I), although the relative smallness of the sample of isoleucine sites that are so far known in globins prevents one from attributing significance to this finding.

Table IV and Fig. 1 show that the proportion of different types of substituents generally decreases as the hydrophobic side chains increase in bulk. If we draw the best straight line through the points on Fig. 1, the

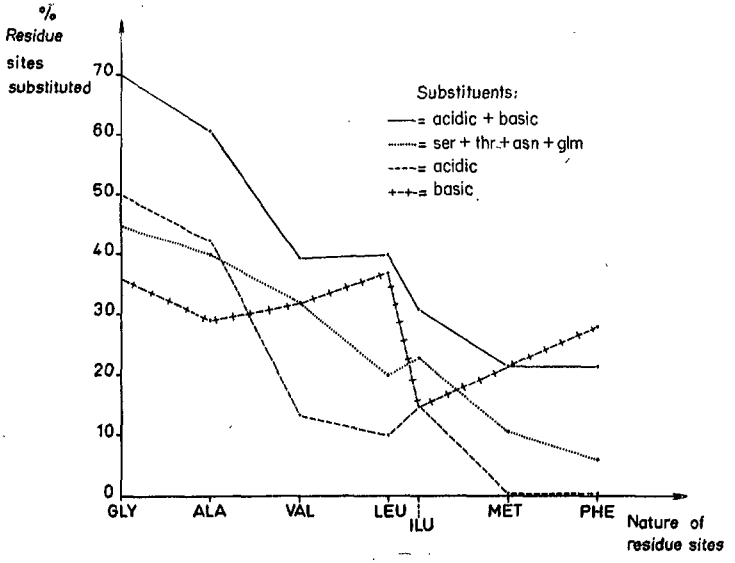


FIG. 1. Frequency of various types of amino acid substituents at residue sites for hydrophobic side chains of increasing bulk. Values from Table IV.

line relative to all charged substituents considered together has a slope not unlike that of the line relative to the hydrogen-bond-forming residues serine, threonine, asparagine, and glutamine considered together, although the latter are less frequent substituents. When we consider acidic and basic substituents separately, an interesting difference between these two categories appears, however. The slope for the acidic substituents is the greatest, whereas that for the basic substituents is the smallest and is, on the average, not far from zero. This observation means that as the side chains of hydrophobic residues increase in length or bulk the sites for such residues become rapidly intolerant to acidic residues, whereas, as far as can be judged from available data, their tolerance to basic residues does not vary significantly. This observation

TABLE IV

CONVERSION FREQUENCIES AT GLOBIN SITES IN RELATION TO SUBSTITUENTS GROUPED ACCORDING TO FUNCTIONAL CHARACTERISTICS[a]

Vector of structural properties	Residue sites	Acidic + basic substituents	Ser + thr	Ser + thr + asn + glm	Acidic substituents	Basic substituents	Ratio of basic over acidic substituents
↓	gly	67(22/33)	39(13/33)	45(15/33)	49(16/33)	36(12/33)	0.7
	ala	61(31/51)	37(19/51)	41(21/51)	43(22/51)	29(15/51)	0.7
	val	39(11/28)	29(8/28)	32(9/28)	14(4/28)	32(9/28)	2.3
	leu	40(16 or 17/40)	12(5/40)	20(8/40)	10(4/40)	38(15/40)	3.7
	ilu	31(4/13)	23(3/13)	23(3/13)	15(2/13)	15(2/13)	1.0
	met	22(2/9)	11(1/9)	11(1/9)	0(0/9)	22(2/9)	2/0
	phe	22(5/18)	6(1/18)	6(1/18)	0(0/18)	28(5/18)	5/0
↓	asp		31(10/32)	53(17/32)			
	glu		32(11/34)	47(16/34)			
↓	lys		26(10/39)	44(17/39)			
	arg		42(5/12)	62(7/12)			
	his		33(9/27)	56(15/27)			
↓	ser	58(19/33)			36(12/33)	42(14/33)	1.2
	thr	52(13/25)			48(12/25)	52(13/25)	1.1
	asn	73(11/15)			53(8/15)	80(12/15)	1.5
	glm	100(10/10)			70(7/10)	90(9/10)	1.3
	pro	79(11/14)	14(2/14)	14(2/14)	79(11/14)	21(3/14)	0.3

[a] The first figure gives the percentage of the residue sites inved. In parentheses, the actual figures of the ratio are given.

is, of course, reflected in the ratio of basic to acidic substituents. At the sites for the smallest residues of the series, glycine and alanine, the acidic substituents predominate. The trend is reversed for the residues with larger apolar side chains. The fact that the larger the apolar side chains the larger, in general, is the proportion of basic to acidic substituents may be related to the difference in size between the acidic and the basic amino acid residues. The basic amino acid residues that occur in proteins are larger than the acidic residues. The large basic residues may have a function beside those linked to their charge and to their capacity to form hydrogen bonds, namely one connected with their capacity to form hydrophobic bonds. A globular protein may be able to tolerate a charged amino acid residue at its inside if, along with the charge, this residue offers a sufficient opportunity for hydrophobic interaction.

The behavior of proline is exceptional. Although this residue is bulky and apolar, the number of sites with acidic substituents is definitely higher than the number of sites with basic substituents. The case of substituents at proline sites is examined more closely in separate papers (72, 73).

Charged residues of either sign occur at one time or another at 90 or 91 different molecular sites, i.e., at over 60% of the sites. The distribution of these sites is such that they cannot be all at the outside of the molecule. Since, however, in any one "edition" of the globin chain the great 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 (38), 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, Table I), 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. (Glycine and alanine residues are not confined to the interior of the myoglobin molecule; 38.) It is plausible that this should be so, since it is at the inside of protein subunits that the requirement for steric fit between residues should be the most generalized.

The transition between glutamyl and aspartyl has been previously found to be a frequent one in evolution (71). It now appears that the transition between glutamyl and lysyl is equally frequent (Tables I and II). This shows, interestingly, that at many charged globin sites (i.e., at sites where charged residues occur in at least one known globin chain) the sign of the charge is of little consequence. Of equally little consequence is the question whether, at many of the charged sites, there is any charge at all in any particular case. Indeed, alanine and

glycine are other very frequent substituents at charged sites, especially at acidic sites. Perhaps this bias in favor of the acidic sites is due to the fact that the basic side chains are larger than the acidic ones and may therefore be replaced with more difficulty by the very small residues alanine and glycine.

What counts primarily may be the distribution of the charges over the surface of the molecule, not their presence at any particular site. The function of alanyl and glycyl seems to be purely negative. The presence of these residues means "no charge at this site," "no obstacle to bend" (72), or "no obstacle to close contact between different parts of the polypeptide chain." More often than not, glycyl or alanyl seem to be no more than molecular spacers along the polypeptide chains. Evolutionarily, only one glycine in globins seems to be absolutely stable and, hence, absolutely required, at 1 out of 33 glycine sites (B6, see above). The situation, from this point of view as from others, is greatly different in cytochrome *c*; 10 out of the 12 glycines of yeast cytochrome *c* seem to have remained stable since the time of the common ancestor of yeast and man. A possible reason for the stability of glycine residues, beside their importance in providing opportunities for short contacts and for sharp bends, has been suggested earlier (Section II.1).

At present, no sequence of more than 3 consecutive molecular sites is left at which no charged amino acid has been found in globin molecules. Five such sequences of 3 are present. Two uninterrupted series of seven "charged sites" are found (E20-EF6, G17-GH4). Except for site D5 all sites of helix D are sometimes charged. Helical and interhelical regions alike show great concentrations of "charged sites."

Of the 57 (or 58) sites at which no charged amino acid has so far been found, 29 sites, one-half of the number, display only the bulky nonpolar side chains of valine, leucine, isoleucine, or phenylalanine. Only once are 2 such sites found consecutively, at G7 and G8, where only leucine and phenylalanine occur. Five times such sites are found 4 residues apart in helical regions (A8, A12; B10, B14; E11, E15; G12, G16; H7, H11). There are 6 further sites at which only bulky apolar side chains *or* alanine are found. Evolutionarily effective conversions leading to charged amino acids are unlikely at sites at which so far only bulky apolar side chains have been found and somewhat more probable at sites where also alanine has been found. One may estimate that the total proportion of sites with charged amino acids tends toward a maximun of about two-thirds of the globin molecule.

Conversions between uncharged polar hydrogen-bond-forming residues and charged residues are very frequent (Table IV). At 52% of the sites where a charged residue has been found, an uncharged hydrogen-

bond-forming residue has also been found. It can be seen by reference to Table IV that there is no significant variation in the ability of the different charged amino acids to be substituted by a member of the group of hydrogen-bond-forming amino acids. On the other hand, there is a variation in the readiness of different types of hydrogen-bond-forming residues to be substituted by charged residues. This readiness increases with the size of the hydrogen-bond-forming side chains. This observation suggests that, for conversions between polar residues, the bulk of the side chain is more important than the presence or the absence of a charge.

Substitution patterns at some individual residue sites may now be briefly discussed. The usual indifference to the sign of the charge at charged sites is well borne out by the fact that basic residues occur at about one-half the aspartyl and the glutamyl sites (Table I). Conversely, acidic residues occur to the same extent at lysyl sites. Acidic residues occur at only one-third of the histidyl sites, and this smaller proportion probably expresses the fact that the basic character of histidyl represents only one of several equally important functions of this residue. Arginyl, on the other hand, a much rarer residue in globins than either of the two other basic residues, is replaced by acidic residues only at 2 out of 12 arginine sites. Its most frequent substituents are either of the two other basic residues. As to histidyl, it is substituted with an approximately equal frequency by either lysyl, aspartyl, or leucyl. The fact that leucyl occurs at 26% of the histidyl sites suggests that histidyl, besides acting by virtue of a charge, is functioning by its bulk through hydrophobic bonding. When histidyl interacts with another ring, π-bonding may be important. A further character of functional importance of histidyl may reside in the polar, uncharged part of the molecule, since seryl and threonyl also occur with a significant frequency at histidyl sites. By the multiplicity of its substituents that occur each with a frequency between 20 and 30% (Table I), none being recorded with a higher frequency, histidine appears to be the amino acid in which the most diverse chemical properties play an equally important role in relation to function in globins and perhaps proteins in general. If histidine shares each of its major properties with some other residue, it should be possible for a protein to dispense with histidine at most sites, without any radical change in functional properties. In fact histidine sites are no more stable than other residue sites (Tables I, II, and IV). Only at 15% of the histidine sites is the number of substituents limited to 0 or 1 (Table I). This number is smaller, or not larger, than at the sites of most other residues that occur frequently in globins (Table I).

Besides histidine, asparagine displays a series of substituents that occur with equal frequency: Lysine, histidine, serine, threonine, and alanine are all found at one-third of the asparagine sites. Only aspartic acid is a more frequent substituent, namely at one-half of the asparagine sites. Although asparaginyl is no more versatile in its substitution properties than other residues, its functional affinities with a number of other residues are more nearly equal than is usually found.

It is no surprise that leucine should be the most frequent substituent of phenylalanine, because of the apolar and bulky characters of both residues. It is of interest, however, that, next in line among the frequent substituents of phenylalanine, although at a percentage value far below leucine, one finds histidine. One might have thought of tyrosine as the likely next most frequent substituent. And, indeed, phenylalanine is the most frequent substituent at tyrosine sites. But tyrosine appears at only about 10% of the phenylalanine sites, although the conversion apparently can occur in a single mutational step (Table II). The relative frequency of histidine at phenylalanine sites is paralleled by its relative frequency at tyrosine sites. Apparently the ring character of the histidine residue is in many cases more prominent than the charge of the imidazole group. This is the more plausible, as numerous imidazole groups, in hemoglobins and myoglobins, are inaccessible to titration (e.g., 65). Because of their relatively low pK, the imidazoles in globins are at any rate expected to be only partly ionized. Charge may be entirely suppressed in some histidine residues, notably when the ionizable nitrogen is surrounded by apolar residues.

Because of the smallness of the sample of tryptophan sites in the globin chains whose amino acid sequence is available to us (5 sites, plus 1 in myoglobin, whose exact position, from the point of view of homologies, is not certain), it is difficult to deduce from the nature of the most frequent substituents the type of properties of the tryptophan residues that is mostly made use of in the protein. It appears, however, that at some residues, at least, the action of tryptophan is likely to be linked to apolar bond formation and bulkiness, or to ring character (cf. 50, p. 44ff). This seems to be the case at residues A12, CD4, and H7. On the other hand, at residues C3 and E6, charged or hydrogen-bond-forming substituents are recorded in certain chains in lieu of tryptophan.

For serine and threonine the most frequent substituents are the other partner in this couple, plus alanine and glycine, as one might expect. Yet charged substituents, either acidic or basic, also occur with significant frequency (Table I). This fact may be associated with the hydrogen-bond-forming capacity of charged residues, whereas the sub-

stituents alanine and glycine may be selected for because of their small bulk, reminiscent of the bulk of serine and threonine.

To conclude this section, let us briefly consider alanine sites and glycine sites. We count at present 33 glycine sites, although the content of glycine per hemoglobin or myoglobin polypeptide chain, as far as is known at present, varies between 5 (human α chain) and 15 (horse β chain). The figure of 33 glycine sites can be interpreted to mean that nearly one-fourth of the globin sites, and perhaps an even larger proportion, can accommodate at one time or another during evolution a residue whose role can only be not to interact or not to prevent some other interaction. About one-half (18 out of 33) glycine sites are in interhelical regions or at or near the end of a helix, as judged from the distribution of helical regions in myoglobin. Only a few of the glycine sites that are found inside helices can have the function of allowing short contacts between helices, since many occur at positions where no such contacts take place. Among the frequent substituents of glycine, the unexpected ones are aspartic acid, glutamic acid, and lysine (Tables I and II). With the exception of three sites (B5, E6, H13), glycine sites with bulky substituents are in interhelical regions or near the end of helices. It may be that bulky substituents for glycine are predominantly concerned with the stabilization of bends. In particular, lysine seems to occur at glycine sites only in interhelical regions or at the end of helices. In this connection, one may point out that lysine and glycine are among the more frequent substituents of proline.

Of alanine sites we count at present 51. Over one-third of the globin chains is made of alanine sites, and this figure may of course increase further. Sometimes alanine sites are found in a row. Thus, between B2 and B9 nearly every site is an alanine site, and likewise between H1 and H6, and between H15 and H20. The same applies to glycine sites: between B3 and B7 every site is a glycine site. At other times, in helical regions, both alanine and glycine sites are 4 residues removed from one of the preceding sites and point therefore in nearly the same direction. There are also significant stretches with hardly any alanine or glycine sites, notably the interhelical region FG and the first half of the following helix G.

Alanine sites are the most frequent of all residue sites. Next in line are leucine and lysine sites. It is interesting that next to one of the functionally least prominent residues, alanine, the residues occurring at the greatest numbers of sites are one of the largest apolar ones and one of the largest charged ones. This tends to indicate that the principle functions in the globin molecule are "no function," hydrophobic bonding, and charge. Residues whose main function probably resides in hydrogen

bonding come next in importance. In their "no-function" function alanine and glycine seem, in part, simply to act as spacers along the linear dimension of the primary structure. In such positions, wherever alanine or glycine are found there is room for future functional differentiation of the protein molecule.

The present analysis leads to a few generalizations. We note that the substitutions that occur frequently are mostly those that lead to a change in some of the properties of the residue, while other properties of the residue are being preserved. This may be an important rule, to which conform the great majority of amino acid substitutions that are retained by natural selection. It appears that the basis for extensive changes in amino acid sequence without any radical change in tertiary structure and protein function is furnished by the fact that each amino acid residue has several important functional properties, and that the set of amino acids that are coded for is chosen so that changes in one or more of these properties can occur while one or more other properties are maintained constant. By a propitious choice of the properties to be maintained constant, in the case of each particular residue, the amino acid sequence of the polypeptide chain can be transformed, and yet its basic pattern of intramolecular and intermolecular interactions remain the same.

When a given residue is found at contiguous sites in different globin chains, as occurs not only in the case of glycine and alanine but also in the cases of valine (e.g., region G11–G18), leucine (e.g., region G12–G19), serine (e.g., region H12–H21), lysine (e.g., region F2–F6) (there are at present a few interruptions in the regions quoted), histidine (G17–G19), etc., the suggestion is that the residue in question is required to occur not at any one particular site, but at some site or sites in the region. This leads to the concept of regional functional differentiation in polypeptide chains, which has also been established by Margoliash and Smith (this volume) in relation to hydrophobic segments and to basic segments in cytochrome *c*. The evolving functional unit often is not a single amino acid residue, but a small region of the molecule. Nonhelical, bent regions of the polypeptide chains are a case in point. The analysis of proline sites and of their environment (72, 73), indicates that the stabilization of interhelical regions is founded on the collaborative action of several amino acid residues. This is also emphasized by Dr. L. Stryer (personal communication). Quite generally, several contiguous residues may collaborate in carrying out a regional function in the protein molecule. Within the section devoted to this function, the "division of labor" that seems to exist between different amino acid residues may be redistributed a number of times during

evolution, as is the function of different ministers in a typical French government. Within the limits of a given protein function, "evolution" may consist more often in such a reshuffling than in significant functional change. This is not to say that in the case of any particular version of the globin chain the exact position of a certain residue is sometimes indifferent. Whether such an indifference is possible we do not know. The reshuffling in question does not demonstrate it, since its modality and extent may be strictly dependent on changes in other parts of the polypeptide chain.

The distribution of clusters of certain types of residue sites suggests that certain functions are of particular importance for certain regions of the globin molecule. Thus, from the data described above, it is evident that large apolar side chains have a particular role to play in the second part of helix G, and hydrogen-bond-forming residues in the second part of helix H. Helices G and H are the two that have been implicated in the interaction between like chains in tetrahemic hemoglobin molecules (Cullis *et al.*, 13).

3. *Conservatism and Radicalism of Amino Acid Substitutions*

The treatment of this subject is already implicit in the preceding section. The notions of conservatism and radicalism of substitutions are here relative to protein function.

The best criterion for the conservatism of a substitution is the high frequency with which it is found to occur during protein evolution. As emphasized by Professor J. Lederberg (personal communication), the evolutionary change of any given species of protein is marked by the conservation of the basic functional properties of the protein. Therefore the substitutions most frequently adopted by natural selection must be the most conservative ones, whereas rarely occurring substitutions will be relatively radical—only relatively, since it will never be so radical as to interfere with the basic function of the protein. Such an interference can be detected only in abnormal mutants of a given protein, and most of these "radical" mutants may escape observation, because the criteria whereby the protein is normally recognized may not apply any longer in most cases of radical substitutions.

The average degree of conservatism of a given substitution can be considered to be represented numerically by the frequency with which this substitution is found. Although such figures cannot be blindly applied to any particular site—there are indeed sites at which a conversion that is usually very conservative is very radical—these figures will give an estimate of a priori likelihood of conservatism of an observed conversion and will, on the other hand, attract attention to the rare substitutions

that are likely to introduce a significant functional change into the molecule. In the present article no differentiated scale of conservatism and radicalism of conversions is as yet proposed. Meanwhile, and in relation to globin sites, we distinguish between only three sets of conversions, with arbitrarily chosen limits for the frequencies of occurrence, namely "very conservative" conversions (Table V), "fairly conservative" conversions (Table VI), and nonconservative conversions (cf. Table II).

TABLE V

VERY CONSERVATIVE SUBSTITUTIONS IN GLOBINS[a]

ala–ser	asp–asn		glu–lys	gly–ala
ala–thr	asp–glm		glu–pro	
ala–asp	asp–glu			
ala–glu	asp–pro			
ala–pro				
his–arg	leu–ilu	lys–glm	ser–thr	val–leu
	leu–phe	lys–arg		val–ilu
	leu–met			val–met

[a] Conversions occurring at 40% or more of the residue sites of at least one of the members of a couple of residues. The rarely occurring residues tyr, try, cys not included. The least bulky residue is in general listed first.

TABLE VI

FAIRLY CONSERVATIVE SUBSTITUTIONS IN GLOBINS[a]

ala–val	asn–lys	asp–lys	glm–his	glu–glm	gly–ser
ala–asn	asn–his	asp–his			gly–asp
ala–ilu					gly–glu
					gly–lys
					gly–pro
leu–his	lys–his	phe–tyr	ser–asn	thr–asn	tyr–his
		phe–try			

[a] Conversions occurring at more than 25% of the residue sites of at least one of the members of a couple of residues. The least bulky residue is in general listed first.

The "very conservative" conversions are defined as those that occur at 40% or more of the sites of at least one of the members of any couple. Most of the rarely occurring residues are not included. In Tables V and VI the smaller residue in any couple is quoted first. We count 22 different "very conservative" transitions in globin chains. A number of these transitions would not, so far, have been considered a priori as conservative, notably ala–thr, ala–glu, ala–pro, asp–pro, glu–pro, glu–lys, glm–lys. The inadequacy of a priori views on conservatism and nonconservatism is patent. Apparently chemists and protein molecules do not

share the same opinions regarding the definition of the most prominent properties of a residue.

To the 22 types of "very conservative" transitions we may add 21 "fairly conservative" ones, that occur at more than 25% of the sites of at least one member in each couple (Table VI). Here we find surprising transitions such as ala–ilu, gly–glu, gly–lys, gly–pro. Whereas only one histidine transition is "very conservative" (his–arg), four more histidine transitions are "fairly conservative" (his–lys, his–asp, his–leu, his–tyr). It has been contended that any substitution of histidine must be a "radical" substitution. In fact, however, some of the histidine substitutions appear, like the average American radical, astonishingly conservative (cf. Section II.2).

Among the reputedly radical substitutions that turn out to be conservative are those of proline by some other residues. According to presently available data on globins, glutamic acid occurs at 57% of the proline sites, and some charged amino acid (negatively or positively charged) occurs at 79% of the proline sites. If any substitution is conservative, it must be that of a charged residue for proline. It is clear that glutamic acid has properties different from those of proline. One prolyl residue will break an α helix; one glutamyl residue cannot, by itself, achieve this. In fact, however, as already mentioned, neither proline nor glutamic acid works alone in the stabilization of the conformation of a nonhelical, bent region of the polypeptide chain. Either is able, apparently, to contribute effectively to the collaborative work among several residues that results in such stabilization.

The intuition of the chemist, based on what appear to him prominent properties of amino acid residues, is correct with respect to what he thinks should be conservative conversions. It is in relation to nonconservatism that he may go astray. Take the case of substitutions for lysine. The conversion between lysine and arginine was considered as conservative. This is indeed so. The conversion occurs at 21% of the lysine sites and at 58% of the arginine sites (Tables I and II). On the other hand, although Šorm (64), for one, has included the conversion lysine–glutamic acid in his list of "standard interchangeable amino acids," it seemed legitimate, up to the present time, to consider the conversion lysine–glutamic acid as mainly nonconservative. Although both residues are charged, weight is given a priori to the opposite sign of the charges. It turns out that this weight is light, from the point of view of the protein, and that the transition lys–glu is, like the transition lys–arg, a very conservative one. It occurs at 31% of the lysyl sites and at 41% of the glutamyl sites. Finally, few transitions would appear a priori less conservative than the transition lysine–leucine. Yet leucine

occurs at 21% of the lysine sites, and lysine at 23% of the leucine sites. A substitution that is found at one-fifth of the sites of a frequently occurring residue (there are 39 lysine sites and 40 leucine sites) is not an exceptional, but a systematic occurrence. Therefore it cannot be a very radical substitution, within the framework of observed protein evolution, which is, as mentioned, one of preservation of protein function.

Of course, any substitution by a different residue excludes absolute conservatism. And the survival of a given type of protein excludes absolute radicalism. Examples of nonconservative substitutions, which have been found either rarely or not at all in globins, may be read off Table II. The status of shifts toward rarely occurring residues such as tryptophan and cysteine cannot be evaluated a priori. Some types of conversions may occur only rarely, not because they are radical, but because they have only rarely the opportunity to be conservative. This may apply in particular to tryptophan, a residue that seems to be able to take the place of phenylalanine in a conservative fashion, but only rarely so.

The conversion glutamic acid–valine is fairly radical. It occurs at 2 out of 34 glutamyl sites and at 2 out of 28 valyl sites. The example of this conversion allows one to show how the "radicalism" of a substitution can vary according to the environment of the molecular site involved. At site E9 the lamprey chain has glutamyl, and the sperm whale myoglobin chain valyl. At site G3 the human α chain has valyl, and the human β chain glutamyl. The conversion therefore can occur in normal globin chains, although it occurs very rarely. A considerable interference with hemoglobin function appears in the shift glutamyl–valyl at position A3 that characterizes sickle cell hemoglobin (31, 28). A still more radical interference is observed for the reverse transition, valyl–glutamyl, at E11, which was found in an abnormal human β chain and defines Hb $M_{Milwaukee}$. In this modified chain the capacity of reversible combination with oxygen is lost, the iron of the heme being oxidized to the trivalent state (20).

It is striking that most of the other structurally known abnormal human hemoglobins are characterized by conservative substitutions. Of a total of 22 substitutions in abnormal human α and β chains (Table VII), 8 are "very conservative," 11 "fairly conservative," and 3 "radical." The third radical substitution is that found in the α chain of Hb Shimonoseki, where an arginine allegedly replaces a glutamine (21). Several causes may be responsible for the predominance of conservative substitutions in abnormal human hemoglobins: Firstly, nonconservative substitutions may lead to hemoglobin chains that cease to be recognizable as such. Secondly, most abnormal hemoglobins are detected on account of a change in electrostatic charge, and we saw that at charged sites changes

to uncharged residues or oppositely charged residues are frequent and therefore "conservative." Finally, the genetic code may well favor, for one-step transitions, the conservative over the nonconservative transitions.

There is at present a tendency among some workers interested in the evolution of proteins to propose that a "very conservative" substitution such as that of a glycine for an alanine may often be strictly neutral in relation to natural selection. It is to be noted that, judged from the evolutionary frequency of the conversion, the glycine–alanine conversion is not expected to be more closely neutral in its functional effects than, say, the shift between lysine and glutamic acid, or between glutamic acid and histidine. Glycine occurs in place of alanine, lysine instead

TABLE VII

DEGREE OF CONSERVATISM OF THE SUBSTITUTIONS IN ABNORMAL HUMAN HEMOGLOBIN CHAINS[a]

Conversion	Number of occurrences	Degree of conservatism[b]
ala↔glu	2	v.c.
asn→lys	1	f.c.
asp←lys	1	f.c.
glm→arg	1	rad.
glu↔glm	3	f.c.
glu→lys	5	v.c.
glu↔val	2	rad.
gly→asp	2	f.c.
gly←glu	1	f.c.
his→arg	1	v.c.
his→tyr	3	f.c.

[a] The table includes Hbs C, S, $G_{San\ Jose}$, E, M_{Emory}, $M_{Milwaukee}$, D_{Punjab}, O_{Arabia}, $C_{Georgetown}$, Zurich, $G_{Baltimore}$, Coushatta, Seattle, I, $G_{Honolulu}$, Norfolk, M_{Boston}, $G_{Philadelphia}$, Shimonoseki, $O_{Indonesia}$, Mexico, $M_{Kankakee}$ (refs. 4–9, 20, 21, 25, 28–31, 35, 44, 45, 51, 56, 66).

[b] v.c = very conservative; f.c. = fairly conservative; rad. = radical (cf. Tables II, V, VI).

of histidine, and glutamic acid instead of lysine at one-third of the corresponding sites. No one has as yet claimed that a substitution such as that of lysine by glutamic acid or of histidine by lysine is not likely to be acted upon by natural selection. On the basis of the frequency data, the glycine–alanine conversion should not be judged differently.

It is of interest to see how the notion of conservative and nonconservative substitutions, as derived for the globins on the basis of frequency data, fares when applied to the cytochromes. If we exclude yeast and tuna fish cytochromes *c* from the comparison and consider only the cytochromes *c* from the more closely related forms—man, monkey, dog,

horse, pig, cow, rabbit, and chicken (41)—we find, by considering all possible conversions at sites with multiple substitutions, that there are 31 possible conversions. Of these, 8 are nonconservative by exclusion from the lists of very conservative and fairly conservative substitutions as established for the globins (Tables V and VI). Let us briefly examine the "nonconservative" substitutions in this group. At site 12, the conversion glm–met has not been found in globins, but since there are only 10 met sites and 9 glm sites so far known, this conversion may yet be discovered in globins and not be as radical as it appears at present from its apparent zero frequency in globins. At site 25, the conversion pro–lys occurs at 21% of the proline sites. It is at present below the limit arbitrarily chosen as that of "fairly conservative" substitutions, yet it is not a very radical substitution, and we may discount it as such. At site 58 the conversion thr–ilu is fairly radical in terms of globin. It appears in globin only at $2/13 = 15\%$ of the ilu sites. At site 82 the conversion phe–ilu is not on the lists of conservative ones. This circumstance is likely to be an accident attributable to the small number of known ilu sites in globins, on account of the fact that leu, a residue closely related to ilu, is a very conservative substituent of phe. It is thus likely that, in the future, the conversion phe–ilu will have to be added to the list at least of the fairly conservative substitutions, and we may not count this shift here as clearly radical. At site 88, the conversion thr–lys has been found at 24% of the threonine sites and is therefore, at this point, just below the arbitrary limit of conservative substitutions. It may be discounted as radical. At site 89, the shift ser–asp, one of the possible substitutions, does not have to be postulated, since the observed change may have occurred over another channel. Besides, this shift occurs in globins at 24% of the serine sites. Again, we may not count it among the established radical substitutions in the cytochromes *c* under consideration. At site 92 the shift glu–val occurs in globins at only 7% of the val sites and at 6% of the glu sites. We have referred to it earlier. It is a radical substitution, but its occurrence does not have to be postulated here, since it is possible to account for the observed situation by proposing that alanine found at this site in several of the cytochromes *c* has been an intermediary substituent between valine and glutamic acid. We are thus left with only two substitutions, at sites 12 and 58, that are at present legitimately to be considered as radical, in terms of substitutions in globin chains. This figure represents 6% of the possible conversions that can be envisaged in this group of cytochromes. In both cases the Primate chains are implicated in carrying the exceptional evolutionarily effective mutation. Since conservatism is what one should expect, we

may conclude that the notion of conservatism as defined in relation to the globins seems to be applicable also to cytochrome *c*, and may be applicable to globular proteins in general.

Let us emphasize the following general conclusion: The properties of charge, hydrogen-bond-forming capacity, apolarity, and bulk, and perhaps others, are distributed over various amino acid residues in various combinations. Therefore conversions within a large number of couples of amino acid residues will be conservative with respect to one or more of these traits, and nonconservative with respect to others. The fact that the set of amino acid residues that are coded for by the genetic code forms a network of overlapping properties is probably the basis for the extensive change in amino acid sequence that may occur without a change in the fundamental traits of tertiary structure and protein function. A certain new function—say, associated with charge—may be introduced while the former function—say, apolar interaction—is maintained. This simultaneity of conservatism and nonconservatism may well also be one of the basic conditions of protein evolution and organic evolution in general. There might indeed not be sufficient opportunities for the invention of new functions of polypeptides if the chemical relationship between the different amino acid residues were not such that a number of sequence patterns are compatible with one given polypeptide function. The possibility of extensive variations of the primary structure within the limits of a given function probably provides the richness in combinatory resources that is necessary for making mutations with radical structural effects sometimes successful in relation to a novel function.

4. Conservative Amino Acid Substitutions and the Genetic Code

The possibility was referred to that, when living matter first evolved toward its present form, the genetic code itself went through a phase of evolution during which those transitions between amino acid residues in polypeptides that are most frequently retained by natural selection were made to correspond to relations between codons such that the transitions could be accomplished in single mutational steps.

It is probable, at any rate, that the most frequently observed substitutions do occur in one mutational step. The only data on amino acid substitutions in metazoa that have so far been used for establishing relationships between codons are those bearing on mutants abnormal in relation to the "wild type." When normal chains from different animals are compared, even when their homology is duplication-independent (for the definition of this term see Section III), and unless the number of differences between the chains is limited to one or a few, the proba-

bility that any particular observed conversion has occurred in more than one step is indeed not negligible. On the other hand, when all molecular sites of different homologous chains are examined, as we have done, it is unlikely that conversions that are found with particular frequency require more than one mutational step.

If we consider conversions that occur evolutionarily at 30% or more of the sites of at least one of the amino acid residues in the couple, 38 conversions, defined in the matrix of Table II, require probably no more than one mutational step according to the frequency data. As many as 10 of these conversions require a minimum of two mutational steps according to Eck's proposal for a complete genetic code (16). Of these 10, 5 appear, however, as allowed one-step conversions in the set of code triplets of Wahba *et al.* (see Jukes, 37, Table I), namely his–lys (conversion frequency = c.f. = 30% at his sites), ilu–ala (c.f. = 39% at ilu sites), met–leu (c.f. = 89% at met sites), asn–ala (c.f. = 33% at asn sites), and glm–asp (c.f. = 56% at glm sites). Not allowed by either code as one-step conversions are try–phe (c.f. = 60%, nonsignificant because of there being so far only a few tryptophan sites known in globins) and met–val (c.f. = 45%). The most remarkable discrepancy with both systems of code triplets occurs for proline conversions. The conversion frequency at prolyl sites is 57% for pro–glu, 43% for pro–asp, and 43% for pro–gly. All three conversions require a minimum of two mutational steps according to either the experimental results of Wahba *et al.* or the systematization and extrapolation of Eck. It is not unlikely that, with respect to these proline conversions, the genetic code will have to be revised. The code under its present form is compatible with the conversions observed at proline sites, if it is postulated that alanine is consistently the ancestral residue. From alanine it is possible to reach proline as well as most other proline substituents in one mutational step, according to available lists of code words. The likelihood of the generality of this relationship decreases, however, with an increasing number of times it has to be postulated.

As shown in Table II, of the 190 conversions that are possible, regardless of their direction, among 20 different amino acids, 89, nearly one-half, occur relatively rarely or very rarely (conversions that occur with a significant relative frequency in one direction but not in the other are not counted), and 52, i.e. 27%, have so far not been found at all. Thus, at least in globins, roughly one-quarter of the possible types of substitutions do not occur. It is interesting to compare these figures with the minimum numbers of mutational steps as defined by Eck's code. We are indebted to Professor J. Lederberg for a table prepared by him, with the help of an IBM 7090 computer, and used in our Table II. According

to Professor Lederberg's matrix, there are, in Eck's code, 83 allowed one-step mutational conversions, which corresponds to 44% of the possible pairs of amino acids. There are 97 pairs separated by at least two mutational steps, and 10 pairs (5% of the total number of pairs) separated by at least three mutational steps. Since Eck's code is the most degenerate among those proposed on the basis of the experimental and observational data, the 107 conversions between amino acid residues that require more than one mutational step may be considered a minimum number, and the 83 allowed one-step conversions a maximal number. Only about one-half this latter number have been classified here as conservative. This suggests that there are a number of one-step conversions allowed by the genetic code that are only rarely compatible with the preservation of a given molecular function.

Indeed, a number of one-step conversions have so far not been found in globins and are therefore radical. There are 18 of these, according to a count of the squares in Table II that contain both a dot and a cross. 29 further one-step conversions have been found with a frequency below 20% in both directions of the conversion. If the genetic code perhaps favors conservative substitutions by allowing them to proceed in one step, it does not on the other hand avoid all conversions that are radical in globins. Because of the provisional character of Eck's code these indications are tentative. It is of interest to ascertain which types of conversion, although presumably of easy access mutationally, are not retained by evolution.

Some of the conversions that require more than one mutational step may be achieved by isosemantic substitution (75), which Sonneborn (this volume) calls effects of synonymy. In such cases two successive single base substitutions occur in a codon, but only the second base substitution leads to an amino acid substitution. Selective value of isosemantic base substitutions might be associated with a change in the quantitative regulation and timing of polypeptide synthesis.

The comparison between homologous "normal" polypeptide chains, rather than between mutants and wild-type, can contribute to the elucidation or to the confirmation of the genetic code, not only, as mentioned, because the most frequently observed substitutions are most likely to occur through a single mutational step, but also because substitutions may be examined at sites that are almost, but not quite, invariant. The more seldom a given site is found to vary, the greater the chances that a rare substituent be introduced via a single mutational step.

Freese (19) has defined transitions as substitutions, in codons, of a purine by a purine, or of a pyrimidine by a pyrimidine, and transversions as substitutions of a purine by a pyrimidine or the opposite. If we con-

sider the most frequently occurring substitutions listed in Table V, which are most likely to be accomplished in single mutational steps, we find that the substitutions most frequently retained by evolution comprise 60% or 63% of transversions, according to whether one uses the list of code words of Wahba *et al.* (cf. Jukes) or Eck's system. If we count only the substitutions that represent either a transition or a transversion to the exclusion of the other term, we find, according to Eck's code, 4 transitions and 8 transversions (67% transversions), and, according to Wahba *et al.*, 5 transitions and 9 transversions (64% transversions). Thus, in the case of vertebrate globins both transitions and transversions are frequently retained by evolution, transversions perhaps somewhat more frequently, although the bias cannot at this point be considered as significant.

III. The Time Dependence of Evolutionary Transformations at the Level of Informational Macromolecules

1. The Issue

At this meeting, as well as at a preceding one on the evolution of proteins that took place in Bruges earlier this year (71), objections have been raised to the ambition to express evolutionary transformations of informational macromolecules as a simple function of time. We shall propose one such formulation presently. Professor Ernst Mayr expressed the view here that evolution is too complex and too variable a process, connected with too many factors, for the time dependence of the evolutionary process at the molecular level to be a simple function. The measure of wisdom of the opposite assumption depends on its measure of success. So far, the refutations of the time function have been weaker than its formulations. For instance, to ridicule the contention that the number of differences between two homologous polypeptide chains is roughly proportional to the phyletic distance of the forms in which these chains are found, it was pointed out at this meeting that the number of differences between cytochromes *c* are 12 in the man–horse comparison and only 8 in the man–kangaroo comparison. Hence, the speaker concluded, the kangaroo, on the basis of molecular evidence, would appear to be more closely related to man than the horse, in flagrant contradiction to solid knowledge. The speaker's conclusion was, however, unwarranted, since the probable fluctuation in every comparison may be evaluated a priori as equal at least to plus or minus the square root of the number of differences. On this basis the observed ratio 12/8 for the relative distance of horse and kangaroo from man might as well be taken to be 9/11. It would be unreasonable to require that any such single measure be accurate in terms of phyletic distance. An ac-

curate measure should be approached as an increasing number of different types of polypeptide chains from two organisms are being analyzed. Counting numbers of differences in amino acid sequence is only one stage of the analysis, and recording the nature of the differences is a necessary further step in the establishment of a molecular phylogeny. When both operations are combined, chemical paleogenetics, in its aptitude to determine phylogenetic relationships, should not be fundamentally different from the immunological method; it is only potentially less equivocal, more accurate, suited for absolute instead of only relative evaluations, and able to extrapolate from the present to the past. Anyone who recognizes the value of the immunological approach for estimating phyletic distance within certain limits should find it impossible to deny that the comparison of amino acid sequences is potentially an even better tool.

In this connection one may also recall the following comment of Simpson (60) about there having been found only one difference in amino acid sequence between the human and the gorilla hemoglobin β chains. This number, he says, "has nothing to tell us about affinities or indeed tells us a lie." To be sure, the difference in question tells us little about the affinities of man and gorilla—it tells us only that they are very closely related. But it does not tell us a lie. The sensitivity of any polypeptide chain as a measuring rod for phyletic distance will increase with its length, with the magnitude of the phyletic distance measured (up to a point), and with the rate of evolutionarily effective amino acid substitutions, a rate that is expected to differ for different polypeptide chains. As to accuracy, it cannot be expected to be satisfactory, except by coincidence, on the basis of one single type of polypeptide chain. Obviously, the fact that one single, rather slowly evolving type of polypeptide chain does not tell us much about the phyletic distance of two closely related organisms does not demonstrate a disadvantage of the molecular approach to phylogeny.

Ernst Mayr recalled at this meeting that there are two distinct aspects to phylogeny: the splitting of lines, and what happens to the lines subsequently by divergence. He emphasized that, after splitting, the resulting lines may evolve at very different rates, and, in particular, along different lines different individual systems—say, the central nervous system along one given line—may be modified at a relatively fast rate, so that proteins involved in the function of that system may change considerably, while other types of proteins remain nearly unchanged. How can one then expect a given type of protein to display constant rates of evolutionary modification along different lines of descent?

One of us (71) has recently drawn a generalization that, if valid,

should, in principle, vindicate Professor Mayr's misgivings. We are referring to the following postulate: A contemporary organism that morphologically closely resembles an ancestor of another contemporary organism also closely resembles that ancestor with respect to the amino acid sequence of most of its polypeptide chains. If no considerable morphological difference between two organisms is observed, whatever their relationship in time, no considerable change in the majority of their polypeptide chains is expected. The consequences that result from this postulate do not appear to be borne out by most of what is now becoming apparent about evolutionarily effective rates of amino acid substitutions in hemoglobin chains. The evidence is so far not extensive enough for a final conclusion to be reached, nor is it altogether unequivocal. We find it worth while to discuss the issues involved.

To this effect, a system of nomenclature of hemoglobin chains that is appropriate for the study of their evolution must first be adopted.

2. *A System of Nomenclature of Polypeptide Chains in Relation to Gene Duplication during Evolution*

The different human hemoglobin chains have originally been defined by their N-terminal sequence (57). This sequence may change partially or totally during evolution, and yet the structural genes that control a changed and an unchanged chain of a given type may still be homologous. We need to be able to refer to homologous structural genes and to the corresponding polypeptide chains irrespective of the actual sequences involved. We shall therefore redefine operationally the human hemoglobin chains and the corresponding genes in terms of a particular gene lineage rather than in terms of a particular amino acid sequence. The common ancestor of two or more of the human hemoglobin genes is designated by juxtaposition of the symbols that relate to the genes derived by the postulated duplication. Thus the common ancestor of the β chain gene and the δ chain gene will be called the β–δ gene, the common ancestor of the β, δ, and γ genes, the β–γ–δ gene, and so forth.

The different α chains found in higher Primates have presumably arisen not by gene duplication but by simple filiation from a single gene, and similarly for the different β chains, and so on. They thus possess another degree of homology than the α chain and the β chain, which have presumably become distinct through the duplication of an ancestral gene. Thus it is useful to distinguish duplication-dependent homology from duplication-independent homology. Genes related by duplication-dependent homology occupy nonhomologous chromosomal loci; i.e., they are nonallelic or at most pseudoallelic (very closely linked). Genes related by duplication-independent homology may or may not occupy

homologous chromosomal loci, since they may or may not have changed place by translocation in the course of time. In fairly closely related forms, however, it is probable that no evolutionarily effective translocation of a given gene has occurred since the time of their common ancestor.

The homology of chromosomal loci could so far be demonstrated only in organisms that give rise to viable crosses, i.e., in closely related forms. Hoyer, McCarthy, and Bolton (26) have now developed a method whereby the degree of homology of stretches of DNA from any source can be tested *in vitro*. This method seems promising in this connection and in others. When the common ancestor of two contemporary forms is too remote, such a thorough reshuffling of the hereditary material may have occurred that the notion of homologous genic loci has become meaningless. Duplication-independence of homology can in such cases be deduced only through the examination of relatively unmodified descendants of forms that were intermediate between the very distant ancestor and his very modified descendants. The validity of such a procedure will rest on the validity of the above-mentioned postulate, according to which a near-constancy in the morphological traits of a type of organism is accompanied by a high average stability of its genome.

Thus, in our studies, hemoglobin structural genes or polypeptide chains, as designated by Greek letters or groups of Greek letters, are defined by their participation in a pedigree in which the original ancestor directly results from the last gene duplication that one must assume to have occurred in the ascendancy of the genes and polypeptide chains. For instance, when we say that two hemoglobin polypeptide chains are both β chains, they are so called, not on account of an identity of their N-terminal amino acid sequence, nor because they both function as the adult major-component non-α chain and are thus under the same type of control as the human β chain as to period and amount of synthesis, but because between the common chain ancestor and the two descendent chains no evolutionarily effective gene duplication presumably occurred.

On this basis, the adult major-component non-α chain of, say, horse should not be called a β chain, since the duplication of the β–δ gene, to yield the β gene and the δ gene, has presumably occurred in the line of descent of the Primates *after* the time of the common ancestor of the Primates and the Ungulata (see Section III.6). Let us call any adult major-component non-α chain a "β" chain (in quotation marks). The horse "β" chain may be a β–δ chain, but this is not necessarily so. Between the gene duplication that gave rise to the β–δ gene and the γ gene, and the later duplication of the β–δ gene that gave rise to the

β gene and the δ gene, one or more further gene duplications may conceivably have occurred, and the horse "β" chain may be related by duplication-independent homology to one of the daughter genes resulting from these latter duplications. How shall we confer on the nomenclature the flexibility that will allow one to insert gene duplications, as they are discovered, at places where the accepted lettering leaves no room for them? We shall refer to genes resulting from such duplications by the Greek letters, with the index "prime," "second," etc., that pertain to the genes resulting from the oldest previously known duplication. This procedure is illustrated on Fig. 2, where dotted lines represent a hypothet-

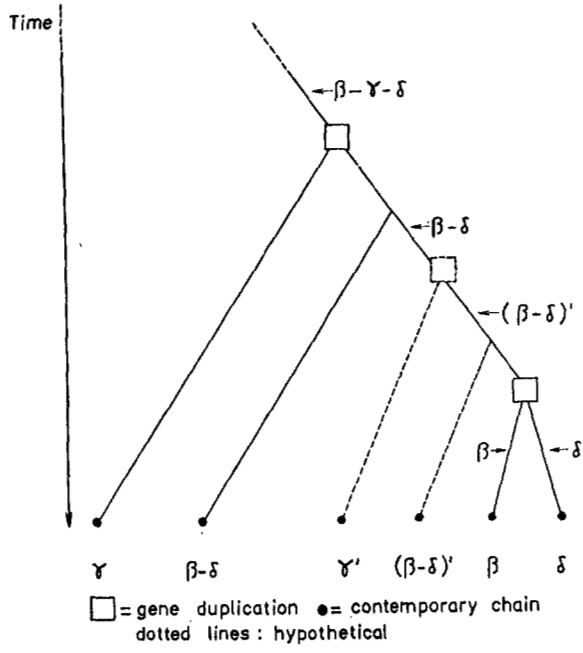


FIG. 2. Nomenclature of hemoglobin polypeptide chains.

ical gene duplication, considered for the purpose of explaining the nomenclature. This nomenclature is imperfect because the topological relationship of a gene with the index prime, second, etc., cannot be deduced from the nomenclature, but must be read off a graphically presented tree. The present system of nomenclature is therefore provisional.

Further specification will be provided by the subscripts E, F, and A for early embryonic, later fetal, and adult hemoglobin chains, and by the subscripts M and m, for major and minor components in relation to their proportions found.

Since, according to present usage, adult non-α chains are frequently called β chains, and fetal non-α chains, γ chains, we shall, when referring to this type of nomenclature, put the Greek letters between quotation marks. For instance a chain or gene designated as $(\beta\text{–}\delta)_{FM}$ is a "γ" chain or gene.

3. *A Priori Considerations on the Rate of Change of Primary Structure in Proteins*

We shall now return to the issue of the time-dependence of evolutionarily effective mutations. The postulate relative to the correlation between morphological stability and stability of the base sequence in structural genes appears to imply that the rate of evolutionarily effective substitutions in homologous polypeptide chains will in general be very unequal along two lines of descent, one characterized by a great evolutionary stability of the organism as a whole, the other by a relatively high rate of evolutionary change of the organism as a whole.

What is meant in this connection by "organisms as a whole" remains, however, undefined. It is possible that striking evolutionary changes may be correlated with changes in genes that represent only a very small to moderate fraction of the genome. If this is so, there is only a very slight to moderate increase in probability that any given type of gene will be substantially changed along lines of descent characterized by a rapid evolutionary rate. Thus, if we compare hemoglobins or cytochromes *c* from organisms whose ancestors we have reason to believe to have evolved at very different rates, the rates of change in the hemoglobins or cytochromes may not share in this difference. This may explain the apparent approximate constancy of the rate of evolution of the "mammalian type" of cytochrome *c* over very long evolutionary times (41). Complete data about hemoglobins and related globins in relation to a lapse of evolutionary time of comparable magnitude are not yet available. One may expect a priori that the rate of evolution of hemoglobin will have varied more frequently than the rate of evolution of cytochrome *c*. The functional requirements to be met by cytochrome *c* are probably constant over a wider variety of organisms and of ecological niches than the functional requirements to be met by hemoglobin. However, even in the case of the hemoglobins, the expectancy of a relative constancy of rate of change during evolution may be justified on the basis of some considerations to be developed below.

To play it as safe as possible when we want to postulate comparable rates of evolution of a given type of polypeptide chains along different lines of descent, we ought preferably to compare homologous polypeptide chains, either as found in organisms whose ancestors are known to

have evolved at comparable rates (or, better, in organisms in which the function the polypeptide is participating in, say the respiratory function, has evolved at comparable rates—if rates are definable in such terms), or as found within the same organism. The latter type of comparison, say between the human α and β hemoglobin chains, or between the human α and γ hemoglobin chains, may be the safest one to be used for an evaluation of an absolute time of common molecular ancestry. Indeed, in that case, there is no difference in rate of change between two organisms to be taken into account. Although it is by no means thereby implied that the past rates of evolution of the two chains that are being compared have been constant, at least the chances seem best that the factors that influence the rate of evolution will have influenced the evolution of the two chains similarly.

Further potential factors of inequality of rates of polypeptide evolution are in the difference between major and minor components of a given type of polypeptide chain—this applies to the comparison between the human hemoglobin β and δ chains—and the Ingram effect, namely a probable difference in rates of evolution between, on the one hand, polypeptide subunits that form components of several versions of a protein with a given quaternary structure and, on the other hand, polypeptide subunits that participate in only one such version. The latter, indeed, have to be adapted to but one different type of partner chain, whereas the former have to sterically fit several different types of partner chain. The magnitude of the Ingram effect may, however, be expected to be small. It will increase with the proportion of amino acid residues that are involved in the interaction between unlike chains in the quaternary structure, but this proportion will usually not be considerable. Moreover, if we take hemoglobin as an example, the relative expected invariance of the α chain will in turn decrease the variability, along the surfaces of contact, of all the non-α chains. Like chains will also interact over certain surfaces, and so the more significant generalization may be that polypeptide chains of a given type should in general evolve more rapidly when they exist as monomers than when they exist in association. This would make it difficult, for instance, to evaluate the time of the common molecular ancestor of a human hemoglobin chain and human myoglobin. This comparison may be further compromised by the possible significance in relation to evolutionary rates of the fact that the functions of hemoglobin and myoglobin are slightly different and that the two types of molecules carry out their function in different tissues. We must be aware of the factors that may influence the rate of evolutionarily effective amino acid substitution, although all these effects may turn out to fall within the limits of the statistical fluctuations. For

instance, the comparison of the amino acid sequences of the human (23) and sperm whale (58) myoglobin chains does *not* suggest that the myoglobins have evolved faster than the hemoglobins.

4. *The Pattern Presented by the Number of Differences between Various Hemoglobin Polypeptide Chains*

In spite of the objections against using numbers of differences between polypeptide chains for evaluations of the absolute time at which their common chain ancestor presumably existed, available data show an encouraging measure of consistency. This applies to the cytochrome *c* data discussed by Margoliash and Smith (this volume), and it applies to the hemoglobin data presented in Table VIII. In this matrix the numbers of differences between various mammalian hemoglobin chains are recorded. (Deletions or insertions are not counted, and where a "hole" is observed in one chain residues that fill the hole in other chains are disregarded.) Many figures are only approximate, because the analysis of a given polypeptide chain has not yet been completed, or because amino acid sequence was not established directly but deduced mostly, with good probability, from the amino acid composition of tryptic peptides, by reference to a homologous tryptic peptide of known sequence.

Although an approximate constancy in rate of evolution of different hemoglobin chains is not borne out by all the data in this matrix, it is strongly supported by the majority of them. The agreement between the numbers of differences that distinguish all α chains from all non-α chains is surprisingly good. This seems to imply that the fetal non-α chains have evolved at an average rate that is not significantly different from the average rate of evolution of the adult non-α chains. More generally, it also seems to imply that, since the rather remote time of their common ancestry—several hundred million years (see Section III.6)—all α chains and non-α chains recorded in the table evolved at similar average rates.

Some of the values relative to the cattle "β" and "γ" chains will be discussed in Section IV. The most aberrant value recorded in the table, which is far outside the expected range of statistical fluctuations, is the number of differences between the horse and cattle α chains. There may be something peculiar about the evolutionary history of the cattle α chain.

Aberrant values are expected and are not very disturbing, as long as they appear occasionally rather than generally, as seems to be the case on the basis of our rather small sample of data. The difficulty lies, however, in the interpretation of such inconsistencies. The most striking inconsistency that has so far turned up in the hemoglobin field involves

TABLE VIII

NUMBER OF DIFFERENCES BETWEEN SOME MAMMALIAN HEMOGLOBIN CHAINS[a]

	α man	α horse	α cattle	β man	"β" horse	"β" cattle[b]	γ man	"γ" cattle
α man	0	17	~27	74	81	≈75	79	82
α horse	17	0	~38	77	75	≈77	77	77
α cattle	~27	~38	0	~81	~83	~83	~81	~88
β man	74	77	~81	0	26	≈27	39	32
"β" horse	81	75	~83	26	0	≈35	43	33
"β" cattle[b]	≈75	≈77	≈83	≈27	≈35	0	≈45	≈28
γ man	79	77	~81	39	43	≈45	0	~40
"γ" cattle	82	77	~88	32	33	≈28	~40	0

[a] Differences due to deletions are not counted.

[b] Estimated on the basis of 65% of the cattle β chain (composition of tryptic peptides).

the adult non-α chains of primitive Primates, analyzed by Hill and the Buettner-Janusches (24). They find, for instance, 6 differences between the α chain of *Lemur fulvus* and the human α chain, and 27 differences between the corresponding β chains. It is significant that the observed number of differences between the α chains is in line with other data. We do not believe that the Ingram effect can account for the unexpectedly high number of differences between the β chains. From the small sample of data presented in Table IX it appears that the Ingram

TABLE IX

MINIMUM NUMBER OF AMINO ACID SUBSTITUTIONS IN ANIMAL HEMOGLOBIN CHAINS AS COMPARED TO THE CORRESPONDING HUMAN HEMOGLOBIN CHAINS[a]

Species	α chain	β chain	Reference
Horse	17	26	(12, 62)
Pig	≈18[b]	≈14[b]	(10)
Cattle	≈27	≈27	(53)
Rabbit	~27		(14)

[a] Mean number of differences for the pooled α chains and β chains: 22. Approximate time for common ancestor of man, horse, pig, cattle, and rabbit: 80 million years. Mean period of time between evolutionarily effective amino acid substitutions: 7×10^6 years according to this sample of data.

[b] Computed from peptide composition studies relative to 72% of α chain and 81% of β chain.

effect, if present at all, may be blurred by other factors and may be disregarded at least in certain groups. Table IX also presents a re-evaluation, which is provisional, like our first evaluation of this type (74), of the mean length of time that elapses during two successive evolutionarily effective amino acid substitutions in hemoglobin chains. The order of magnitude has remained the same.

The main alternative in interpreting the large difference between the *Lemur* "β" chain and the human β chain appears to be as follows. Either the number of evolutionarily effective amino acid substitutions is not even roughly proportional to time in the case of a comparison between lower Primates and man, or the *Lemur* "β" chain and the human β chain are not related by duplication-independent homology.

One may accept very large fluctuations in the rate of evolutionarily effective amino acid substitutions and reconcile them with the consistency of the cytochrome *c* results and with that of the comparisons between the hemoglobin α chains and non-α chains on the basis of the proposal of Margoliash and Smith (41) that over long stretches of evolutionary time fluctuations in the rate of structural change of polypeptide chains will tend to cancel out. If, however, some especially potent selective pressure caused the major adult hemoglobin component of the an-

cestors of *Lemur* to evolve at an especially rapid rate, why did the α chain in these ancestors not also evolve more rapidly than expected? In view of the interactions between the two types of chains, such a discrepancy in evolutionary behavior would be puzzling.

On the other hand, the postulate according to which the *Lemur* adult non-α chain is controlled by a structural gene that resulted from a gene duplication other than the one that gave rise to the human β gene is, at the moment, illegitimate, although perhaps correct, because this type of postulate may be used to explain away any uncomfortable discrepancy. It is perhaps significant that the N-terminal amino acid in the adult non-α chain of *Lemur fulvus* is threonine (23). N-Terminals in vertebrate hemoglobins are stable, and threonine had so far not been found in that position. Perhaps the *Lemur* and human "β" chains are related by a duplication-dependent homology. A reasonable explanation of the observations is that the *Lemur* "β" chain is in fact a γ chain.

5. *Interpretation of the Apparent Constant Rate of Evolutionary Change in Most Polypeptide Chains*

If we provisionally accept the conclusion to be drawn from the data recorded in Table VIII—namely that the mean rates of evolutionarily effective amino acid substitutions are usually comparable, not only along different lines of descent, but also for different hemoglobin polypeptide chains—we must explain why this is so, in spite of reasons to expect the opposite. Several tentative explanations may be proposed.

It may be that differences between organisms in terms of amino acid sequence of polypeptides are more nearly a simple function of time than are differences due to changes in the control of rate and period of synthetic activity of structural genes. Some of the especially rapid evolutions, say along the hominid line, may be due to a significant extent to changes in the control of gene activity (69, 70) and not be reflected in significant changes in the rate of evolution of most polypeptide chains.

It may also be, as mentioned by Prof. Mayr, that during phases of exceptionally rapid evolution only certain systems and, hence, probably only certain types of polypeptide chains are structurally more affected than would be expected from their normal rate of change. The great majority of the genes may continue to change at a rather slow and perhaps relatively regular rate. A small sample of proteins may, by accident, show that evolution has been exceptionally fast along a given line of descent, but a larger sample of proteins may establish the basic stability of the organism in spite of the important phenotypic variation that is observed. A vastly predominant number of unmodified or nearly unmodified genes may participate in the formation of an organ that appears

significantly changed with respect to relative dimensions and other properties. At the level of the structural genes the constancies may far exceed the changes when, as judged from the "looks" of two organisms, the changes seem to exceed the constancies. Independently of the question concerning the relative contributions of structural genes and controller genes (which are distinct at least in their action, if not otherwise), it is therefore likely that the ratio of differences to similarities can be determined at the polypeptide level with a type of significance that does not obtain elsewhere and cannot be matched at the level of the phenotype as seen by the organismal biologist.

Perhaps the most important consideration is the following. There is no reason to expect that the extent of functional change in a polypeptide chain is proportional to the number of amino acid substitutions in the chain. Many such substitutions may lead to relatively little functional change, whereas at other times the replacement of one single amino acid residue by another may lead to a radical functional change. That this is so is at present amply documented (cf. Section II). Therefore an abnormally rapid change in phenotype along a given evolutionary line need not imply an abnormally high rate of evolutionarily effective amino acid substitutions even in those proteins that are most directly involved in the evolutionary change. It is the type rather than the number of amino acid substitutions that is decisive. Of course the two aspects are not unrelated, since the functional effect of a given single substitution will frequently depend on the presence or absence of a number of other substitutions. But if, for bringing about significant functional changes, the emphasis is on type rather than on number of amino acid substitutions, periods of rapid evolutionary change need not be expressed by a substantial increase in rate of amino acid substitution at the polypeptide level.

If this view is correct, it leads to an interesting consequence. It would then appear that the changes in amino acid sequence that are observed to be approximately proportional in number to evolutionary time should not be ascribed to vital necessities of adaptive change. On the contrary, the changes that occur at a fairly regular over-all rate would be expected to be those that change the functional properties of the molecule relatively little—namely the so-called isogenetic or conservative substitutions (see Section II). As we saw, this point of view is confirmed—and was independently suggested—by the great majority of substitutions that we actually observe in the comparison of hemoglobin polypeptide chains whose presumed common molecular ancestor is not excessively remote in evolutionary time.

There may thus exist a molecular evolutionary clock. The basic rate

of evolutionarily effective substitutions may express the degree of plasticity or of looseness in the relationship between a given rather narrowly defined molecular function and amino acid sequence. If this plasticity is great, the observed rate of evolutionarily effective change will be high, because a relatively large proportion of the amino acid substitutions that occur by chance mutation will be slightly advantageous. If the plasticity of the amino acid sequence in relation to molecular function is very restricted, if only a few substitutions will change molecular function slightly rather than radically, only a smaller proportion of the mutations will be advantageous, and the average time elapsing between two evolutionarily effective substitutions will be larger. *Superimposed* on this basic rate of molecular change are then the functionally highly significant changes. Since the latter changes are relatively rare, the rates of evolutionarily effective amino acid substitutions during periods of rapid evolution may not substantially differ from the rates that obtain during periods of slow evolution. The molecular evolutionary clock would work through a shuttle motion at a certain number of molecular sites from one to the other among a small number of amino acid residues whereof, according to circumstances, the ones or the others are slightly more advantageous for the organism. Hence there arise the multiple molecular coincidence or "convergence" effects that are observed (see Section V). The basic rate of evolutionary molecular change would in a sense be comparable to moderate thermal motion, in that the molecule is reversibly altered structurally without loss of function and at a constant average rate. Each type of polypeptide chain would, in this sense, have its specific evolutionary "temperature."

What is the relationship between this theory and the above-mentioned postulate according to which most polypeptide chains in morphologically stable organisms have stable amino acid sequences? These two tentative generalizations are compatible only if the postulate is amended in the following way: Along lines of descent marked by high evolutionary stability, the shuttle motion between functionally similar amino acid residues will also occur. The changes in amino acid sequence will, however, be limited almost exclusively to the functionally nearly neutral changes. The rate of the functionally nearly neutral evolutionarily effective amino acid substitutions may be lower in organisms that live in a very stable environment than in organisms that live in a more variable environment, because in the very stable environment only few nearly neutral substitutions may be slightly advantageous and thereby spread in the population. In organisms exposed to a more variable environment, a larger proportion of nearly neutral amino acid substitutions may turn out to be slightly advantageous at one time or another,

even during periods marked by a relative over-all evolutionary stability.

Which of the two ideas, if any, applies to a greater extent—the idea of an over-all stability of the amino acid sequence of polypeptide chains along lines of descent where little change in "organismal" phenotype occurs, or the idea of a molecular evolutionary clock in which all organisms participate, irrespective of the rate of evolutionary change of the phenotype? The verdict of the facts will be looked forward to.

6. *Quantitative Relationship between Number of Observed Differences in Homologous Polypeptide Chains and the Time of Their Common Ancestry*

On the basis of the postulate encouraged by observation, that the rate of evolutionarily effective amino acid substitutions in hemoglobin chains is comparable along different lines of descent, the relation between the number of observed differences between two homologous polypeptide chains and the time of their common ancestry may be formulated as follows.

Let the chance of a mutation's affecting one amino acid site of a polypeptide chain in time t be αt. For N sites, the total number of mutations is about $N\alpha t$.

Let us consider the time, t, at which $N\alpha t$ mutations have occurred. We assume equal probability for all sites. The probability that the first site is unchanged after the first mutation is

$$P_1 = \frac{N-1}{N} = 1 - \frac{1}{N}$$

and after $N\alpha t$ mutations is

$$P_{N\alpha t} = P_1^{N\alpha t} = \left(1 - \frac{1}{N}\right)^{N\alpha t}$$

At the limit, as N increases to infinity,

$$\lim_{N \longrightarrow \infty} \left(1 - \frac{1}{N}\right)^{N} = \frac{1}{e}$$

Hence[2]

$$P_{N\alpha t} \cong \left(\frac{1}{e}\right)^{\alpha t} = e^{-\alpha t} = A^t$$

where $A = e^{-\alpha}$.

Let us use the expression $P(t)$ as the average probability for any given site to be unchanged at time t. Let τ be a unit lapse of time. Then

$$P(n\tau) = P^n$$

---

[2] This equation is an application of well-known probability theory. An equivalent equation is used by Margoliash and Smith (this volume).

In the present treatment back mutations are ignored because of uncertainty of their probability and because accuracy of experimental points does not justify a refinement of the theory.

Consider a type of protein such as the globins with practically all sites changeable. Table IX shows that the mean difference between some mammalian hemoglobin chains and the corresponding human

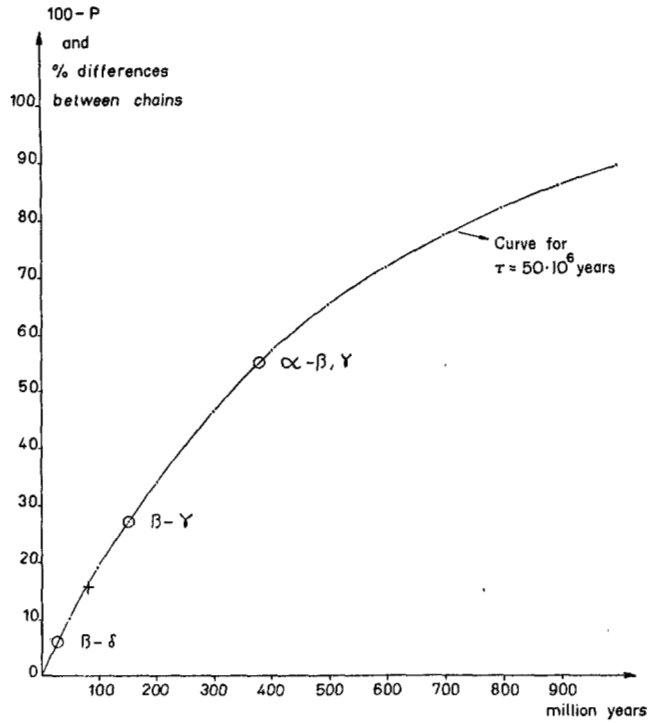


FIG. 3. Epoch of common chain ancestors as a function of the proportion of differences between polypeptide chains. + = point for difference between human and some other mammalian hemoglobin α and β chains (cf. table IX), with common ancestry assumed at —80 million years. O = observed per cent differences between hemoglobin chains, inserted on curve as defined in part by the above assumption.

chains is 22. This figure represents 15% of the hemoglobin chains, in relation to the mean length of the α and β chains (143.5 residues). We assume the mean epoch of separation of man and that of each of the mammals listed in Table IX to be at —80 million years. Thereby τ is defined as on Fig. 3. It is approximately equal to 50 million years. The average α–β and α–γ change is 55%. On the curve of Fig. 3 this gives the α–β,γ separation at —375 million years. This figure is one-third lower than

the one obtained by a previous preliminary evaluation (74). The epoch of separation of other chains is also indicated on Fig. 3. The more remote the epoch of the common molecular ancestor, the larger is the error due to ignoring back mutations.

The cytochrome *c* data will fit this curve and lead to a reasonably remote epoch for the common ancestor of yeast and man only on the basis of the assumption that a large proportion of the molecular sites in cytochrome *c* are evolutionarily invariant for a reason other than that of statistical accident. The necessity of this assumption in relation to the calculation constitutes an argument in favor of the assumption (cf. Section II.1).

It is appropriate to compare one protein with another by reference to the rate of evolutionarily effective mutations *per amino acid residue.* The observed rate (the probability) of amino acid substitutions is to be divided by the number of residues in the polypeptide chain. In hemoglobin chains, if we accept the provisional figure (Table IX) of one evolutionarily effective mutation per 7 million years, we find that the chance for any given site to undergo an evolutionarily effective mutation along a given line of descent is on the average about 1 in 1 billion years. One may refer the rate of evolutionarily effective amino acid substitution to the number of evolutionarily changeable residue sites, rather than to the total number of sites. Since, according to present counts, there are about 117 evolutionarily changeable sites in hemoglobins (including the lamprey chain), the chance for any given changeable site to be changed evolutionarily is on the average about 1 in 800 million years. In fact, the chances vary, of course, for different sites of the molecule, but the average figures can be used in comparisons between proteins.

IV. Gene Duplication in Globins

Gene duplication may be associated with the following distinct processes: (a) duplication of the whole genome during cell division; (b) intralocus multiplication of some or all genes in the course of cell differentiation; (c) interlocus duplication of individual genes or duplication of whole individual chromosomes in the course of evolution.

The three types of processes differ, in part, by the order of magnitude of the time during which the duplicate genes are maintained in a given type of cell.

We need not comment on the first process and shall refer only briefly to the second hypothetical process.

Itano, Neel, and Wells (cf. 33) evaluated the proportions of different hemoglobin components in members of families that included heter-

ozygotes for normal adult hemoglobin, HbA, and an abnormal hemoglobin, namely HbS or HbC. The HbA components found in different families appeared to be structurally identical as far as could be determined by the available criteria, yet the proportions of the normal and the abnormal hemoglobins in heterozygotes was found to be different. As heritable characteristics, these different proportions were assumed to be under genetic control. It was further suggested that the proportions of hemoglobin components as found in the red cells of the peripheral circulation reflect relative rates of synthesis of these components. The apparently identical alleles that differed by the rate of their synthetic activity were called "isoalleles." In a seminar given at the California Institute of Technology in 1961, one of us (E.Z.) had noted, on the basis of the available family studies, that the ratios of hemoglobin components found in the offspring are in accordance with expectation, if one assumes that the relative amounts of chain synthesis in the parents represent small heritable integral multiples of a basic polypeptide production unit. This way of "quantizing" the results of Itano and his collaborators introduced the question of a fit between the situation found in the parents and that found in the offspring without reference to the "modifying genes" considered by Neel *et al.* (47). The presentation of the data pointed to the possibility, also considered by Nance (46), that several contiguous duplicates of the same gene are usually active in hemoglobin synthesis. We shall go neither into the examination of the data nor into the reasons to doubt that the gene duplication hypothesis, as developed in detail by Nance, offers the correct explanation of the facts. What we want to point out is the following: If the apparent ratios of integers, whose possible existence is suggested by the data of Itano *et al.*, should turn out not to be due to coincidence, then they might be attributed tentatively not only to genic multiplication resulting in the creation of new loci on the chromosome, but at least equally well to intralocus genic multiplication. This process is conceived as a somatic occurrence during development, which is directed differentially and heritably for different genes and in different tissues and represents a means of control of rate of protein synthesis. Some genes may be thought to display intralocus multiplication in one tissue, and other genes in other tissues. Possibly the giant chromosomes in salivary glands of Diptera represent the extreme case in which all loci undergo a high order of intralocus multiplication, although it is also possible that this phenomenon is qualitatively different from the one envisaged here. Support for the idea of intralocus genic multiplication may be derived from the observation of DNA "puffs" in certain regions of polythene chromosomes of cells engaged in active protein synthesis (17, 67). A

slight structural alteration of the gene, as occurs in most abnormal hemoglobin genes, might reduce intralocus genic multiplication to a smaller order. Isoalleles might be alleles that are structurally identical but differ in their degree of intralocus multiplication. This degree would be a heritable property of the locus and at the same time a function of cellular conditions. It is plausible that simple ratios between orders of intralocus multiplication could be maintained. Certainly it is also possible that cryptic substitutions in the gene, either expressed in an undetected change in the amino acid sequence of the corresponding peptide, or isosemantic (75) and not expressed in this sequence, might affect the order of intralocus multiplication.

Less doubtful, although not directly demonstrated, but almost unavoidable, is the postulate of interlocus gene duplication or chromosome duplication as the basis for the existence of isogenes—i.e., nonallelic, structurally different, yet homologous structural genes. The detailed application of this postulate to the evolution of the hemoglobin chain genes was probably first made in a lecture given by one of us (E.Z.) at Dalhousie University (Halifax) in 1960 (unpublished). At about the same time Ingram (32) independently arrived at the same conclusion. Some further consequences of the existence of isogenes were examined later (74, 69).

The frequency of gene duplications, and that of evolutionarily effective gene duplications, namely of those that are retained by natural selection in a given line of descent, are unknown. In the latter connection we may examine some evidence in favor of a further gene duplication in mammals, beyond those that have already been presumed to have occurred during mammalian evolution. This evidence rests on data kindly made available to us, prior to publication, by Drs. Donald Babin and Walter A. Schroeder of the California Institute of Technology.

The tabulation of the number of differences in amino acid sequence between several mammalian hemoglobin polypeptide chains as given in Table VIII contains suggestive evidence to the effect that the fetal chain of cattle separated from the β chain at a later date than the fetal chain of man. The cattle γ chain does not appear to be a true γ chain, nor the cattle β–δ chain a true β–δ chain. The two chains probably arose through a duplication of the β–δ chain gene, distinct from the one that, in the line of descent of the Primates, gave rise to the β gene and to the δ gene. According to the system of nomenclature described in Section III of this article, the cattle "γ" chain will be called a γ′ chain, and the cattle β–δ chain, a (β–δ)′ chain.

Schroeder *et al.* (59) have shown that there are 39 differences in sequence between the human β chain and γ chain. We would expect

to find between the cattle β–δ chain (that we may call "β" chain) and the cattle γ chain a difference of 39, plus or minus the square root of 39, i.e., between 33 and 45 differences. Instead we find only roughly 28 differences. This is roughly the same number of differences as that between the β chain of man and the "β" chain of cattle, and also between the β chain of man and the "β" chain of horse. Thus, from the point of view of the number of differences in primary structure, the fetal chain of cattle behaves like a β chain in relation to other mammalian β chains. It also behaves like a β chain in relation to the γ chain of man. Indeed, there are about 40 differences in sequence between the cattle fetal and the human γ chains, i.e., the same number of differences as between the human β and the human γ chains (Table VIII).

The tentative conclusion about the evolutionary relationships of the cattle fetal non-α chain that we draw from the data in Table VIII is confirmed by some specific structural features. Drs. Babin and Schroeder found methionine at the N-terminus of both the fetal and the adult non-α chain in cattle, a character that had before been encountered only in the "β" chains of the related sheep and goat. Cattle has preserved the habitual valine at the N-terminus of its α chain. Horse α and "β" chains have valine at their N-terminus, as do all other mammalian α and "β" chains, except cattle "β," sheep "β," and goat "β." The introduction of methionine at the N-terminus hence seems to have occurred in the immediate ancestry of the Artiodactyla. The finding that methionine is present at the N-terminus of the cattle fetal non-α chain furnishes an important character shared by the cattle adult and fetal non-α chains, and an important distinction between the human and cattle fetal non-α chains.

Furthermore the cattle adult and fetal non-α chains both lack one residue in their N-terminal tryptic peptide (a peptide split off by the action of trypsin) in comparison with all other known "β" chains and with the only other γ chain whose primary structure has been established, the human γ chain. The lack of this globin site is an α chain character that is found in all structurally known α chains. It does not follow that the cattle non-α chains are phyletically close to the α chains. There is convincing evidence to the contrary. We do believe it to be significant, however, that, in a character as important from the evolutionary point of view as a deletion, the cattle fetal non-α chain is identical with a cattle adult non-α chain and different from the human fetal non-α chain.

A number of characters of sequence are common to the cattle adult and fetal non-α chains while different from the human γ chain, for instance at sites E15 (phe–leu), H7 (phe–try), and H10 (val–met); however, a few other characters of sequence are common to the human

and cattle fetal non-α chains, and differ in the cattle adult non-α chain. Therefore the situation, from this point of view is not entirely unambiguous. However, methionine and tryptophan are rare residues in hemoglobins, and their presence in the human γ chain and absence from the other non-α chains under consideration is more meaningful than any of the other pertinent relationships that are observed.

On the basis of the evidence, we propose the relationship between chains shown in Fig. 4. In this figure, the vertical dimension is propor-

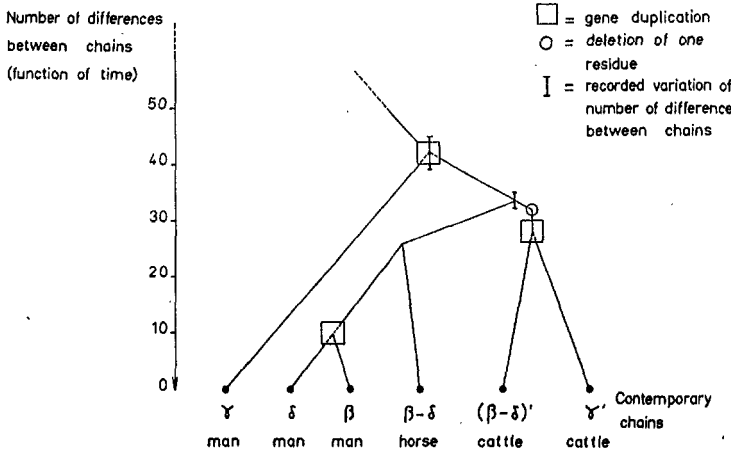


FIG. 4. Probable evolutionary relationship of some mammalian hemoglobin chains.

tional to the number of differences between chains as given in Table VIII. An ancestral gene has duplicated to yield two daughter genes, one of which has become the human γ gene, and the other the horse β–δ gene and the β–δ gene present in the descent of the Primates prior to the appearance of the β and δ genes. Not a great many million years before the "Artiodactyl duplication," the ancestor of the cattle non-α chains lost a residue at or next to the N-terminus (which term of this alternative obtains cannot at present be ascertained) and adopted methionine as its N-terminus. The "Artiodactyl duplication" then yielded two daughter genes, one of which continued to be used as an adult major-component non-α chain in cattle, whereas the other was adopted for use as the fetal non-α chain in cattle. Figure 4 suggests that man and horse are slightly more closely related than man and oxen, but this piece of molecular evidence cannot be taken seriously as long as it remains single.

As mentioned, the absence of one residue at or next to the N-terminus of non-α chains seems to be limited to a relatively small group of mammals. We may therefore assume that we are dealing with a deletion

of the residue in this small group, rather than with the addition of a residue in all other known non-α chains.

On the other hand, an *addition* of a residue at this molecular site may actually have taken place in a molecular ancestor of the β chain. This ancestor should be remote, but more recent than the gene duplication that led to the differentiation between α chains and non-α chains. Indeed, the very ancient lamprey hemoglobin chain (whose partial sequence was recently made public by Dr. Rudloff from Prof. Braunitzer's laboratory, 52), according to the most probable homology relations with other hemoglobin chains, displays at globin sites 1β, 2β, and 3β the sequence valine–leucine–serine, identical with the N-terminal sequence of the human α chain, although in the lamprey chain the first of these globin sites is preceded by nine others that had not so far been found elsewhere. Hence it seems that, some time after the duplication that led to the α gene on the one hand and to the β–γ–δ gene on the other, the addition of one residue in position 2β, namely histidine, occurred in the β–γ–δ line of descent. Later on, after the β–δ line had separated by a new gene duplication from the γ line, and the bovine β–δ line had separated, without gene duplication, from other mammalian β–δ lines, this added residue was again lost. The identity with respect to the absence of this residue of the α chains and the cattle non-α chains is the first example of a probable evolutionary convergence or coincidence at the molecular level that involves a deletion or an addition. It also illustrates a point that is a priori probable, namely that evolutionary changes at any given site of an informational macromolecule are potentially reversible. The rule of Dollo relative to the irreversibility of evolution is indeed not expected to apply at this level.

We noted earlier (Section II.2) that the lamprey chain differs from the other structurally known globin chains at very invariant sites at which all other known hemoglobin and myoglobin chains are identical. This, in conjunction with the presence of a unique N-terminal "tail," suggests the possibility that the common ancestor of the lamprey chain and mammalian hemoglobin chains may be even more remote than the very ancient common ancestor of mammalian hemoglobin and myoglobin chains. On the basis of apparent evolutionary rates of hemoglobins, it would not be surprising if the common molecular ancestor of mammalian and lamprey hemoglobin chains had existed at a time more remote than the Silurian, the geological era from which the oldest known remains of vertebrates have come down to us. This would mean either that the vertebrates are actually an older phylum than one thinks on the basis of present geological evidence, or that the common molecular ancestor under consideration existed in a prevertebrate. In the latter case the descent of the vertebrates

would not be monophyletic, and the Cyclostomes would descend from a prevertebrate independently of the mammals.

One may presume that the contemporary lamprey hemoglobin chains have arisen by one or more evolutionarily effective gene duplications in the direct ancestry of the lamprey. Three is the probable minimum number, in view of the likelihood that the four distinct lamprey chains are controlled by nonallelic genes (1, 2). In fact, six structurally distinct components have recently been identified in *Petromyzon marinus* (40).

The recent discovery by Huehns *et al.* (27) of the human ε chain added a fifth unit to the group of nonallelic human hemoglobin genes. The lack of allelism between the ε gene and any other human globin gene can be inferred, with some probability, from the fact that the ε chain differs from other chains by more than one character of sequence, and more safely from its unique behavior with respect to its period of synthetic activity. Thus, a further evolutionarily effective gene duplication (or chromosome duplication) must be postulated to account for the existence of the ε gene.

The total number of probable evolutionarily effective gene duplications of globin genes, on the basis of the data available at this writing, is about 10, and it is 6 in the human line of descent if we include the common ancestor of man and lamprey.

Suppose that the time at which the β–δ gene duplicated is close to that of the most recent evolutionarily effective globin gene duplication in the human line of descent, and suppose that this length of time is not very far from a mean figure for evolutionarily effective duplications of globin genes per unit time; we may then evaluate the order of magnitude of the number of evolutionarily effective gene duplications as 10 in 400 million years (cf. 74) per line of descent, i.e., since the approximate time of the rise of the Vertebrates. The figure checks quite well with the number of presumed gene duplications in the human line of descent since the common ancestor of man and lamprey—a number that will no doubt be increased by at least several units as more information about hemoglobin sequences and nonallelism of genes becomes available. In facts, the true number of evolutionarily effective duplications of globin genes may be significantly larger, since neither of the two estimates given above takes into account the disappearance of globin genes, either by bodily elimination, or by their conversion into dormant genes (74), or by their transformation into a gene so different that its polypeptide product is not recognizable as a globin, although selected for because fit to carry out a novel, useful function.

If we assume that the evolution of man has started with a single gene 2 billion years ago, and if we assume that man has a complement of about 100,000 different genes as estimated by one of us (L.P.), and,

furthermore, that gene duplication accounts for most of the growth of the genetic material, neglecting chromosome multiplication, then 17 generations of duplicating genes are required, each of the daughter genes duplicating again at each generation. The mean lapse of time, over the whole period of evolution, per generation of a duplicating set, would be roughly 120 million years. This figure, which would apply to any single gene with average behavior, turns out to be of the same order of magnitude as the two independent estimates given above. This rough agreement supports the idea that gene duplication has played a major role in the increase of the genetic material per cell in the course of evolution toward higher forms.

V. Evolutionary Convergence and Coincidence in Protein Molecules

Evolutionary convergence is the movement that goes from different ancestral structures to similar descendent structures with similar functions. When descendent residues at homologous sites of polypeptide chains differ from the ancestral residue, they may either differ among themselves, or be identical. If we examine evolutionary periods over which a two-step mutation in any single line of descent is unlikely to occur, then the number of different descendent residues will be limited to a maximum equal to the number of one-step conversions between amino acid residues that are allowed by the genetic code. If more than that number, p, of independently evolving polypeptide chains are compared, namely $p + n$, it is evident that identical residues will be found at the molecular site under consideration in at least n descendent polypeptide chains. On the other hand, the nature of the residues retained by evolution is believed not to be determined by the chance effect of mutations. Over reasonably long periods of evolutionary time, most or all allowed one-step transitions will have had an opportunity to occur, as can be shown by a rough calculation. The residue actually found will be the one that is functionally the most favorable among those proposed to natural selection. When the residues present at a given molecular site in the two lines of molecular descent are identical, but differ from the presumed ancestral residue, this identity is therefore probably established for functional reasons.

The observed identities of this type are not necessarily to be called evolutionary convergence effects. Indeed, different functional reasons may lead to the selection of the same residue, especially when the number of allowed one-step transitions is smaller than the number of observed independently evolving homologous polypeptide chains. Not only does the function of a residue depend on its environment in the molecule; moreover, the distribution of functional properties over the

twenty amino acids provided for by the genetic code is such that, on the one hand, as we saw (Section II), different residues may carry out the same function, but also, on the other hand, the same residue different functions. In the absence of evidence to the effect that the function and selective advantage of a given residue are the same in two chains, the term convergence is biased and should be replaced by coincidence (see also ref. 69).

It appears a priori probable that a certain fraction of the coincidence effects will be convergence effects. Indeed, the homologous polypeptide chains that are being compared are too similar in structure and function for identical residues at corresponding sites not to have sometimes identical functions. In this sense one may say that the frequency of coincidence effects establishes the existence, although not the extent, of convergence in protein molecules.

Coincident amino acid conversions are very frequent indeed in the course of the evolution of globin chains. One of us has previously discussed some examples of this effect (71). In fact, coincidence effects offer an alternative of three rather than of two terms. Besides selection of the same residue for identical functional reasons (convergence) and selection of the residue for different functional reasons, there may be coincidence effects in the absence of any selection by genetic drift.

Some abnormal human hemoglobins are characterized by substituent residues that also occur as normal residues in some animal hemoglobins. There can be, in such cases, no question of convergence as long as we do not assume that the rare human mutant has been selected for. Aspartic acid instead of glycine at site E6 characterizes the abnormal human hemoglobin Norfolk and probably an orangutan hemoglobin chain (5, 6, 36). Lysine instead of asparagine at position E17 in the α chain is a character of HbG Philadelphia as well as probably of pig hemoglobin (7, 10). Aspartic acid instead of glycine is found at position A13 in the β chain of HbJ Baltimore and of the horse (9, 62). Such occurrences will allow one to study the physicochemical effects of single substitutions that occur in combination with other substitutions in other hemoglobins.

As we saw, alanine occurs at at least 51 out of 148 sites of globin chains (Table I). Conversions to proline have been found at only 9 of these sites. The conversion alanine–proline is one of the 7 allowed one-step conversions from alanine, according to Eck's code. The chances are that mutations to proline will at some time have occurred at most of the 51 alanine sites. Most of these mutations have been eliminated by natural selection and, in the case of alanine sites that are inside α helices, the reason is obvious. The disrupting effect of prolyl residues on α helices presumably is also the reason why proline has been selected for at

some alanine sites in interhelical sections of the polypeptide chain. Thus, when the conversion from an ancestral alanine to a descendent proline has occurred twice or more times independently, as may well obtain at globin site 4α (= A2) (71), proline should be present in the descendent chains for very similar if not strictly identical reasons. Such are cases of convergence.

Whether we witness convergence or coincidences without convergence, the frequency of such effects in globin chains is high. Nevertheless we may, in most cases, be assured that these effects do not lead to the establishment of false homology relations. The progressively increasing number of differences between globin chains as one compares forms further removed from each other on the evolutionary scale sets limits to the amount of changes per unit evolutionary time, and thereby to the amount of similarity between chains that can be generated by coincidence and convergence. Moreover, differences also may be generated by chance and compensate for coincidences.

What may be obscured by secondarily generated identities, at any molecular site, is the correct branching of the phylogenetic tree established for that single site, when one attempts to determine the branching independently from knowledge derived from classical taxonomy. This only means, however, that on account of coincidences with or without convergence, one cannot expect to safely establish phylogenetic relationships from similarities and differences that are observed at a single molecular site. When the information contained in the phylogenetic trees for all molecular sites of a given polypeptide chain is combined, one may expect that the resulting measurement of phyletic distance between the chains will approach reality; the more so, the longer the chain. It is intended to carry out such measurements and to compare the results with those of organismal biology.

When the organismal biologist speaks of convergence—a phenomenon that has been disturbing to taxonomists—he refers to structures of independent phylogenetic origin that have become apparently identical or strikingly similar. To inquire whether a molecular equivalent of this phenomenon exists in informational macromolecules, we may ask the question in the following way: If we consider two polypeptide chains with random amino acid sequences, is there enough evolutionary time available for the chance to be significant that a succession of evolutionarily effective mutations generates significant similarities between the two chains so that the chains appear to be homologous?

By far the largest proportion of evolutionarily effective mutations will be those that change the polypeptide within the limits of its established function and tertiary structure. Only a small, and probably very

small, proportion of evolutionarily effective mutations will be of the kind that alter radically the function and tertiary structure of the polypeptide.

Of these presumably only a small or very small proportion will lead to types of proteins that duplicate in tertiary structure and function already existing types of proteins.

Although countless pathways exist for the transformation of one polypeptide chain into another by point mutations, deletions, additions, intracistronic partial duplications, and recombinations, the assumption that a channel can be found during evolutionary time, such that each step in the pathway is functionally useful and selected for, appears to be a daring one.

The difficulty is not to imagine that the product at each step of transfunctional transformation of a polypeptide chain has some kind of enzymatic activity. Interestingly, Professor Sidney Fox (18) has detected some weak enzymatic activities in his synthetic proteinoids. It seems significant, however, that these activities are weak.

It may be a fallacy to think that, when a polypeptide chain displays a weak enzymatic activity of a given kind, this can be turned into a strong enzymatic activity of the same kind by a succession of evolutionarily effective mutations. In other words, it may be a fallacy to think that weak enzymes are the evolutionary antecedents of strong enzymes with the same type of activity.

It may well be that all polypeptide chains are endowed with weak enzymatic activities of some kind, especially in the presence of trace metals. For natural selection to retain a polypeptide, a weak enzymatic activity may be sufficient only in the beginning phases of evolution. In organisms that are already established, it is probable that a weak enzymatic activity will be selected for only rarely. For most weak activities developed, there will already be a stronger activity available, and therefore the "new" enzyme will not be retained. Moreover a weak enzymatic action of a novel kind may not suffice to encourage the evolvement of a novel function within organisms already endowed with intense metabolic activities.

A strong enzymatic activity is likely to require a much higher degree of specificity, notably in tertiary structure, than a corresponding weak enzymatic activity. From an enzyme with a weak activity to an enzyme with a strong activity of the *same* kind, the evolutionary pathway, in terms of transformations of primary, tertiary, and quaternary structure, may be most of the time as involved as the pathway between two specified functionally *different* types of proteins.

One may thus deem it likely that a new enzyme that is retained by

natural selection arises with a tertiary structure not remote from that which is finally evolved. The evolution of a functionally important new polypeptide chain may well be a saltatory phenomenon and not a progressive one. Once a rare random mutational event leads to a privileged tertiary structure combined with features of primary structure that are, together, associated in determining some intense enzymatic activity, this polypeptide, if retained by natural selection, will undergo, in the early stages of its history, a limited amount of "perfecting" steps to further the function that was attained at once (cf. Section II.1). Thereafter most of the evolutionary history of the polypeptide will be marked by variations on a functional theme rather than by the perfecting of a function.

Synthetic proteinoids seem to be of great interest mainly in relation to very early stages of evolution. A chemical model of an evolutionary process as it occurs in any of the presently existing phyla should allow the production in a test tube of polypeptides with strong enzymatic activities. The chances for this to occur by copolymerization are probably small. It would not be surprising that a template mechanism for polypeptide production is a necessary requirement for strong, specific enzymatic activities to arise, because the rare polypeptide endowed with these properties can hardly be produced reproducibly in sufficient amounts except through a template mechanism of molecular multiplication. In copolymerization experiments, such a rare polypeptide, with a rare tertiary structure, is likely to be only a "contaminant," and the observed enzymatic activity will be weak.

To achieve a pseudo-homology between two polypeptide chains of independent evolutionary origin, it would be necessary, according to the above discussion, that there exist a pathway of functional transformations such that mutations can lead the polypeptide chain in saltatory fashion from one type of strong enzymatic activity to another type of strong enzymatic activity. For this to happen it would be necessary that the number of required functional transformations between, say, a random polypeptide chain and a hemoglobin chain be very small indeed. There is no reason to assume that such is the case.

Whereas one cannot say that pseudo-homologies between polypeptide chains never arise in evolution, it seems very unlikely that such a process has occurred in the case of any particular polypeptide chain one is considering. It is thus likely that proteins as different in amino acid sequence as mammalian hemoglobins and myoglobins indeed derive from a common molecular ancestor, not because of the characters of amino acid sequence they have in common, but because of the common characters of primary structure, tertiary structure, and function taken

together. The ease with which variations of a given type of protein can be produced through duplication and mutation of a gene should be so much greater than the ease of convergent evolution from independent starting points that on the basis of this consideration alone any variants within a given type of tertiary structure and function seem to have a much greater chance to be phyletically related than unrelated.

ACKNOWLEDGMENT

The authors thank Mr. Jean-Pierre Salleï for his help in preparing the tables. This work was supported in part by N.I.H. award # GM 11272-01A1.

REFERENCES

1. ALLISON, A. C., CECIL, R., CHARLWOOD, P. A., GRATZER, W. B., JACOBS, S., AND SNOW, N. S., *Biochim. Biophys. Acta*, **42**, 43 (1960).
2. ALLISON, A. C., personal communication.
3. BABIN, D., AND SCHROEDER, W. A., personal communication.
4. BAGLIONI, C., *Biochim. Biophys. Acta*, **59**, 437 (1962).
5. BAGLIONI, C., *J. Biol. Chem.*, **237**, 69 (1962).
6. BAGLIONI, C., personal communication.
7. BAGLIONI, C., AND INGRAM, V. M., *Biochim. Biophys. Acta*, **48**, 253 (1961).
8. BAGLIONI, C., AND LEHMANN, H., *Nature*, **196**, 229 (1962).
9. BAGLIONI, C., AND WEATHERALL, D. J., *Biochim. Biophys. Acta*, **78**, 637 (1963).
10. BRAUNITZER, G., communicated at Conference on Hemoglobin, November 5-7, 1962, Arden House, Harriman, New York.
11. BRAUNITZER, G., AND HILSE, K., *Z. Physiol. Chem.*, **330**, 234 (1963).
12. BRAUNITZER, G., AND MATSUDA, G., *J. Biochem.*, **53**, 262 (1963).
13. CULLIS, A. F., MUIRHEAD, H., PERUTZ, M. F., AND ROSSMANN, M. G., *Proc. Roy. Soc.*, **A265**, 161 (1961).
14. DIAMOND, M., AND BRAUNITZER, G., *Nature*, **194**, 1287 (1962).
15. DOBZHANSKY, Th., personal communication (1964).
16. ECK, R. V., *Science*, **140**, 477 (1963).
17. FICQ, A., AND PAVAN, C., *Nature*, **180**, 983 (1957).
18. FOX, S. W., this volume.
19. FREESE, E., *Proc. Natl. Acad. Sci. U.S.*, **45**, 622 (1959).
20. GERALD, P. S., AND EFRON, M. L., *Proc. Natl. Acad. Sci. U.S.*, **47**, 1758 (1961).
21. HANADA, M., AND RUCKNAGEL, D. L., *Biochem. Biophys. Res. Commun.*, **11**, 229 (1963).
22. HILL, R. J., AND KRAUS, A. P., *Federation Proc.*, **22**, 597 (1963).
23. HILL, R. L., personal communication.
24. HILL, R. L., BUETTNER-JANUSCH, J., AND BUETTNER-JANUSCH, V., *Proc. Natl. Acad. Sci. U.S.*, **50**, 885 (1963).
25. HILL, R. L., SWENSON, R. T., AND SCHWARTZ, H. C., *J. Biol. Chem.*, **235**, 3182 (1960).
26. HOYER, B. H., MCCARTHY, B. J., AND BOLTON, E. T., *Science*, **140**, 1408 (1963).
27. HUEHNS, E. R., DANCE, N., BEAVEN, G. H., KEIL, D. V., HECHT, F., AND MOTULSKY, A. G., *Nature*, **201**, 1095 (1964).
28. HUNT, J. A., AND INGRAM, V. M., *Nature*, **184**, 640 (1959).

29. HUNT, J. A., AND INGRAM, V. M., *Biochim. Biophys. Acta*, **42**, 409 (1960).
30. HUNT, J. A., AND INGRAM, V. M., *Biochem. Biophys. Acta*, **49**, 520 (1961).
31. INGRAM, V. M., *Nature*, **180**, 326 (1957).
32. INGRAM, V. M., *Nature*, **189**, 704 (1961).
33. ITANO, H. A., *Advan. Protein Chem.*, **12**, 215 (1957).
34. ITANO, H. A., Symposium on Abnormal Hemoglobins, Ibadan, 1963. Blackwell, Oxford.
35. JONES, R. T., personal communication.
36. JONES, R. T., AND ZUCKERKANDL, E., unpublished results.
37. JUKES, T. H., *in* "The Origins of Prebiological Systems," (S.W. Fox, ed.), p. 407, Academic Press, New York, 1965.
38. KENDREW, J. C., *Brookhaven Symp. Biol.*, **15**, 216 (1962).
39. LANNI, F., *Proc. Natl. Acad. Sci. U.S.*, **47**, 261 (1961).
40. LOVE, W. E., *Federation Proc.*, **22**, 597 (1963).
40a. MARGOLIASH, E., *Brookhaven Symp. Biol.*, **15**, 266 (1962).
41. MARGOLIASH, E., AND SMITH, E. L., this volume.
42. MONOD, J., CHANGEUX, J. P., AND JACOB, F., *J. Mol. Biol.*, **6**, 306 (1963).
43. MUIRHEAD, H., AND PERUTZ, M. F., *Cold Spring Harbor Symp. Quant. Biol.*, **28**, 451 (1963).
44. MULLER, C. J., AND KINGMA, S., *Biochim. Biophys. Acta*, **50**, 595 (1961).
45. MURYAMA, M., *Federation Proc.*, **19**, 78 (1960).
46. NANCE, W. E., *Science*, **141**, 123 (1963).
47. NEEL, J. V., WELLS, I. C., AND ITANO, H. A., *J. Clin. Invest.*, **30**, 1120 (1951).
48. PATTEE, H. H., *Biophys. J.*, **1**, 683 (1960).
49. PAULING, L., AND ZUCKERKANDL, E., *Acta Chem. Scand.*, **17**, 9 (1963).
50. PERUTZ, M. F., "Proteins and Nucleic Acids," Elsevier, Amsterdam, 1962.
51. PIERCE, L. E., *New Engl. J. Med.*, **268**, 862 (1963).
52. RUDLOFF, V., communicated at the 6th Intern. Congr. Biochem., New York, 1964.
53. SATAKE, K., AND SASAWAKA, S., *J. Biochem. (Tokyo)*, **52**, 232 (1962).
54. SCHACHMAN, H. K., *Cold Spring Harbor Symp. Quant. Biol.*, **28**, 409 (1963).
55. SCHELLMAN, J. A., AND SCHELLMAN, Ch., *in* "The Proteins," (H. Neurath, ed.), Vol. II, p. 1, Academic Press, New York, 1964.
56. SCHNEIDER, R. G., HAGGARD, M. E., MCNUTT, C. W., JOHNSON, J. E., BOWMAN, B. H., AND BARNETT, D. R., *Science*, **143**, 697 (1964).
57. SCHROEDER, W. A., *Fortschr. Chem. Org. Naturstoffe*, **17**, 322 (1959).
58. SCHROEDER, W. A., *Ann. Rev. Biochem.*, **32**, 301 (1963).
59. SCHROEDER, W. A., SHELTON, J. R., SHELTON, J. B., CORMICK, J., AND HONES, R. T., *Biochemistry*, **2**, 992 (1963).
60. SIMPSON, G. G., *Science*, **146**, 1535 (1964).
61. SIMPSON, G. G., *Science*, **143**, 769 (1964).
62. SMITH, D. B., *Can. J. Biochem.*, **42**, 755 (1963).
63. SMITH, E. L., personal communication.
64. ŠORM, F., *Collection Czech. Chem. Commun.*, **27**, 994 (1962).
65. STEINHARDT, J., AND BEYCHOK, S., *in* "The Proteins" (H. Neurath, ed.), Vol. II, p. 140, Academic Press, New York, 1964.
66. SWENSON, R. T., HILL, R. L., LEHMANN, H., AND JIM, R. T. S., *J. Biol. Chem.*, **237**, 1517 (1962).
67. SWIFT, H., *in* "The Molecular Control of Cellular Activity" (J. M. Allen, ed.), p. 73, McGraw-Hill, New York, 1962.

68. VOGEL, H. J., BRYSON, V., AND LAMPEN, J. O., eds., "Informational Macromolecules," Academic Press, New York, 1963.
69. ZUCKERKANDL, E., *in* "Classification and Human Evolution" (S. L. Washburn, ed.), p. 243, Aldine Publishing Co., Chicago, 1963.
70. ZUCKERKANDL, E., *J. Mol. Biol.*, **8**, 128 (1964).
71. ZUCKERKANDL, E., *in* "Protides of Biological Fluids," Proceedings of the 12th Colloquium (H. Peeters, ed.), p. 102, Bruges, 1964.
72. ZUCKERKANDL, E., *Abstr. 6th Intern. Congr. Biochem.*, New York, Vol. III, p. 210 (1964).
73. ZUCKERKANDL, E., in preparation.
74. ZUCKERKANDL, E., AND PAULING, L., *in* "Horizons in Biochemistry," (M. Kasha and B. Pullman, eds.), p. 189, Academic Press, New York, 1962.
75. ZUCKERKANDL, E., AND PAULING, L., *in* "Problems of Evolutionary and Industrial Biochemistry," p. 54, Science Press, Academy of Sciences, USSR, 1964, and *J. Theoret. Biol.*, **8**, 357 (1965).
76. ZUCKERKANDL, E., AND SCHROEDER, W. A., *Nature*, **192**, 984 (1961).
77. ZUCKERKANDL, E., AND SCHROEDER, W. A., unpublished observations.