

MUTATION-ABSORPTION MODEL OF THE ENZYME

■ MICHAEL CONRAD

Department of Computer and Communication Sciences,
University of Michigan,
Ann Arbor, Michigan 48109, U.S.A.

Gradual changes in function of proteins in response to single changes in primary structure are often observed to occur and are a necessary condition for evolution by variation and natural selection at the protein level. A probabilistic (entropy theory) analysis of the effect of changes in primary structure on three-dimensional shape and function shows that such gradualism is based on the presence of a control system in the molecule involving a definite general form of structure-function degeneracy. The assumptions of the analysis are that primary structure determines tertiary structure (or a thermal distribution of tertiary configurations and allosteric forms), tertiary structure determines function (characterized by rate and other parameters), and that certain features of tertiary structure may be specialized for particular functions. The main conclusion is that embodied in the molecule is a subsystem which serves as a buffer, absorbing mutation or other forms of genetic variation and expressing these as graceful variations in features of the shape critical for function. This buffer system may be realized by numerical redundancy of amino acids or other mechanisms which increase the redundancy of weak interactions responsible for folding, utilization of amino acids having a greater number of analogs with redundant features, or local and global structural formats which allow for more effective utilization of redundancy. The mutation-absorption model has implications for the interpretation of structure-function relations in biology, the topology of the adaptive landscape, the interpretation of isoenzymes and allozymes, the relationship between selection and neutralism in evolution, and the relation between the complexity of and energy required by biological systems and the effectiveness of evolutionary optimization.

1. Introduction. Central to the study of enzymes and proteins generally is the idea that function is implicit in tertiary structure and tertiary structure is implicit in primary structure (or amino acid sequence). For the evolution of proteins, however, an additional consideration is of fundamental importance. This is that shape and function must in many cases change only gradually with single changes in the primary sequence. The argument is that if this were not the case each protein would be isolated atop some

adaptive peak, with no easily traversable pathways for reaching other adaptive peaks (cf. Maynard-Smith, 1970). The simple probabilistic consideration underlying this argument is that the probability of jumping from one adaptive peak to another in a single step depends on the product of probabilities of all the independent genetic events required for this jump; whereas the probability of moving from peak to peak in a series of single steps is in essence additive (Conrad, 1972a).

In this paper an attempt will be made to develop a conceptual model of structure-function relations in the enzyme based on the above consideration, viz. on the consideration that it is on a gradualistic relation between function change and primary structure change that all protein evolution depends and therefore that this relation must be the decisive historical influence on macromolecular structure-function relations. From this point of view there must be added to the list of functional properties of the enzyme the function of contributing to amenability to the evolution process itself. This evolutionary function may be described by studying the effect of primary on tertiary and hence on functional variability, the object being to determine the conditions under which mutational events at the primary level are most likely to appear as slight changes in some feature of shape critical for function and therefore in terms of slight changes in function. The model which emerges may be called a mutation-absorption model since its essential feature is that implicit in the protein structure is a buffer system which absorbs mutations and expresses them gracefully in terms of slight changes in those features of three dimensional shape critical for the conventional functions, e.g. rate processes, binding, mobility, cooperativity, control, structure formation. According to the mutation-absorption paradigm the structure of the enzyme cannot be interpreted solely in the framework of these conventional functions.

2. Determinate Relationships between Primary and Tertiary Properties. Among the various aspects of enzyme structure and function certain quite definite relationships are generally believed to obtain:

(i) *Folding relation.* Primary structure determines tertiary structure (three dimensional shape), or more precisely a (thermal) distribution of possible tertiary structures in a given milieu and (in some cases) under given initial conditions.

(ii) *Lock-key relation.* Tertiary structure determines function (characterized by various rate constants, binding constants, etc.).

(iii) *Function-environment relation (allosteric property).* Enzymes are capable of undergoing shape (and therefore function) changes in response to the substrate or to control chemicals in the environment.

(iv) *Functional specialization.* Certain features of the tertiary structures

are generally associated with particular aspects of function. This is sometimes expressed by describing the enzyme in terms of recognition sites, control sites, active sites, and binding sites. In some cases specialized features of the molecule have been shown to be particularly involved with or critical for these functions, though this of course does not mean that these features would have any functional significance in isolation from the molecule (Rosen, 1973a).

Undoubtedly statements (i)–(iv) are subject to future refinement and possibly qualification by important exceptions. However, it is probably fair to say that they are the commonly made generalizations about enzymes (cf. Perutz, 1962) and that their usefulness is established at least at a first level of approximation. It should be noted, however, that the significance of shape for function derives from an analogy (the lock–key analogy of Ehrlich) and is therefore in some degrees metaphorical. The functions of the enzyme undoubtedly depend to a considerable degree on its dynamical properties. From a strictly physical point of view a more accurate statement would be that rate constants and other specific functional properties are implicit in certain quantum numbers of the entire system, which in turn are implicit in primary structure in the same sense that tertiary structure is (i.e. in the sense that enzymes with the same primary structure would have the same set of possible quantum numbers). The condition should be added that these implicit dependencies are in general conditioned on the environment. For the sake of definiteness it may be supposed that the latter can be specified in terms of classical physical chemical variables, e.g. mole numbers and macroscopic variables (such as temperature, pressure). The adequacy of such a classical description may certainly be questioned (especially *in vivo*), but the in-principle existence of a set of possible environments is a safe assumption and is all that is really necessary for the present purposes.

An analogy can be made between primary structure and genotype (since the gene codes form primary structure) and tertiary structure and phenotype of the enzyme, a point made in the context of the origin of life by H. H. Pattee. From the present point of view the phenotype may be regarded as consisting, for the purposes of practical description, of distinguishable structural and behavioral properties. Hence tertiary structure refers to the structural aspect of enzyme phenotype and rate constants to the functional (or behavioral capability) aspect. Undoubtedly at a deeper level of analysis, structure and behavior would have to be viewed as abstractions approximating a more complete dynamical description. However, in practice it is possible to specify structural properties and rate constants independently and the distinction can be taken as valid at the level of analysis which gives rise to statements (i)–(iv).

3. Statistical Relations between Primary and Tertiary Properties

A. *Measures of primary and tertiary variability.* To develop statistical relations between primary and tertiary properties (and also to formally express the determinate relations discussed in the previous section) it is convenient to adopt the following notational conventions:

(i) The ensemble of primary sequence encoded by a given gene locus is denoted by G and members of the ensemble by G_1, \dots, G_n (each member is encoded by some allele of the gene).

(ii) The ensemble of tertiary structures encoded by the locus is denoted by T and members of the ensemble by T_1, \dots, T_n .

(iii) The ensemble of functional properties of proteins coded by the locus is denoted by K and members of this ensemble by K_1, \dots, K_n . Each K_i is itself a tuple of parameters, i.e. $K_i = (k_{i1}, \dots, k_{in})$, where the k_{ij} are rate constants, binding constants, etc.

(iv) The ensemble of possible environments (milieus) is denoted by E , and the particular environments by E_1, \dots, E_p .

Note that (ii) and (iii) correspond to structural and functional aspects of the phenotype. Thus the assumption of separable structural and functional aspects is tantamount to the assumption that the phenotype (denoted by P_i) is equivalent to some pair (T_i, K_i) , or alternatively that the ensemble P of possible phenotypes is given by the Cartesian product $T \times K$.

As genotypic, phenotypic, or environmental variability increases, the uncertainty of the primary sequence, the phenotypic state, or the environmental state increases. Thus a good measure of variability is the entropy measure (cf. Woodward, 1953).

$$H(X) = - \sum_{i=1}^n p(X_i) \log p(X_i) \quad (1)$$

where $X = G, T, K, P$, or E and $p(X_i)$ is the probability of X_i . Conditional entropies $H(X | Y)$ or $H(X | Y, Z)$ can also be defined, e.g.

$$H(X | Y) = - \sum_{i=1}^n \sum_{j=1}^m p(X_i, Y_j) \log p(X_i | Y_j) \quad (2)$$

where $Y = G, T, K, P$, or E and $p(X_i | Y_j)$ is the probability of X_i given Y_j . The conditional entropy is a measure of the variability of one type of property (e.g. tertiary structure) given complete information about another type (e.g. primary structure). Clearly the summations in definitions (1) and

(2) subsume a finite, discrete set of states. This (I believe mild) assumption avoids technical difficulties associated with a choice of measure (i.e. avoids the necessity of a formalism which handles both the discrete and continuous cases at once) and therefore eliminates some complications which have no material bearing on the conclusion. (An argument might also be made that it is in fact the realistic assumption from the more strictly physical point of view indicated in the previous section. The set of possible primary structures is of course in principle finite, though perhaps extremely large).

In the case of the gene, $p(G_i)$ may be taken as the relative frequency of allele (or allozyme) G_i in an infinite (or very large) population of organisms carrying this gene and altogether there are n such alleles. The variability in the gene may come from any genetic process, e.g. point mutation, duplication, or intragenic crossing over. In the case of T and K the probabilities may be taken, in the first instance, as the relative frequencies of the tertiary structures of enzymes coded by this gene locus, and in the second instance, by the relative frequencies of rate constants associated with these enzymes. In the case of E , the $p(E_i)$ are the relative frequencies of the occurrence of the various possible milieus and, for the present purposes, there is no necessity to distinguish the temporally determinate and temporally indeterminate changes in the environment (but see Conrad, 1972b, 1977a).

The entropy measure can be related to the possible extent of similarity among the members of an ensemble, hence the possible degree of gradualism of function change. The relation is: *a small value of $H(K)$ is a necessary condition for a high degree of gradualism in function change.* The argument is that in order for $H(K)$ to be large the number, n , of elements must increase or the $p(K_i)$ must become more equal. In the former case the number of possible distinguishable tuples of rate constants, (k_{i1}, \dots, k_{ih}) , must increase and in the latter case (which can only account for a delimited entropy increase) the number of tuples whose frequency of occurrence is significant increases. Now define the range of values of K as the maximum distance between any two tuples in the ensemble (where the distance between any two tuples, $d(K_i, K_j)$, can be taken as $\max\{|k_{i1} - k_{j1}|, \dots, |k_{ih} - k_{jh}|\}$). Since it can reasonably be assumed that the number of rate constants required to specify K (i.e. the dimensionality of the tuple) does not increase, it follows that the range of either possible or significantly occurring values of K can only be small if $H(K)$ is small, with the minimum possible range growing larger as $H(K)$ grows larger. However a small value of $H(K)$ is a necessary condition for a high degree of gradualism.

Note that if the ensemble of possible tuples is small, $H(K)$ will be small

whether or not the range of values is small, whereas if $H(K)$ is large the range of values cannot be small. Thus the above condition is necessary but not sufficient for gradualism. However, from the standpoint of the argument to be developed, a necessary condition is more valuable since any conclusion derived from it will apply to all proteins.

B. *Restrictions on conditional terms.* The determinate relations between primary and tertiary properties (relations (i)–(iv), Section 2) can now be expressed in terms of the following restrictions on conditional terms:

(i) *Unreliability terms.* According to the folding relation primary structure determines tertiary structure, therefore G necessarily provides information about T . This means that

$$H(T) - H(T|G) > 0, \quad (3a)$$

where $H(T|G)$ is the shape uncertainty arising from allosteric or other environmentally induced transitions (since the state of the environment is unspecified) and also from thermal configurational changes. If the environment is specified it is possible to write

$$H(T|G, E) \leq H(T|G), \quad (3b)$$

where $H(T|G, E)$ is the shape uncertainty arising from thermal configurational changes *per se*. Such terms will be called unreliability terms since their increase is concomitant to an increase in the number of possible configurations implicit in a given primary structure. (An alternative way of handling the unreliability term would be to eliminate it by defining entropies over the equivalence classes of configurations associated with any primary sequence, the assumption being that a unique equivalence class is implicit in the primary structure.)

(ii) *Phenotype-genotype degeneracy.* The folding relation implies that each protein shape, if specified in sufficient detail, is uniquely associated with a single primary structure (not in the sense that this primary structure could be computed, but in the sense that there would be no question as to the primary sequence in any new case once the association had been established *a posteriori*). Thus

$$H(G|K, T) = H(G|K, T, E) = 0. \quad (4)$$

It should be noted that this non-degeneracy does not extend to the DNA sequence which codes for the primary structure (because of the degeneracy of the code). Also, if the protein consists of different polypeptide chains, there are alternate ways in which the primary structure could be written, thereby giving rise to some degeneracy. For the organism as a whole, significant degeneracy of the genotype-phenotype relation is inevitably large since the number of alternative gene sequences which are equivalent in their final effect is potentially enormous.

(iii) *Functional dependence on structure.* According to the lock-key relation tertiary structure determines function, therefore T necessarily provides information about K . Rate constants are by definition average properties of a process and therefore not ordinarily regarded as fluctuations. Thus

$$H(K | T) = H(K | T, E) = 0. \quad (5)$$

According to the folding relation primary structure determines a thermal equivalence class of tertiary structures in a given milieu and therefore, if functional parameters are *bona fide* average properties of the various equivalence classes, it is possible to write

$$H(K | G) = H(K | G, E) = 0. \quad (6)$$

The consideration here is that if the averaging assumption is good, the association between primary structure and rate constants is always definite, even if the particular tertiary configuration can only be specified probabilistically. Also, the specification of the milieu provides no extra information since K represents the complete set of functional parameters, not the subset relevant to any particular environment.

C. Effect of primary on tertiary variability. Consider the joint entropy $H(K, T, G, E)$ and expand (in analogy to the law of multiplication of probabilities, cf. Shannon and Weaver, 1949) to give

$$H(K, T) + H(E | K, T) - H(K, T | G) - H(E | K, T, G) - H(G | K, T, E) = H(G), \quad (7)$$

where $H(K, T) = H(P)$, but it is here convenient to maintain the distinction

between function and structure. According to equation (4) $H(G|K, T)=0$ and thus it is possible to write

$$H(E|K, T) - H(E|K, T, G) = 0. \quad (8)$$

With these terms eliminated (7) simplifies to

$$H(K, T) - H(K, T|G) + H(G|K, T, E) = H(G). \quad (9)$$

According to (9) the uncertainty in the primary structure must either appear as phenotypic uncertainty (i.e. in the $H(K, T)$ term), where the actual amount of phenotypic uncertainty depends on the reliability of the relation between phenotype and genotype (i.e. on the term $H(K, T|G)$), or must be absorbed in the phenotype-genotype degeneracy term (i.e. in $H(E|K, T, G)$). For a single protein the degeneracy term is zero (cf. equation 4) and therefore absorption is not possible. In what follows it is shown that the restrictions discussed in the previous section also imply that gradualism in function change is possible if primary uncertainty is absorbed in tertiary structural uncertainty, but that this requires a definite form of structure-function degeneracy.

To isolate structural and functional uncertainties (9) may be expanded

$$\{H(K) + H(T|K)\} - \{H(K|G) + H(T|G, K)\} = H(G), \quad (10)$$

where the genotypic degeneracy term has been set to zero in accordance with (4). This may be further simplified, using restriction (6) to give

$$H(K) + H(T|K) - H(T|G, K) = H(G). \quad (11)$$

Thus $H(K)$ can be small (allowing for gradualism of function change) if $H(T|K)$ absorbs the primary uncertainty, where the actual magnitude of this term depends on the magnitude of the unreliability type term, $H(T|G, K)$.

To interpret $H(T|K)$ consider the identity

$$H(K) + H(T|K) = H(T) + H(K|T), \quad (12)$$

where the right-hand side is the alternative expansion of $H(K, T)$. In the extreme case of no function change $H(K) = H(K|T) = 0$ and therefore

$$H(T|K) = H(T). \quad (13)$$

This means that in order for tertiary structural uncertainty to completely protect functional properties from primary uncertainty, the functional properties must give no information about tertiary structure. If $H(K)$ is greater than zero

$$H(K) = H(T) - H(T|K), \quad (14)$$

since $H(K|T)$ is in any case zero (cf. equation 5). This means that function change can be made gradual to the extent that specification of function gives no information about tertiary structure. Under these circumstances the structure-function relation may be said to be degenerate.

Structure-function degeneracy, taken in the sense that many structures can perform the same function, is an important general feature of biological and other function performing systems (cf. Rosen, 1973b). Taken in the specific form required by (11), however, it would appear to be physically unrealistic. Two different proteins cannot reasonably be expected to be identical in all their rate constants. However, this means that an *a posteriori* detailed specification of all rate constants is just as good for specifying an equivalence class of tertiary structures as specification of primary structure, therefore that

$$H(T|K) = H(T|G, K) \geq 0. \quad (15)$$

In combination with (11) this gives

$$H(K) = H(G), \quad (16)$$

implying that under the given assumptions there is no way of buffering function from variation in primary structure.

The one assumption not considered so far is the assumption of functional specialization, *viz.* that different features of tertiary structure are more strongly associated with particular aspects of function than others (point (iv), section 2). Structure-function degeneracy is possible in the presence of such specialization provided that certain of the rate constants are irrelevant to any specific physiological function. Consider a protein which can be described in terms of two subsystems P^I and P^{II} (or two sets of features of the tertiary structure, T^I and T^{II} , along with sets of functional parameters, K^I and K^{II} , implicit in these features). Equation (9) can now be written

$$H(K^I, T^I, K^{II}, T^{II}) - H(K^I, T^I, K^{II}, T^{II} | G) = H(G), \quad (17)$$

where $H(K^1, T^1, K^{\text{II}}, T^{\text{II}}) = H(P^1, P^{\text{II}})$ and the genotype–phenotype degeneracy has again been set to zero. Expanding to a pattern analogous to (10)

$$\begin{aligned} & \{H(K^1) + H(T^1 | K^1) + H(T^{\text{II}} | K^1, T^1) + H(K^{\text{II}} | K^1, T^1, T^{\text{II}})\} \\ & - \{H(K^1 | G) + H(T^1 | G, K^1) + H(T^{\text{II}} | G, K^1, T^1) + H(K^{\text{II}} | G, K^1, T^1, T^{\text{II}})\} \\ & = H(G). \end{aligned} \quad (18)$$

Simplifying with restrictions of the form in (5) and (6), and also with the structure–function degeneracy restriction in (15)

$$H(K^1) + H(T^{\text{II}} | K^1, T^1) - H(T^{\text{II}} | G, K^1, T^1) = H(G). \quad (19)$$

The remaining structure–function degeneracy has the form $H(T^{\text{II}} | K^1, T^1)$ and the condition that $H(K^1)$ be small is now

$$H(T^{\text{II}} | K^1, T^1) > H(T^{\text{II}} | G, K^1, T^1). \quad (20)$$

There is no physical reason why this should not be possible since K^1 is the set of rate constants previously associated with T^1 and therefore would be expected to give very little information about T^{II} , whereas G specifies T^{II} up to thermal noise.

According to (19) it is possible to make the rate constants (or other functional attributes) implicit in one substructure of the protein change gradually, with the degree of gradualness being dependent on the extent to which specification of this substructure and its functional attributes fail to give information about some second substructure, *viz.* the one which absorbs the primary variations. In order for the function change of the enzyme as a whole to be gradual, the attributes implicit in this second substructure must be irrelevant to any specific function. The conclusion, so far as the structure–function relation in enzymes is concerned, is that there must be features of the structure irrelevant to any specific function if there is to be a mechanism for buffering function change in response to primary structure change (cf. Figure 1).

D. *The enzyme in an uncertain environment.* So far primary uncertainty has been considered the main source of noise. However, it is possible to shift the point of view, considering the environment as the main source of noise (in the sense of unexpected environmental disturbance) and primary uncertainty (along with its buffered effects on function) as providing one mechanism for coping with this noise. To do this expand $H(K, T, G, E)$ to give

$$H(K, T) - H(K, T | G, E) + H(E | K, T) - H(G | E) = H(E), \quad (21)$$

where the genotype-phenotype degeneracy term, $H(G|K, T, E)$ has, as usual, been set to zero, $H(G|E)$ is the uncertainty of the primary sequence once the environment (milieu of the enzyme) has been specified, and $H(E|K, T)$ represents the degree to which both structure and function of the enzyme forget (or are indifferent to) this milieu. High indifference means that the enzyme is capable of performing the same function in a variety of milieus and without allosteric shape change.

Now, taking the possibility of functional specialization into account

$$H(K^I, T^I, K^{II}, T^{II}) - H(K^I, T^I, K^{II}, T^{II} | G, E) + H(E | K^I, T^I, K^{II}, T^{II}) - H(G | E) = H(E). \quad (22)$$

Expanding

$$H(K^I) + \{H(T^I | K^I) - H(T^I | G, E, K^I)\} + \{H(T^{II} | K^I, T^I) - H(T^{II} | G, E, K^I, T^I)\} + H(E | K^I, T^I, K^{II}, T^{II}) - H(G | E) = H(E), \quad (23)$$

where the terms $H(K^{II} | K^I, T^I, T^{II})$, $H(K^I | G, E)$, and $H(K^{II} | G, E, K^I, T^I, T^{II})$ have been set equal to zero in accordance with (5) and (6). Note that the structure-function degeneracy restriction (equation 15) cannot in general be used to eliminate terms in the leftmost bracket (as was the case when the environment was not specified, cf. (18) and (19)). This is because specification of all relevant rate parameters does not determine the particular allosteric form, whereas specification of the genotype and the milieu does. Thus these terms represent the amount of environmental uncertainty that is absorbed in allosteric transitions.

The entire left-hand side of (23) may be called the adaptability of the enzyme (actually of an enzyme species). This splits into two parts, phenotypic adaptability associated with short term changes in the milieu and evolutionary adaptability associated with long term changes (i.e. changes originating in long term changes in the external environment or from changes at other gene loci). The phenotypic part consists of allosteric adaptability and indifference. If these are not sufficient to absorb the short term component of uncertainty, the only other possibility is that this is absorbed in functional uncertainty which is divorced from genetic change and therefore concomitant to denaturation or other diminution of function. The evolutionary component of adaptability consists of (genetically dependent) functional and buffer system adaptability. The latter is included in the terms in the second set of brackets in (23). However, note that in

general

$$H(T^{\text{II}} | G, K^{\text{I}}, T^{\text{I}}) \geq H(T^{\text{II}} | G, E, K^{\text{I}}, T^{\text{I}}), \quad (24)$$

implying that the terms in the second brackets may also include a contribution from environmentally induced shape change. These shape changes might also be concomitant to denaturation or may reflect functionally significant shape changes in T^{I} . Potentially they could also serve to absorb short term environmental changes. However, this would reduce the effectiveness of buffering since features of T^{II} would then be assuming significance for aspects of specific function. It should also be noted that if G were the genome of the whole organism $H(G) - H(G|E)$ would represent the information which the selecting environment provides about the primary structure; but in the present case E is a local milieu, buffered by many levels of biological control, and therefore the difference between $H(G)$ and $H(G|E)$ only indirectly reflects selection.

If $H(K^{\text{I}}) > 0$, this means that the enzyme species undergoes a function change which meets the environmental change. If both the environmentally conditioned uncertainty, $H(G|E)$, and the buffer system term, $H(T^{\text{II}} | K^{\text{I}}, T^{\text{I}})$, are larger, this function change can be more gradual in the sense that the degree to which T^{I} and K^{I} forget K^{II} can increase for a given value of $H(K^{\text{I}})$, essentially allowing for greater absorption of mutation in T^{II} and therefore finer modulation of the expression of this mutation in T^{I} . In the extreme case this absorption is complete and $H(K^{\text{I}}) = 0$. However, any long term change in the environment would then have to be met by change in T^{II} and therefore K^{II} . In this case the initially physiologically functionless buffer system assumes a specific physiological function, or some subunit of this system assumes such a function. This is of course evolution by transfer of function, except that the transfer is from an evolutionary (buffer system) function to a physiological function rather than from one physiological function to another. If, in the course of evolution, the enzyme assumes more particular functions, the relative importance of the buffer system would have to increase in order to maintain the same degree of gradualism, or more effective forms of buffering would have to develop.

4. Mechanisms of Mutation-Absorption. A flow diagram of the mutation-absorption model is illustrated in Figure 1. The enzyme is pictured (as in the preceding analysis) as consisting of two subsystems, with the first involved with traditional functions such as recognition, control and catalysis, and the second involved with buffering the effects of mutation

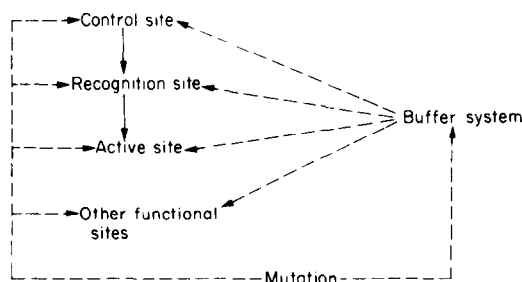


Figure 1. Functional diagram of mutation-absorption model. Solid arrows represent interactions among features of enzyme structure specialized for specific physiological functions. Dotted arrows represent the influence of mutation. The buffer system absorbs some of the effects of mutation, thereby modulating its expression in functionally specific features. This system thus subserves an evolutionary function, implying that enzyme structures should be interpreted as embedding evolutionary as well as physiological functions

and therefore modulating the evolutionary transformation of these functions. A mechanical, balls and springs analog is pictured in Figure 2 (with balls representing amino acids and springs representing strong or weak bonds). As the number of interconnected balls increases, a mutation from one size ball to another has a smaller effect on a feature of geometry critical for function (represented by the distance between two triangles). The perturbation of the structure could also be decreased by increasing the number of interconnections for a given number of balls or by increasing the number of types of balls (e.g. allowing for more graduations of size). In molecular terms these mechanisms translate to:

(i) *redundancy of weak bonding*, i.e. utilization of amino acids or amino acid combinations which increase the number of weak bonds influencing critical features of the shape;

(ii) *quantitative redundancy of amino acids*, i.e. insertion into the primary sequence of amino acids whose principle contribution to function is to increase the number of weak bonds which influence critical features of the shape;

(iii) *redundancy of amino acid types*, i.e. insertion into the primary sequence of amino acids which have a greater number of close structural analogs, thereby making it possible for substitutions to give rise to more graded perturbation of the weak bonding influencing critical features;

(iv) *specific organizational formats*, i.e. utilization of local or global structural formats which increase the effectiveness of mechanisms (i)–(iii).

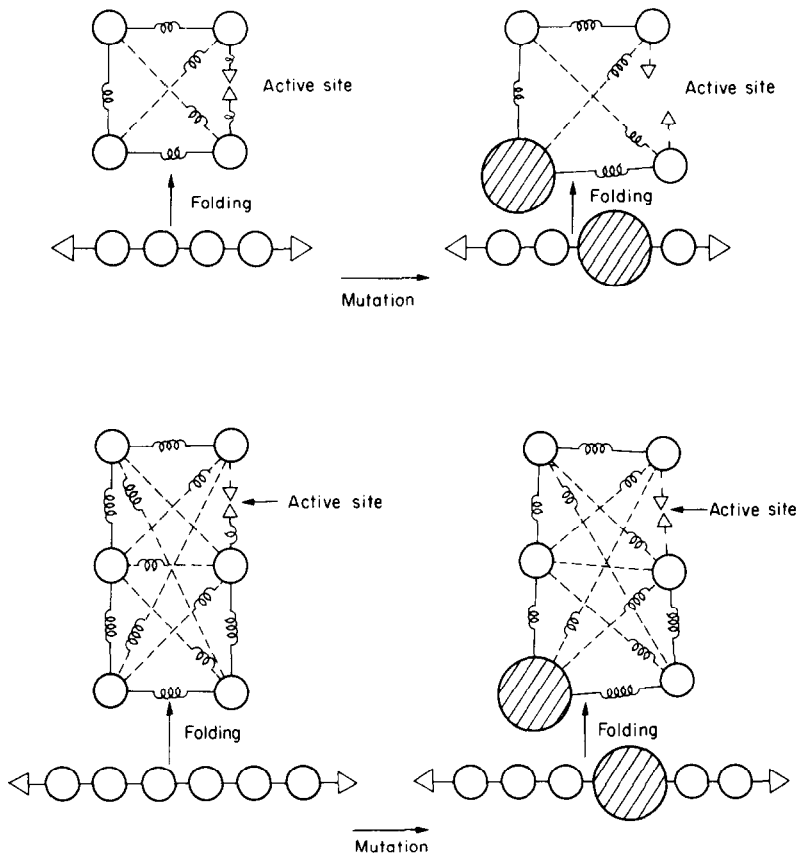


Figure 2. Balls and springs analog of the mutation-absorption model. Springs on solid lines represent strong bonds and springs on dashed lines weak bonds. Small balls, large balls, and triangles represent different types of amino acids. The feature of folded shape critical for function (e.g. catalytic specificity) is the relative position of the triangles. If more amino acids are added to the structure, it is possible for typical, but noncritical mutations (involving the replacement of a small ball by a large ball) to produce smaller modifications in the active site (i.e. smaller changes in the distance or angle between the triangles) as well as a larger variety of modifications. The increase in gradualism derives from an increase in the redundancy of weak bonding. However, it could also be based on the utilization of amino acids with a greater number of structurally similar (or redundant) analogs and also on specific organizational formats which allow for effective utilization of redundancy. The balls and springs analog schematically illustrates mechanisms of mutation buffering and is not intended as a realistic model of protein structure

Mechanisms (i) and (ii) both imply an increase in the free energy of folding and mechanisms (iii) and (iv) may be associated with such an increase as well (if closer packing or more ordered interactions become possible). Thus these interactions contribute to the stability of the folded

structure as well as the degree of gradualism with which this structure typically changes in response to changes in primary structure. These two roles are mixed and in this sense what has been called a buffer system is also a stabilization system. However, it is possible to distinguish unambiguously the extent of overall contribution to stability from the contribution to gradualism. This is because stability can be identified with the reliability of the tertiary–primary structure relation and therefore with the actual magnitudes of the terms $H(T^{\text{II}}|K^{\text{I}}, T^{\text{I}})$ and $H(T^{\text{II}}|G, K^{\text{I}}, T^{\text{I}})$, whereas the degree of gradualism depends on the difference between these terms. This is what justifies a conceptual distinction between buffer system and features of the shape critical for function.

The most remarkable known format allowing for gradualism is undoubtedly the “lobster” type structure of the immunoglobulins. What are believed to be opposing loopings of each claw (light chain and hinged segment of heavy chain) make it possible to generate an enormous variety of structures in response to variation at variable and hypervariable locations on the claws, yet within the framework of the same basic structure (cf. Roitt, 1974). However, loopings mean redundancy of amino acid numbers (as far as the specific function of each particular immunoglobulin is concerned) and also increased redundancy of weak bonding. The immunoglobulins are clearly an extreme case since graded variation in structure is the *sine qua non* of their physiological function. However, they provide a graphic instance of an aspect of function which is less obvious but of fundamental importance in other proteins.

5. Implications of the Mutation–Absorption Paradigm. The enzyme concept developed in the previous section has a number of implications. These include:

(i) *Interpretation of structure–function relations in proteins.* The buffer system function must be added to the list of functions embedded in enzyme structure. The functional significance of the number of types of amino acids, structural similarities among these amino acids, local formats of folding, and overall global organization cannot be adequately understood without considering their contribution to buffering the expression of mutation.

(ii) *Effect of natural selection on the buffer system.* So far the argument has been that a buffer system of a definite type and realized by certain definite mechanisms is a requirement for gradualism. The assumption of gradualism can be justified on the basis of the experimental observation that slight changes in function and structure frequently follow stepwise changes in primary structure (e.g. Rossman *et al.*, 1975; Wills, 1976), in which case the buffer system can be considered to be implied by experi-

ment. The argument can also be made that gradualism is a precondition for evolution and therefore that the presence of a buffer system can be inferred from the fact of evolution. A third, more self-contained argument is that the degree of gradualism, like any other enzyme property, is subject to selection and therefore the degree of buffering must be an evolved property, adjustable in the course of evolution (Conrad, 1972a, 1977b, c). Furthermore, it is inevitably subject to strong selection since selection for gradualism will always be tied to selection for any new protein property whose likelihood of appearance is increased by this gradualism. According to the evolutionary argument, one version of a protein may be favored over another not because they differ in any physiological properties, but only because they differ in the degree to which they buffer mutation. This is why the relation between structure and function cannot be understood without considering the buffering function.

(iii) *Effect of buffering on the structure and smoothness of the adaptive landscape.* Suppose an adaptive landscape is defined in terms of a fitness axis, a set of environment axes, and a set of axes for specifying primary structure (e.g. using separate axes for each possible amino acid at each possible position or, alternatively, distinct amino acid axes for each amino acid type). The fitness "hypersurface" will have a certain global structure (the topography of the adaptive peaks and valleys) and also a local structure, which determines the ease with which particular peaks can be climbed by an evolutionary system. Thus, a particular pathway, either up or between peaks, may be smooth, in the sense that it can be traversed by a series of single step changes in primary structure which do not involve any significant decrease in fitness, or it may be rough in the sense that multistep changes in primary structures are necessary to jump local crevices in fitness.

Now, suppose that buffering increases. This increases the gracefulness with which mutation is typically absorbed and expressed in protein shape and therefore the gradualness of function change in response to primary structure change also increases. In the adaptive landscape picture this must appear as movement to a region of the landscape where the pathways are smoother and perhaps more numerous. The evolution of buffering corresponds to an evolutionary movement to a region of the adaptive landscape where evolutionary hill climbing is most effective, and furthermore (recalling the mechanisms of mutation absorption discussed in the previous section), such an evolutionary movement can always take place in a gradual way.

(iv) *Energy required for smoothing.* Smoothing of the adaptive landscape (in the sense defined above) involves a certain expenditure of energy (since utilization of more amino acid types, longer primary chains,

and organizational formats especially suitable for gradualism all involve extra energy expenditures). Thus the mutation-absorption model implies that an increase in the energy invested in the performance of specific functions may be exchanged for a decrease in the energy which must be expended in evolutionary search. Since the specific functions could in general be performed with less energy investment, the ascended peaks are not in general the highest possible (in terms of the fitness axes). However, they are the ones which are most easily climbable and therefore the ones which allow the most rapid increase in fitness. In effect, the pathway of most rapid ascent is in general a pathway to more nearly optimal gradualism.

(v) *Relation between buffering, neutralism, and selection.* The mutation-absorption model provides a framework for interpreting the apparent contradiction between selectionist and neutralist, so-called non-Darwinian theories of evolution. According to the classical selectionist theory the decisive directing influence in evolution is natural selection. According to the neutralist theory, this is contradicted by the statistical evidence for neutral drift of primary sequence (cf. King and Jukes, 1969). However, the condition for an effective evolutionary response to selection is gradualism of function change in response to primary structure change. According to the mutation-absorption model the buffer system makes this possible and the buffer system is effective when $H(G|E)$ and $H(T^{\text{II}}|K^{\text{I}}, T^{\text{I}})$ are both large. Under these circumstances many variations of the primary structure have the same or similar functional properties and therefore the model implies that selection and neutral phenomena are inherently linked—since buffering is a necessary condition for the efficacy of the former and a sufficient condition for the occurrence of the latter.

The above considerations are also relevant to interpreting the surprisingly large number of variations on primary structure (allozymes and isozymes) found in single species and single organisms. These are also inevitable concomitants of mutation absorption. However, as in selective evolution, neutral and non-neutral variations grade into one another and therefore subtle but significant differences among at least some of the variations cannot be precluded. On the contrary the presence of a mutation buffering system makes it practically impossible for the organism or the species to escape carrying a reservoir of potentially significant variants, even if only a minority of these variants have current selective value.

6. *Conclusions and Extensions.* Commonly made generalizations about structure-function relations in proteins have implications for the relation between change in primary structure and change in function. The main

implication is that gradualism in function change is possible only if the protein can be divided (conceptually at least) into two subsystems, one critical for specific function and one buffering the transformation of this specific function in response to mutation. This buffer system is implemented by redundancy of weak bonding (which may vary with the format of the molecule or with the number of amino acids) and also by redundancy of amino acid types (which may vary if the number of amino acids with close structural analogs increases). Thus the enzyme is conceived as a control system in the sense that these mechanisms control the expression of mutation. This is in addition to the control mechanisms and other mechanisms which implement the various specific functions in terms of which enzymes are normally described (e.g. recognition, allosteric shape change, catalysis). Indeed, the buffer system may be picturesquely described as a submolecular organ system of adaptability, both from the standpoint of evolution and from the standpoint of maintaining reservoirs of enzyme types within the organism.

The analysis can be extended to the organism as a whole, in the direction of a general model of genotype–phenotype relations based on the requirement of gradualism. The situation here is clearly more complicated, partly because the simple general statements of the type made about enzymes are not available and many of the simplifying assumptions used in the present analysis (e.g. elimination of tertiary–primary degeneracy, consideration of only mutation or intragenic recombination) are no longer valid. However, the point of view does generalize in the sense that it is reasonable to expect mechanisms of development and morphogenesis to be such that gradual transformation of structure (as in the topological transformations of D'Arcy Thompson, 1917) and function (as in the relational mappings of Rashevsky, 1960) is possible, implying smoothing of pathways in the adaptive landscape and investment of energy to achieve smoothing.

The mutation–absorption paradigm has a peculiar epistemological consequence which perhaps captures the essence of its implication for the interpretation of structure–function relations in proteins or in any higher level system to which it can be extended. As gradualism increases to the point where single genetic changes are likely to give rise to gradual changes in function, the pathways in the adaptive landscape become more easily traversible. This reduces the difficulty of evolutionary hill climbing for the system and also simplifies the problem of describing the sequence of events in evolution. However, this simplification is at the expense of expanding the buffer system and therefore complicating the structure required to perform any specific function and also complicating the task of describing the relation between structure and function.

LITERATURE

- Conrad, M. 1972a. "Information Processing in Molecular Systems." *Currents in Modern Biology* (now *BioSystems*), **5**, 1-14.
- . 1972b. "Statistical and Hierarchical Aspects of Biological Organization." In *Towards a Theoretical Biology*, Ed. Waddington, C. H. Edinburgh: Edinburgh University Press.
- . 1977a. "Functional Significance of Biological Variability." *Bull. Math. Biol.*, **39**, 139-156.
- . 1977b. "Evolutionary Adaptability of Biological Macromolecules." *J. Molec. Evolution.*, **10**, 87-91.
- . 1977c. "Evolution of the Adaptive Landscape." To appear in *Theoretical Approaches to Complex Systems*, Eds. R. Heim and G. Palm. Lecture Notes in Biomathematics. Heidelberg: Springer-Verlag.
- King, J. L. and T. H. Jukes. 1969. "Non-Darwinian Evolution." *Science*, **164**, 788-798.
- Maynard Smith, J. 1970. "Natural Selection and the Concept of a Protein Space." *Nature*, **225**, 563-564.
- Perutz, M. F. 1962. *Proteins and Nucleic Acids*. Amsterdam: Elsevier Publishing Company.
- Rashevsky, N. 1960. *Mathematical Biophysics*, 3rd Ed. 2 vols. New York: Dover.
- Roitt, I. 1974. *Essential Immunology*, 2nd Ed. Oxford: Blackwell Scientific Publications.
- Rosen, R. 1973a. "On the Relation Between Structural and Functional Descriptions of Biological Systems." In *The Physical Principles of Neuronal and Organismic Behavior*, Eds. M. Conrad and M. Magar, London: Gordon and Breach.
- . 1973b. Discussion, p. 232. In *The Physical Principles of Neuronal and Organismic Behavior*, Eds. M. Conrad and M. Magar. London: Gordon and Breach.
- Rossmann, M. G., A. Liljas, C. Branden and L. J. Banaszak. 1975. "Evolutionary and Structural Relationships among Dehydrogenases." In *The Enzymes*, 3rd ed. vol. 11. Ed. P. D. Boyer. New York: Academic Press.
- Shannon, C. and W. Weaver. 1949. *Mathematical Theory of Communication*. Urbana: University of Illinois Press.
- Thompson, D'Arcy Wentworth. 1917. *On Growth and Form*. Cambridge: Cambridge University Press.
- Wills, C. 1976. "Production of Yeast Alcohol Dehydrogenase Isoenzymes by Selection." *Nature*, **261**, 26-29.
- Woodward, P. M. 1973. *Probability and Information Theory with Applications to Radar*. Oxford: Pergamon Press.

RECEIVED 8-10-77

REVISED 2-6-78