J.Mol.Evol.7,167-183 (1976)

Journal of Molecular Evolution

Evolutionary Processes and Evolutionary Noise at the Molecular Level

I. Functional Density in Proteins

EMILE ZUCKERKANDL*

Marine Biological Laboratory, Woods Hole, Mass. 02543, and Department of Biological Sciences, University of Delaware, Newark, Del. 19711, USA

Received November 25, 1974; January 15, 1975

Summary. The distinction between molecular sites that mainly carry out general functions and sites committed to specific functions is analyzed, notably in terms of different evolutionary variabilities. Functional density is defined as the proportion of sites involved in specific functions. Weighted functional density, by representing the relative variability at specific-function sites is to some extent a measure of the specificity of molecular interactions. The relationship between general- and specific-function sites on the one hand and the covarions of Fitch on the other is discussed. The functional "degeneracy" of amino acids is described as increasing the interdependence of general functions. It is predicted that proteins that do not possess general-function sites besides their specific-function sites tend to "freeze" their primary structure, according to an evolutionary process that is an autocatalytic function of the decrease in site variability. This limits the use of weighted functional density as an indicator of the overall degree of interaction specificity of a protein to values that are not close to unity.

Key words: Protein Evolution/Protein Functions/Functional Density/ Covarions/Functional Degeneracy

All evolutionarily effective amino acid substitutions do not have the same impact on protein evolution. Many accepted mutations probably lead only to minute functional variations in the protein. They are essentially evolutionary noise. Other fixations represent more important functional changes. To be able to measure the evolutionary significance of amino acid fixations, we shall consider different categories of functions as carried out by single protein molecules. This analysis will

*Directeur de Recherche at Centre National de la Recherche Scientifique, Paris.

furnish the basis for a second, connected paper, which will deal with a mechanism for the most frequent, although not evolutionarily most important, amino acid fixations.

EVOLUTIONARY VARIABILITY IN RELATION TO SITE FUNCTIONS

1. General and Specific Functions

We may call "functional property" or "function" any molecular property of a polypeptide chain that is discernable by natural selection. This definition implicates any property of the polypeptide that contributes to a certain equilibrium state in molecular ecology, a state that expresses the adaptation of the organism to its environment and of components of the organism to the organism itself.

A polypeptide selected by nature will at all times exert general as well as specific functions, and the first will be, so to speak, at the service of the second. A specific function is one that can be carried out by a certain amino acid residue at a certain molecular site with little or no leeway as to the residue and very little or none as to the site. A general function can be carried out by any among several residues at any among a certain number of sites. Several sites may share in one specific function. They must then act each in its particular way. On the other hand a number of sites always interact in general functions but they may do so in different and flexible combinations. The active site of an enzyme is an example of a set of residues carrying out a specific function. General functions are, for instance, solubility, charge density, isoelectric point, pK, mean polarity of residues.

Each general function can be represented by a single figure giving its overall value for the protein and there are in principle a very large number of possible molecular solutions for obtaining this value. Specific functions are expressed not only by algebraic but also by geometric parameters. Each specific function is characterized by several parameters, whereas each general function can be characterized by just one.

One single amino acid substitution will always have a relatively modest incremental effect (from negligible to significant) on the value of a general function of the molecule. In the case of specific functions, the effect of one single substitution is never negligible and is highly variable; it may impair the function completely.

All protein sites participate in general functions. Sites involved in general functions accomodate a number of different substituents when the sites are not also engaged in a specific function. When they are, their variability, though not necess-

arily abolished, is reduced. The dichotomy between general-function and specific-function sites obviously is not absolute. There may be no sites and no substitutions that are without any effect on specific functions of a protein. But at some sites (the most variable) and for some substituents these effects should be very small.

Even though specific-function sites will play a role with respect to general functions also, the specific functions should be dominant, as being the protein's primary business. Protein sites nearly exclusively concerned with general functions will have to make up for whatever effect the filling of the specific functions has on adaptation to general functions. Adaptation with respect to several general functions simultaneously, in addition to adaptation to the specific functions, no doubt is difficult unless a certain proportion and minimum absolute number of sites are nearly exclusively geared to general functions. Thus, as a rule, proteins must be large enough for mustering a sufficient number of sites nearly exclusively involved with general functions, so as to compensate for undesirable effects on general functions exerted by sites primarily connected with specific functions. This circumstance should help explain the selective value of the large size of protein molecules that appear to be needing only part of their structure for carrying out their specific functions (Schultze, 1964). (The large size of proteins appeared as an enigma in the past also because the specific functions of intermolecular interactions were not properly understood.)

Sets of sites engaged in different specific functions may be expected mostly not to overlap. In contrast, sets of sites primarily engaged in different general functions not only overlap, but should be largely identical. One and the same residue will have a bearing on several general functions. For instance, when an alanyl residue is replaced by a glutamyl residue, the result may concern at once solubility, isoelectric point, and structural stability (see paper II of this series).

2. Structure Functions of Residues

Is structural stability of the macromolecule to be listed under general functions of sites? One of the obvious assignments of many molecular sites and residues is to maintain the secondary, tertiary, and, if pertinent, quaternary structure of a protein. This is part of what may be called the structure function of sites and residues. Structure functions include a class of general as well as a class of specific interactions. They thus represent a special case by their ambivalence.

The implication of a residue in the overall ratio of polar to apolar amino acids, or its stabilizing or destabilizing action on a helical segment, etc., belong to the category of general site and residue functions. The generality lies in the fact that the amino acid may be replaced by others at the same site, or various replacements at various other sites may achieve the same effect. One amino acid substitution usually only has a limited incremental effect on the function, and the function itself can be represented by a single figure.

Beside their general structure function, in which in fact all residues share, an important proportion of sites and residues exert a specific structure function that may be termed *contact function*. It consists in making specific contact either with other parts of the same polypeptide chain (or poly nucleotide chain, for that matter) or with parts of other molecules, whether other polypeptide chains, polynucleotide chains, substrates, cofactors, prosthetic groups, allosteric ligands, etc.

To the notion of contact function a negative component can be added, namely the specific or nonspecific *avoidance* of intermolecular binding. The set of functional properties that includes the contact functions and this negative component may be referred to under the term of *relational functions*. Table 1 summarizes the different types of protein functions and their designations as used here.

Specificity of intra- and intermolecular interactions by no means precludes variability of primary structure. This is suggested by the many differences in sequence that are compatible with one and the same overall tertiary structure, as is well known; and is demonstrated by the evolutionary variability of residues at interchain contact sites in hemoglobins (Zuckerkandl & Vogel, 1972; Goodman et al., 1975).

Nevertheless, on the average, residues at sites involved predominantly in general functions are more variable than residues at sites that exercise a contact function, whether of intra- or intermolecular competence. There is an evolutionary slowdown due to contact function (Zuckerkandl & Vogel, 1972). In turn, sites engaged in a contact function are generally more variable than groups of residues forming together the active site of an enzyme.

There are thus three main categories of site variability:

1. The most variable sites are those involved, to a first approximation, with general functions only. They are at the surface of globular proteins and include general structure functions such as the ratio of polar to apolar amino acids. Sites at the interior of the protein molecules also carry out such general structure functions. However, at the same time, most of them exercise specific contact functions. Hence, they are not in this group of most variable sites.

Table 1
Types of functions exercized by polypeptide chains

General	Specific
Single physical parameters, with many different molecular "solutions" for any single value. All sites are involved	Complex sets of parameters with few or no interchangeable solutions. A particular subset of sites is involved

I. Indirectly linked to primary function of molecule

Structure Functions

General structure functions	Contact functions	
(polar/apolar amino acid ratio, helical stabilization/ destabilization, etc.)	intramolecular	intermolecular
		-with evolutionarily variable ligands
OR	OR	
		-with invariant ligands
Nonspecific relational functions (includes general structure functions plus negative component: avoidance of nonspecific intermolecular interactions such as aggregation)		Specific relational functions (includes contact functions plus a negative component: noncombination with certain molecules)

II. Directly linked to primary function of molecule

Chemical functions (active enzymatic sites).
Dynamic (actin) and static (silk fibroin) *mechanical functions, transport functions,* etc.

2. Next come indeed sites responsible for structural specificity in contact function. There probably are two subclasses from the point of view of variability:

a) The more variable subclass is concerned with specific interactions either with other polypeptide chains or with other parts of the same polypeptide chain. Since in either case the

interacting partner sequences are themselves variable during evolution, and mutual adaptation will follow variation, the variability of such specific contact sites is not negligible. Protein sites making contact with polynucleotide chains may well fall into this same group.

b) There is a subclass of sites that interact with invariable molecular partners, such as prosthetic groups and cofactors. Even in this case some variability of the polypeptide chain is retained in most cases. There should indeed be more than one way to specifically interact with any molecule, - in fact there may be many ways as soon as the ligand is not small. Whether more than one way will be adopted during evolution should depend on whether the premium is on binding per se, irrespective of how exactly it is performed (the combination between antibody and antigen should be an example here), or whether function requires that a certain narrowly defined kind of interaction, or set of interactions, should take place with the invariable molecule. (This should apply to the interaction between some histones and certain invariable features of DNA. An obligation for the ligand to interact specifically with *more than one* structural state of the partner molecule, as may well be the case of histones, should greatly contribute to restricting the variability of the polypeptide chain.)

3. At the extreme pole of invariability are the active sites of enzymes. Here some residues must have a certain identity and be in a certain mutual spatial relationship for reasons of *chemical function.* These residues presumably are only a few per chemically active polypeptide chain (one in the case of the globins, hemoglobin and myoglobin: the "proximal histidine"). Many polypeptide chains have no such active sites. Other primary functions of proteins (Table 1) include dynamic or static mechanical functions, transport functions, and others. How invariable these functions render the residues that are directly implicated is unknown.

3. Functional Density

Maximal evolutionary variability is represented by the variability of sites of type (1) that are involved nearly exclusively in general functions. The extent to which the rate of evolution of a given type of protein departs from this maximal rate is no doubt linked to what may be called *functional density* of a polypeptide chain. This is defined by the proportion of sites concerned with specific functions. This proportion in turn is determined by the number of different specific functions carried out per unit length of a polypeptide chain and by the number of sites involved in each of these functions. Since, as stated, sites implicated in specific functions may

in general not overlap (though they sometimes do, see example below), these numbers are mostly additive. General site functions cannot be part of a definition of functional density, since, as mentioned, general functions are associated with one hundred per cent of the sites in all proteins.

On the basis of the data of Goodman et al. (1975) on evolutionary rates at different types of sites in vertebrate tetrahemic hemoglobins, functional density in the case of the β-chain

$$FD = \frac{N_s}{n} = \frac{76}{146} = 0.52$$

with N_s, number of sites committed to specific functions[1] and n, total number of sites.

Since the evolutionary variation in primary structure of a protein is slowed down by its specific functions, there should be a simple relation between evolutionary rates and functional densities of different kinds of proteins. Departures from this relation might be supposed to be due to a variation in the proportions of residues concerned with the two types of specific structure functions, in regard to variable (macromolecular) and invariable (smaller) partner molecules, and with chemical function. Since the proportion of residues involved in this last type of function is low, the question arises as to whether the proportionality between functional density and rate of evolution of the informational macromolecule might in practice be affected only by changes in the ratio between the number of sites concerned with the two categories of specific contact functions, namely interaction with respect to evolutionarily variable and invariable ligands. According to the analysis of Goodman et al. (1975), contact sites for variable ligands number approximately 49 in Gnathostome hemoglobin chains (there is some slight ambiguity; see also footnote below) and contact sites for invariable ligands number 25 (haem contacts plus 2,3-diphosphoglycerate binding sites). Yet, as Dr. Goodman points out to me, these numbers do not correspond to the faster and slower evolving contact sites, respectively. Some of the contact sites for variable ligands, namely the $\alpha_1\beta_2$ (and $\alpha_2\beta_1$) subunit contact sites are in reality highly invariant. This is so, says Goodman, because these sites modulate oxygen affinity and thus exert a strong influence on the chemical function of the molecule. On account of functional overlaps, the distinction between contact sites for variable and invariable ligands thus cannot be used to advantage in the present connection without a very thorough

[1] Dr. Morris Goodman points out to me that the figure of 76 sites involved in specific functions is a minimum, since specific binding sites for haptoglobin are not taken into account.

knowledge of functional relationships throughout the molecule. Moreover the number of sites actually engaged in chemical function does not seem negligible.

Functionality implies a certain measure of invariance and this measure is insured by natural selection. With respect to general function sites, the contribution to variability per residue may be anticipated to be approximately constant for all proteins and all times after the appearance of the contemporary types of cells, since the kinds, limits, and tolerance ranges of general-functions are unlikely to have changed. This likens general function sites to Fitch & Markowitz's (1970) covarions (Fitch, 1973; see further discussion below). For specific function sites the degree of invariance changes not only, as stated, with the degree of invariance of the interacting partner molecules, but also with the degree of specificity of the interaction. By this is meant the latitude left to the molecule to adopt either any of the many possible solutions to the problem of achieving some intermolecular binding, or only a restricted set out of these many. The degree of this restriction represents the degree of interaction specificity. The variability of residues engaged in specific functions is, in part, a measure of this specificity. For weighting functional densities, the variability or invariability of ligands then does not furnish the most adequate parameter. It seems preferable to use the distribution of the site variabilities in the protein under consideration (only sites engaged in specific functions being of course considered). If, taking into account evolutionary rates at specific-function sites it turned out that the rates of evolution of the remainder of the proteins are equal, the expectation that general-function sites evolve at identical average rates in different proteins would be verified.

A weighted functional density, WFD, may thus be a functional density weighted by the mean variability of sites engaged in specific functions. We may assume that sites engaged in a single specific function, barring overlaps between specific functions with respect to sites, are characterized by a limited variance of their variability and represent one variability set. A precise molecular analysis would allow one to define the different variability sets, one per specific function (e.g. in hemoglobin the binding of heme and O_2, of partner chains, of the proton that is instrumental in the Bohr effect, of 2,3-diphosphoglycerate, of CO_2, of haptoglobin would represent six specific functions) plus supplementary variability sets for sites with overlapping specific functions. If sites with overlapping specific functions are set apart, a distinction with respect to evolutionary rates between sites binding variable and invariable ligands may be reestablished.

A weighted functional density may be represented by

$$WFD = \sum_i \frac{N_{s_i}}{n} \frac{\bar{\nu}_g - \nu_{s_i}}{\bar{\nu}_g}$$

with N_{s_i}, number of specific-function-sites of variability i, n, total number of molecular sites, $\bar{\nu}_g$, mean rate of evolutionarily effective amino acid replacement (variability) at general-function sites, and ν_{s_i}, variability of magnitude i, characteristic of a variability set of sites. Thus, the lower the variability at specific-function sites with respect to the variability at general-function sites, the higher is the weighted functional density. Maximal functional density, if all sites were involved in specific functions and all sites were totally invariant would be 1 according to this expression. Weighted functional density is always smaller than functional density. It would be equal to functional density if all sites involved in specific function were totally invariant. This is obviously but one possible formulation of a weighted functional density.

In most cases an evaluation of the mean evolutionary variability of all specific-function sites may be more easily obtainable than the variabilities for different residue sets. On the basis of the postulate of a universal rate for the evolution of general-function sites, an approximation to the mean variability of specific function sites may be obtained, provided a count of specific function sites (and therefore general function sites) is available:

$$\bar{\nu}_s = \frac{n\bar{\nu}_t - N_g\bar{\nu}_g}{N_s}$$

with $\bar{\nu}_s$, mean variability at specific-function sites, $\bar{\nu}_t$, mean variability at all sites (evolutionary rate of the polypeptide chain), N_g, the number of general function sites and the other symbols as above.

The value of the weighted functional density is then obtained by

$$WFD' = \frac{N_s}{n} \frac{\bar{\nu}_g - \bar{\nu}_s}{\bar{\nu}_g}$$

On the basis of the data of Goodman et al. (1975; their Table 10), taking their rate value for "remaining exterior positions" (0.20 nucleotide replacements per position per hundred millions of years) as an approximation to the mean variability at general-function sites, WFD = 0.32 and WFD' = 0.34. For histone IV (cf. Dayhoff, 1972), the weighted functional density should be rather close to 1.0.

Weighted functional density is to some extent a measure of the overall degree of interaction specificity of a protein[2]. According to the definitions and relationships used here, histone IV would thus be almost three times as specific as hemoglobin chains from higher vertebrates. However, very high values of WFD no longer can safely be considered to measure interaction specificity, as will be shown below.

Could not the ratio of general-function site variability to variability at all sites be taken as a simple measure of interaction specificity? This is not so, because mean variability can be lowered to the same extent either on account of many sites with slightly reduced variability (high functional density, low specificity), or fewer sites with greatly reduced variability (lower functional density, higher specificity). To what extent the latter situation actually occurs is not established. But the answer is to be found out and not to be given implicitly. The degree of invariance of a set of sites engaged in one given specific function (treating sites committed to several specific functions at once as separate variability classes) measures the specificity of interaction in the case of each specific function, and the mean of these specificities, taking into account the fraction of the molecule to which they relate, is the mean interaction specificity of the molecule. WFD may thus indeed be considered a measure of molecular specificity.

General function sites are taken here as equivalent, and this should be legitimate to a first approximation. It is unlikely that they really are and that they thus evolve at exactly identical rates. Their position on the molecule should make some difference on account of a varying fractional par-

---

[2] *Enzymatic* specificity is usually defined as the range of compounds which are substrates for a given enzyme (e.g. Citri & Pollock, 1966). *Interaction* specificity is here defined as the range of molecular solutions for binding a given substrate or other ligand. The first definition can be conceived as involving sometimes an extension of the second: the larger the range of different substrates that are bound by an enzyme, the larger is the number of different constellations of protein binding sites brought into action. Therefore a lower enzymatic specificity will imply a lower interaction specificity. Weighted functional density, as expressed above, is an imperfect measure of interaction specificity, in that a part of the evolutionary variability of a protein will be a direct function of the evolutionary variability of the ligand, when the ligand is another protein. Thus, at equal interaction specificity, the evolutionary variability of the specific interaction sites of a protein may be more or less high. A more fully satisfactory measure of interaction specificity would demand the introduction of the evolutionary variability of the ligand as a further weighting factor.

ticipation in specific functions of the sites. It would seem that sites in certain external loops are the fastest evolving of any polypeptide sites and get close to the figure of 1 amino acid replacement per hundred million years (Barnard et al., 1972; see also Corbin & Uzzell, 1970). Even in those extreme cases only about one fifth or less of the mutations as they occur seem to be accepted (Zuckerkandl, 1975). At other general-function sites the rate of evolution appears to be smaller than it is, for instance, in the ribonuclease loop. The value of 0.2 substitutions per amino acid site per hundred million years, taken from the paper by Goodman et al. (1975) and used above in an illustrative calculation, should be an underestimate on two counts. It is an "unaugmented" value and should be about double according to the augmentation procedure of Goodman and Moore (M. Goodman, personal communication). On the other hand, the set of sites designated by the authors as "remaining exterior positions" in hemoglobin chains should still include some specific-function sites.

It would thus seem that the rates at fixations of general-function sites vary by a factor of 2. Yet it also seems that the highest rates at general-function sites are rather exceptional, so that the mean rate to be used in a general computation of weighted functional densities in proteins should be somewhat closer to the lower than to the higher extreme. An intermediate value is more representative of general-function sites as they *are*. The higher extreme is more representative of an abstract "pure" general-function site. Indeed, since the committment to general functions is unavoidable for any site within a molecule, no sites can "evolve" faster than general function sites. Even the fastest "evolving" sites have been shown to exclude a certain proportion of the mutations that must occur there and therefore to change below mutation rate (Zuckerkandl, 1975). Thus no *site* can of course be neutral, and among the substituents most frequently accepted at a highly variable site none may be *generally* nearly neutral, except perhaps, though this is unlikely, the most frequently accepted one, since the other substituents (if not also the most frequently accepted one) apparently are more or less often eliminated as they turn up by mutation.

Substitutions affect most general functions by small increments, as stated, and the change will be favorable or unfavorable. The chances of a favorable effect are relatively high. On the other hand, most substitutions that affect a specific function will change it in the direction of impairment or elimination. The measure of correctness of the assertion according to which most mutations are deleterious will thus in part depend on the value of the weighted functional density. The smaller the weighted functional density, the greater the frequency of occurrence of nondeleterious mutations and the smaller the genetic load.

It should be stressed that, from the point of view of selection, a single change in net charge may be on the average more significant than an incremental change in other general properties of the molecule such as mean polarity. Thus substitutions that involve a change in charge may be among those that are least often neutral. Demonstrations based exclusively on electrophoretically detected mutants are not sufficient for establishing the true contribution of neutral behaviour of alleles in evolution. This argument is not effectively answered by pointing out that charge changes may be fractional, as shown by the fact that the absolute numerical value of the charge often is not an integer. It has been the experience of workers in the field that relative distances between bands obtained by electrophoresis of proteins nevertheless usually correspond to integral or nearly integral units of charge, as was pointed out by Walter Fitch (personal communication).

SITE FUNCTIONS AND COVARIONS

The present analysis of functionally distinct types of sites in proteins was foreshadowed a long time ago (Zuckerkandl, 1963). The basis for Fitch's concept of concomitantly variable codons, "covarions", was indicated in the same paper. It was pointed out that the range of permissible fixations at one site must change as a function of fixations at other sites, and that the consequence of one mutation often depends on the occurrence of others. A given amino acid substitution may be incompatible with function, it was pointed out, but this incompatibility may cease if a second change occurs. Conversely, a favorable substitution may cease to be favorable when a second substitution occurs somewhere else in the molecule.

These statements imply that the set of protein sites that can accept substitutions should vary as fixations take place. Fitch & Markowitz (1970) have proposed and skillfully demonstrated that this is so.

A critical variation in the variability of amino acid sites cannot be expected to occur at sites that are and remain general function sites. General function sites, for the time they so remain, should be permanent covarions. Yet Fitch (1971) found a high turnover rate (0.75) in covarions of cytochrome c. This means that a fixation in any covarion affects most of the others to a considerable degree.

Should covarions then represent predominantly the more variable specific-function sites? It is plausible that some of these very frequently become variable as others become invariable. For instance, as a more critical role in a contact function is taken over by one site, this site should become less variable and some other site, correlatively, more so.

However, covarions cannot predominantly relate to specific function sites, since general-function sites obviously must be covarions. Moreover, the rate of evolution indicated by Fitch (1973) for covarions in several proteins, including cytochrome c and hemoglobin chains, is in excellent agreement, as far as the precision of present data allows one to judge, with the mean rate of evolution of general-function sites as discussed above.

An alternative is to propose that the high turnover rate of covarions discovered by Fitch points to a constant turnover between general-function sites and specific-function sites. For instance, during the evolution of vertebrate hemoglobins from the monomeric to the tetrameric state, many of the new specific contact sites may have evolved out of general-function sites. Yet such events do not occur regularly with successive amino acid substitutions, but rather rarely, if at all, during relatively recent evolution. When they do, they represent important evolutionary stages. This explanation for the high turnover rate of covarions is therefore to be dismissed also.

As a way out of the difficulty, one may predict that the turnover rate of covarions in hemoglobin chains should be much smaller than the rate indicated for cytochrome c. This is likely to be the case (Fitch, personal communication), in view of the much larger number of covarions in hemoglobin chains (Fitch, 1972a,b). As to cytochrome c itself, the high turnover rate of covarions might mean that there are hardly any general-function sites, and that general functions and specific functions have to be satisfied simultaneously by practically all residues. The "normal" figure for the rate of fixations in cytochrome c covarions (Fitch, 1973), which is moreover close to the rate of fixations at general-function sites in hemoglobins, contradicts this hypothesis. There are few general-function sites in cytochrome c, but apparently there are a few. Perhaps turnover rate of covarions was overestimated. That in slowly "evolving" proteins such as cytochrome c general functions must be largely satisfied by specific-function sites is however obvious.

General-function sites *have* to be covarions, and in view of the evolutionary rarity of a switch in status of general-function sites pointed out above, a switch that would often be measurable by a change in functional density, an important fraction of the covarions should be stable over rather long evolutionary periods of time.

The number of general-function sites in hemoglobin chains from higher vertebrates was given above as approximately 70, a figure that was considered an overestimate. Even if we supposed that the real number is smaller, it should still be somewhat higher than the number of covarions, which is given

as equal to 50 and 39 for the α- and β-hemoglobin chains respectively (Fitch, 1972a,b). Since there should not be a larger number of general-function sites than there are covarions, the number of general-function sites, as indicated in this paper, has been grossly overestimated, or the number of covarions is underestimated. It is unexpected to find that there are more covarions in the α-chains than in the β-chains, since α-chains evolve more slowly than β-chains (Derancourt et al., 1967; Langley & Fitch, 1973). As the analysis given above indicates, a slower evolutionary rate should imply a smaller proportion of general-function sites and therefore a smaller number of covarions. The difference in number of covarions in the two chains suggests that sites other than general-function sites are indeed included in the set of covarions, and to a larger extent in the case of the α-chain than in that of the β-chain. Yet, on the basis of such an interpretation, the number of covarions available for general-function sites in hemoglobin chains shrinks even more, and the gap between number of apparent covarions and apparent general-function sites in hemoglobins becomes even wider. Further analysis of this situation is clearly required.

FUNCTIONAL "DEGENERACY" OF THE AMINO ACIDS

By selecting the particular set of amino acids that are coded for, evolution insured that in many cases one and the same amino acid would bear significantly on several general functions. Coded amino acids thus mostly display functional "degeneracy", in the sense of functional multiplicity of individual side chains, and at the same time are functionally overlapping, in that any pair of amino acids may have some of these functions in common (Zuckerkandl & Pauling, 1965). In other words, the same chemical or stereochemical function, such as introducing into the protein a significant increment in polarity, charge, or bulk, can be carried out by several distinct amino acids. A single amino acid, on the other hand, may combine in different ways several such functions and thus be, in different ways, functionally composite. Subgroups of coded amino acids exist that are very similar in one respect and differ in others. To achieve such a condition, the number, and number of types, of different coded amino acids had to be significant. Contrary selective forces, favoring another degeneracy, namely that of the code, probably limited these numbers during early evolution. The final state of the code presumably was a compromise between the trends toward two opposite favored kinds of "degeneracy".

Early evolution of the coded set of amino acids thus probably increased the possibilities of placing at a given amino acid site a residue that satisfied several general functions simultaneously. In combining different functional qualities, the structure of coded amino acids allows proteins to use fewer sites for approaching an overall optimal state in relation to general functions. At the same time *it renders the general functions highly interdependent*. Many amino acid substitutions have a chance of changing the overall state of the molecule with respect to several general functions simultaneously.

Some replacements may, on the other hand, affect essentially one general function at a time. For example, by substituting glutamine for glutamic acid, the charge is changed, whereas the contribution of apolar atomic groups is maintained nearly constant. Conversely, by substituting aspartic for glutamic acid, the charge is kept constant, while the contribution of apolar groups is slightly reduced. A reduction in hydrophobicity, as brought about by a single substitution of glutamic by aspartic acid, belongs to the category of changes in general functional properties of the protein molecule that, so we shall assume, cannot, as a rule, be sensed by natural selection, except if a certain directional accumulation of changes in the same property has taken place. This will be of pertinence in relation to the topic of the second of these two related papers.

AN AUTOCATALYTIC FREEZE OF PRIMARY PROTEIN STRUCTURE

It was stated that in slowly evolving proteins, of high functional density, the general functions of the molecule must largely be satisfied by the specific-function sites themselves.

This should however be rather difficult to achieve, since the specific functions require certain constellations of residues or types of residues whose role with respect to general functions must in part be accidental; in part only, since on account of the functional "degeneracy" of amino acids referred to above, specific-function sites do have some possibility to take into account the "needs" for general functions as well.

Nevertheless, there may not be many solutions to the problem that satisfy the complete range of both specific and general functions, when all or nearly all sites are directly involved in specific functions, and especially when these specific functions imply strongly reduced variability. The overall variability of the molecule may then be reduced spectacularly even much beyond the degree that the specific functions per se would imply. In other words, a high percentage of sites with reduced variability should in itself contribute

to the invariability of sites and induce a general freeze of the primary structure of the protein. Thus, beyond a certain point, an increase in functional density, and especially in weighted functional density, will tend to rapidly bring the value of weighted functional density up to nearly 1.0. The value will remain somewhat below 1.0 if, starting out from an established solution for satisfying all the general functions at the same time as the specific ones, there still exist one or a few different solutions that can be reached.

Histone IV may have undergone precisely this freezing process at some point of its past history (perhaps not long after the beginning of the evolution of the eukaryote cell). At that moment, either new specific functions were added to those already carried out by the molecule, or it lost a certain number of general-function sites. This would have meant either a reduction in molecular size, or a fusion between two genes under elimination of a section which, in terms of polypeptide sequence, filled mainly general functions.

According to this concept, as the invariability of a protein increases, it does so autocatalytically. The "rate" of the reaction will indeed, like in an autocatalytic process, be proportional to the quantity of its product. The product, here, is the proportion of specific-function sites, after the mobilization of a further fraction of the general-function sites for specific functions or a loss of general-function sites.

This interpretation accounts for the contrast in variability between a protein as invariant as histone IV and most other proteins. Proteins that are nearly invariant might quite generally be so beyond the requirement of invariability of the specific functions themselves.

Thus the decisive reason why a protein is almost totally invariant is not, as has been sometimes suggested, that its function is particularly "fundamental" and "central" with respect to the cell machinery; it is only in part because its specific functions are indeed highly specific and because it interacts with invariable ligands; it also is, we may presume, because for some structural and functional reason it cannot afford to possess a significant contingent of general-function sites and yet, like all proteins, must satisfy general functions. This situation would not be expected to arise in large, globular polypeptide chains, in which the committment of all sites to specific functions is unlikely to be "technically" feasable.

Plainly, in the case of a protein consisting entirely of specific-function sites, the theoretical optimal adaptation of the molecule with respect to each general function and a corresponding abstractly conceived optimum for the global adaptation of the molecule might be at quite a distance from

the *maximal* overall adaptation that is reached as the structure freezes. Yet, equally plainly, this maximal adaptation is functionally *sufficient*. This touches upon matters to be considered in the following paper.

Acknowledgements. I am indebted to Drs. M. Goodman and W.A. Fitch for their comments and help.

This work was supported by a grant from the American Philosophical Society.

REFERENCES

Barnard, E.A., Cohen, M.S., Gold, M.H., Kim, J.-K. (1972). Nature 240, 395

Citri, N., Pollock, M.R. (1966). Advan.Enzymol. 28, 237

Corbin, K.W., Uzzell, T. (1970). Am.Nat. 104, 37

Dayhoff, M.O. (1972). Atlas of protein sequence and structure, Vol. 5. Washington, D.C.: National Biomedical Research Foundation

Derancourt, J., Lebor, A.S., Zuckerkandl, E. (1967). Bull.Soc.Chim.Biol. 49, 577

Fitch, W.M. (1971). J.Mol.Evol. 1, 84

Fitch, W.M. (1972a). Brookhaven Symp.Biol. 23, 186

Fitch, W.M. (1972b). Haematologie und Bluttransfusion 10, 199

Fitch, W.M. (1973). Ann.Rev.Genet. 7, 343

Fitch, W.M., Markowitz, E. (1970). Biochem.Genet. 4, 579

Goodman, M., Moore, G.W., Matsuda, G. (1975). Nature 253, 603

Langley, C.H., Fitch, W.M. (1973). The constancy of evolution: a statistical analysis of the α and β hemoglobins, cytochrome c, and fibrinopeptide A. In: Genetic structure of populations, N.E. Morton, ed., p. 246. Honolulu: University Press of Hawaii

Schultze, H.E. (1965). Biologically active protein fragments. In: Protides of the biological fluids, H. Peeters, ed., p. 1. Amsterdam: Elsevier

Zuckerkandl, E. (1963). Perspectives in molecular anthropology. In: Classification and human evolution, S.L. Washburn, ed., p. 243. Chicago: Aldine Publishing Co.

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

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

Zuckerkandl, E., Vogel, H. (1972). The evolution of polarity relations in globins. In: Darwinian, neo-Darwinian, and non-Darwinian evolution, Proc.6th Berkeley Symp.Mathematical Statistics and Probability, Vol. 5, J. Neyman, ed., p. 155. Berkeley: University of California Press