Sridhar Govindarajan[1]
Richard A. Goldstein[1,2]
[1] *Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055*

[2] *Biophysics Research Division, University of Michigan, Ann Arbor, MI 48109-1055*

Received 31 December 1996; accepted 1 April 1997

The Foldability Landscape of Model Proteins

Abstract: *Molecular evolution may be considered as a walk in a multidimensional fitness landscape, where the fitness at each point is associated with features such as the function, stability, and survivability of these molecules. We present a simple model for the evolution of protein sequences on a landscape with a precisely defined fitness function. We use simple lattice models to represent protein structures, with the ability of a protein sequence to fold into the structure with lowest energy, quantified as the foldability, representing the fitness of the sequence. The foldability of the sequence is characterized based on the spin glass model of protein folding. We consider evolution as a walk in this foldability landscape and study the nature of the landscape and the resulting dynamics. Selective pressure is explicitly included in this model in the form of a minimum foldability requirement. We find that different native structures are not evenly distributed in interaction space, with similar structures and structures with similar optimal foldabilities clustered together. Evolving proteins marginally fulfill the selective criteria of foldability. As the selective pressure is increased, evolutionary trajectories become increasingly confined to ''neutral networks,'' where the sequence and the interactions can be significantly changed while a constant structure is maintained.* © *1997 John Wiley & Sons, Inc. Biopoly* ***42:*** *427–438, 1997*

Keywords: *protein folding; molecular evolution; lattice models; fitness landscapes; spin glasses*

INTRODUCTION

Proteins are the result of a long evolutionary process. We can better understand the properties of these biological macromolecules by considering the manner in which they arose. For instance, evolutionary considerations can explain how the plasticity of protein sequences can coexist with a tremendous robustness of the resulting structures, why certain native structures are overrepresented among biolog-

Correspondence to: Richard A. Goldstein
Contract grant sponsor: College of Literature, Science, and the Arts; Program in Protein Structure and Design; Horace H. Rackman School of Graduate Studies; NIH; NSF
Contract grant number: LM0577 (NIH) and BIR9512955 (NSF)

CCC 0006-3525/97/040427-12

ical proteins, and what interactions dominate in the folding process and why.[1–8] Considering evolutionary relationships can assist in predictions of the structures of specific proteins (for reviews, see Refs. 9–11). Ancestral protein sequences can be reconstructed based on the sequences of modern existent proteins.[12–15] These ancestral proteins can then be synthesized in the laboratory and their biochemical properties measured, providing information about how their specific functional features arose.[16–18] All of these applications rely on understanding the process of molecular evolution, specifically as it occurs in proteins.

One of the major advances in our thinking about evolution was the introduction of the concept of a *fitness landscape,* representing the fitness of a particular system (such as a protein) as a function of the parameters of that system (the protein sequence).[19] Evolution occurs as movement in this landscape. Just as the dynamics of mechanical motion is determined by the energy landscape, the characteristics of the evolutionary dynamics is determined by the fitness landscape. While some of the more general features of evolution have been investigated through the use of abstract models,[20–23] any attempt to understand the characteristics of biological macromolecules must take into account the nature of the fitness landscape for these particular molecules. Previous studies have focused on the specific landscapes for RNA, and more recently, proteins, considering the mapping of sequence to structure.[24–28] While highly instructive, such studies are not based on fitness in an evolutionary sense, except for the need for the lowest energy conformation to be nondegenerate. As such, it is not possible to consider how *selective pressure,* the pressure due to natural selection, would affect the evolutionary dynamics.

One of the reasons why the fitness landscape approach has mostly relied on abstract models is the difficulty of applying the concept to complicated biomolecules such as proteins in the appropriate biochemical context. Proteins represent a large class of different molecules that fulfill diverse structural and functional roles. The fitness of a protein is complicated and multifaceted, involving function, stability, and survivability, and is specific for each particular protein. One property common to essentially all proteins, however, is that they have to be able to fold rapidly in order to avoid irreversible processes such as proteolysis or aggregation. As pointed out by Levinthal, protein folding is a nontrivial process, involving finding the native conformation in a time too short to sample more than a minuscule fraction of all possible conformations.[29] While ‘‘foldability’’ is not the only factor comprising the fitness function, it must be a major component in the selection process. By understanding how proteins evolve based on the need to maintain an adequate foldability, we can understand a major force governing natural selection as it occurs on the molecular level. In addition, based on the observation that proteins with similar folds often have completely unrelated functional roles, and similar functional roles may be performed by proteins with unrelated structures,[30] we can consider that the selection pressure that acts on structures may be in some sense orthogonal to the selection pressure that is based on functional properties. In such a case, we can understand protein structures by considering only selection pressures that act on structural properties such as the ability to fold.

One major obstacle to studying the folding behavior of biological proteins is that we do not understand the interactions in the protein that govern the folding process, especially interactions between the protein and the solvent. Shakhnovich and co-workers developed a novel approach to answering questions about protein folding and evolution by postulating an alternative world where biological proteins exist on three-dimensional lattices, and where the true energy function is known in advance.[31,32] By modeling evolutionary selection given the exact energy function, it is possible to model the folding behavior of evolutionarily optimized proteins, at least optimized for the ‘‘alternative’’ universe under study. Confining the proteins to a lattice makes it possible to do an exhaustive enumeration of all maximally compact structures, allowing the exact computation of thermodynamic properties. This approach also allows us to take advantage of theoretical work that has greatly increased our understanding of the properties of proteins that allow for rapid folding (for reviews, see Refs. 4, and 33–37). These simple models allow us to compute measures of the foldability, and to explore the nature of the resulting fitness landscape and how the need to fold can influence the course of protein evolution. The hope is that the model energy function is qualitatively similar to the real energy function, and the questions being addressed sufficiently coarse-grained, so that the approximate nature of the theoretical model does not compromise the results.

In recent work, we used simple lattice models to consider the properties of proteins optimized for maximum foldability, and the relationship between the optimizability of different structures and how often these structures would be found among biological proteins.[2,3,6] In this paper, we describe results concerning the evolution of proteins, where the fitness is defined by the foldability. We look at both static and dynamic properties on this foldability landscape. We characterize the nature of the landscape, investigating the mapping between sequences, interactions, native states, and fitnesses. We also describe how we can identify points on the foldability landscape representing specific amino acid sequences, model the changes of the sequence as the protein undergoes evolution, and observe how this is influenced by the degree of selective pressure. We observe that as the selective pressure increases, the evolutionary dynamics become confined to ''neutral networks'' where the structure remains fixed, even as the sequence and interactions are rapidly modified. We also find evidence for the marginal foldability and stability of proteins that arise during evolution.

THEORY

Spaces, Metrics, and Mappings

Our lattice model consists of 27-residue proteins confined to a $3 \times 3 \times 3$ cubic lattice, where each residue occupies one lattice point. The interresidue distance is equal to the lattice spacing.

We are concerned with three different spaces, the *sequence space,* the *interaction space,* and the *conformation space.* The sequence space $\mathcal{A}$, first introduced by Maynard-Smith, represents the space of all possible amino acid sequences $A^k \equiv \{a_i^k\}$, where a_i^k is the amino acid in the ith position in sequence k.[38] Our 27-residue model proteins consist of the 20 naturally occurring amino acids, so the sequence space contains 20^{27} discrete points. The most natural metric between different points in this space is the Hamming distance h_{kl}, representing the number of amino acid changes necessary to go from sequence k to another sequence l. Unfortunately, the interactions between residues in a protein are more complicated than the simple base matching that occurs in other biological macromolecules such as RNA. There are some mutations (a threonine to a serine) that are much more conservative than others (a cysteine to a phenylalanine). The magnitude of the effect of any mutation will also depend upon the location of the mutation in the protein structure and on the identity of the other residues in the protein. For these reasons, the Hamming distance is not the most useful metric for measuring the distance between sequences.

We therefore consider a separate space, the interaction space $\mathcal{I}$, representing the space of all pairwise contact interactions in the protein $\Gamma^k \equiv \{\gamma_{ij}^k\}$, where γ_{ij}^k represents the energetic contribution when residues i and j of protein sequence k come into contact. The use of pairwise contact potentials is motivated by work that showed optimal folding occurs when local structural propensities are relatively weak,[3] and by the fact that local structural propensities are not highly conserved during site mutations.[7] We further assume that the contact potentials only depend on the identity of the residues making contact, so that $\gamma_{ij}^k = \gamma(a_i^k, a_j^k)$. For a 27-residue protein on a cubic lattice, there are exactly 156 possible pairs of residues that can come into contact, so the interaction space is a 156-dimensional space with any specific sequence corresponding to a point in this space. The mapping $\Gamma^k = \Gamma(A^k)$ from the sequence space $\mathcal{A}$ to the interaction space $\mathcal{I}$ is determined by the exact form of $\gamma(a_i^k, a_j^k)$. We use the parameter values originally derived by Miyazawa and Jernigan.[39] As these parameters were derived from a data base of protein structures, they represent potentials of mean force that implicitly include interactions of the protein with the solvent.

In the simple model that we describe below the properties of the protein are only dependent of the relative values of the interaction parameters, and are insensitive to either additive or multiplicative constants. For this reason, we scale all of the parameters so that the average interaction in the protein is equal to zero and the sum of the squares of the interactions is equal to one. This corresponds to projecting the high-dimensional interaction space onto a unit hypersphere. In the present study we use the angular distance θ_{kl} between points on the hypersphere surface as a measure of the proximity of two points in this space.

The properties of the proteins that we are interested in, including the native state $\mathcal{N}^k$ and the foldability-fitness $\mathcal{F}^k$, are functions of the point in the interaction space and thus of the sequence: $\mathcal{N}^k = \mathcal{N}(\Gamma^k) = \mathcal{N}(\Gamma(A^k))$, and $\mathcal{F}^k = \mathcal{F}(\Gamma^k) = \mathcal{F}(\Gamma(A^k))$. In order to compute $\mathcal{N}$ and $\mathcal{F}$ we need to consider the third space, the conformation space, which is the physical space in which the protein exists. For our 27-residue protein confined to a $3 \times 3 \times 3$ lattice, this is simply the set of 103,346 possible conformations (neglecting rotations and reflections) represented by selfavoiding walks through the lattice. The energy of any conformation m is of the simple form

$$E^m = \sum_{i<j} \gamma_{ij}^k \Delta_{ij}^m \qquad (1)$$

where Δ_{ij}^m is equal to one if residues i and j are not adjacent in sequence but are on adjacent lattice sites in conformation m, and zero otherwise. Every compact conformation contains exactly 28 residue pairs in contact, corresponding to 28 nonzero values of Δ_{ij}^m. For any set of interactions Γ^k we can calculate the energy of every possible compact conformation, and identify the native state $\mathcal{N}^k$ as the compact state of lowest energy. The natural metric between native states $\mathcal{N}^k$ and $\mathcal{N}^l$ is the q value, representing the percentage of the contacts that are the same in both states:

$$q_{kl} = \frac{1}{28} \sum_{i<j} \Delta_{ij}^{\mathcal{N}^k} \Delta_{ij}^{\mathcal{N}^l} \qquad (2)$$

Two identical conformations will have a q value of 1. The average q value between two random structures is approximately $q = 0.19$, with a standard deviation of 0.07.

Foldability and Fitness

Recent theoretical work has concentrated on the thermodynamic requirements necessary for rapid folding. In analogy to spin glasses in condensed matter physics, two thermodynamic transitions for the protein are considered: the first a transition to the native state at some temperature T_f and the second to a glassy phase at a temperature T_g.[40] If the glass transition temperature is higher than the folding temperature, then the protein will not be able to fold but will instead get trapped in local minima. Close to the glass transition temperature, the dynamics will be slow and non-Arrhenius. In order for rapid folding to be possible, the protein must be able to fold at a temperature sufficiently high relative to the glass transition temperature. As the physiological temperature must be below T_f in order for the protein to be thermodynamically stable, T_f must be significantly higher than T_g.

Using the Random Energy Model, it can be shown that T_f/T_g is directly related to the ratio of the energy gap between the native state and the random conformations and the width of the distribution of energies of these random conformations.[41–43] We define this ratio as the foldability $\mathcal{F}$, which is the fitness in our simple model. Wolynes and co-workers showed how this ratio could be maximized in order to produce energy functions optimized for tertiary structure prediction.[41–43] A discretized form of this optimization procedure was used by Shakhnovich and co-workers in order to produce sequences that would fold readily in Monte Carlo simulations.[44–47] A criterion related to $\mathcal{F}$ was found by Chan and Dill and by Karplus and co-workers to distinguish random sequences that fold easily from those that do not.[31,32,48] As we can compute the energy of every compact conformation of our model proteins, $\mathcal{F}$ is easily calculated. This assumes that only the compact conformations are thermodynamically relevant. This assumption has been supported by recent theoretical work that shows that the generic forces favoring compaction should be quite strong under optimal folding conditions.[49]

It is possible to solve in closed form for the set of interaction parameters $\Gamma_{\text{opt}}^k \equiv \{\gamma_{ij}^k\}_{\text{opt}}$ that maximizes the foldability for any native structure $\mathcal{N}^k$, and to calculate the optimal foldability $\mathcal{F}_{\text{opt}}^k$, the foldability at this optimum.[41–43] There is roughly a Gaussian distribution of $\mathcal{F}_{\text{opt}}^k$ values among the 103,346 native structures, with a mean value of 12.44 and a standard deviation of 0.37.[2] One corollary of the existence of this optimization procedure is the fact that the optimal point in the interaction landscape for any native structure is unique. The set of optimal parameters can therefore serve as a way of defining the location of the different native structures in the interaction space. Because the number of possible sequences for even a short protein of length 27 is so much larger than the number of possible structures (by 30 orders of magnitude), the interaction space is rather densely populated relative to the topological features in the landscape—the median distance between two nearest neighbors is only $\theta = 0.2\pi$. The distribution of random sequences in the interaction space based on the Miyazawa–Jernigan potentials is indistinguishable from the random distribution of points in $\mathcal{I}$, indicating that the discrete points corresponding to specific sequences are relatively uniformly distributed in the interaction space. This allows us to either consider the discrete points in interaction space corresponding to different sequences, or alternatively to consider the interaction space as a continuous space with smooth peaks, valleys, ridges, and plateaus.

RESULTS AND DISCUSSION

In this section, we analyze the model described above. Because of our ability to rapidly find the set of interactions corresponding to fitness maxima, we first concentrate on the nature of the interaction space and the locations of these maxima, ignoring the fact that real sequences correspond to discrete points in this space and that the various fitness maxima are not likely to correspond to actual sequences. The questions we would like to address include the following:

1. How are the fitness maxima distributed through the space?
2. Do similar native structures have fitness maxima that are near each other in the interaction space?
3. Are maxima corresponding to especially high or low fitness values located close to each other in the interaction space?

We then consider the sequence space as well, and investigate how sequence mutations correspond to movement in the interaction space and how the need-to-fold effects this movement. This allows us to address an additional set of important questions:

1. How does movement in sequence space correspond to movement in the interaction space?
2. How does the fitness landscape influence the movement in sequence space, affecting the relative rates of different mutations?
3. What is the distribution of foldability values among evolving sequences?
4. As the sequences and interactions change, what happens to the structures?

The rest of the paper is an attempt to generate answers to these questions.

Statistics of Fitness Optima

We compute $\{\gamma_{ij}^{k}\}_{\text{opt}}$ for each of the 103,346 compact structures in the lattice, and look at the distribution of these various fitness maxima in the interaction space with respect to each other. Figure 1 shows $g(\theta)$, the pair-correlation function for the maxima, representing the probability that two maxima are located a distance θ apart in the interaction space. This is compared with the distribution of pairs of random points in this space, which follows a $\sin^n\theta$ distribution where n is the dimension of the interaction space; due to the nature of the very high dimensional hypersphere, this function has a relatively sharp maximum at $\theta = \pi/2$.

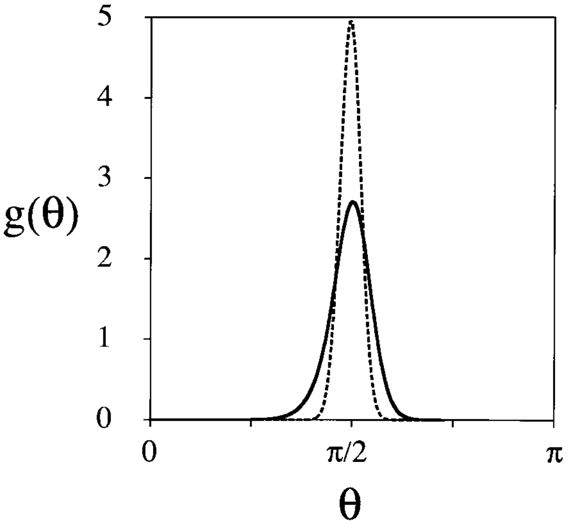

FIGURE 1 Pair-correlation functions $g(\theta)$ for the distribution of optima in interaction space (——) compared with the distribution of random points (– – –).

The significant deviation of the two curves indicates that the optima are *not* distributed randomly throughout the interaction landscape. The long tail to the distribution extending in the direction of smaller θ values indicates that the optimal maxima for different structures are quite clustered. For instance, there is approximately a 3% probability of two structures having optima with $\theta < 0.4\pi$, more than 400 times what would be expected at random. This might be expected, as the interactions that optimize one conformation might be quite similar to the interactions that optimize a different but similar conformation. Interestingly, the distribution also extends further in the direction of larger θ values, suggesting that different optima repel each other. In order for a protein to be able to fold, not only must the correct conformation be stabilized, but incorrect conformations must be destabilized. This effect may explain the tendency of optima corresponding to very different structures to be far apart in interaction space, as the interactions become optimized so that radically different structures are of high energy.

These interpretations are supported by looking at how the distances in interaction space correlate with the differences in the native structure. Figure 2 shows a density plot of $P(q|\theta)$, the conditional probability that two structures whose optima are separated by θ in interaction space correspond to native structures related by a given value of q. As shown, the neighborhood effect shows that similar structures are close together in the interaction landscape, with the similarity of structures gradually

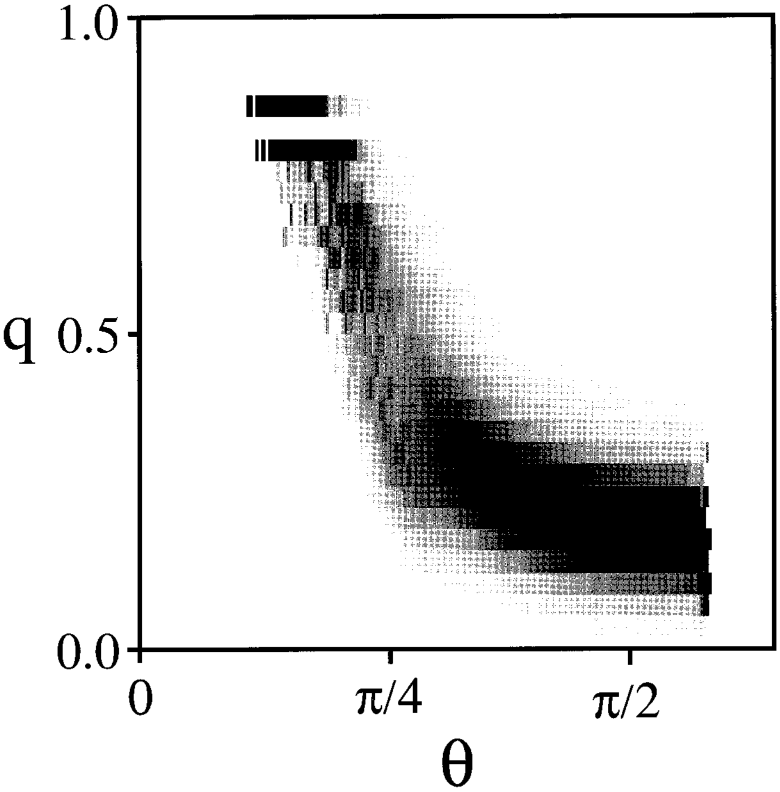

FIGURE 2 Density plot of $P(q|\theta)$, the conditional probability that structures whose optima are separated in interaction space by an amount θ will have similar native structures, as measured by q. As shown, similar structures are relatively close in interaction space, while dissimilar structures are actually more separated than would be expected at random.

decaying with a persistence length of approximately $\theta = \pi/4$. Structures separated by more than $\theta = \pi/2$ have an average q values even smaller than what would be expected for pairs of random structures.

In previous work, we showed that similar structures have similar values of $\mathcal{F}^k_{\mathrm{opt}}$.[2] Given the preceding results indicating that similar structures are located near each other in interaction space, it is reasonable to postulate that there would be a tendency for optima of similar heights to be clustered. Figure 3 shows this to be the case; the heights of maxima are correlated on a scale comparable with the correlations in similarity of structures. Interestingly, this tendency to cluster is especially strong among the most poorly optimizable structures (data not shown). This is because the structures with lower optimal foldabilities have many contacts that are common in random structures, while the structures that are highly optimizable have many rare contacts.[2] For this reason, the poorly optimizable structures have more contacts in common. The consequence of this is that there would be more tendency for the structures that are poorly optimizable to shift from one motif to another, while the highly optimizable structures would be more robust to changes in sequence. Assuming that structural consistency is favorable, this observation would support the link between optimal foldability and the robustness of the native structure to site mutations, as we suggested earlier based on other considerations.[2,6]

Evolution in the Fitness Landscape

One way of exploring the nature of the fitness landscape is to consider the properties of an evolutionary walk in that landscape. This also can serve as a simple model for the evolutionary process where sequence heterogeneity among members of the species is ignored. We have to concern ourselves with real sequences of amino acids corresponding to points in sequence and interaction space, rather than the continuous space, using the parameters derived by Miyazawa and Jernigan to generate the contact interactions as a function of the sequence.[39]

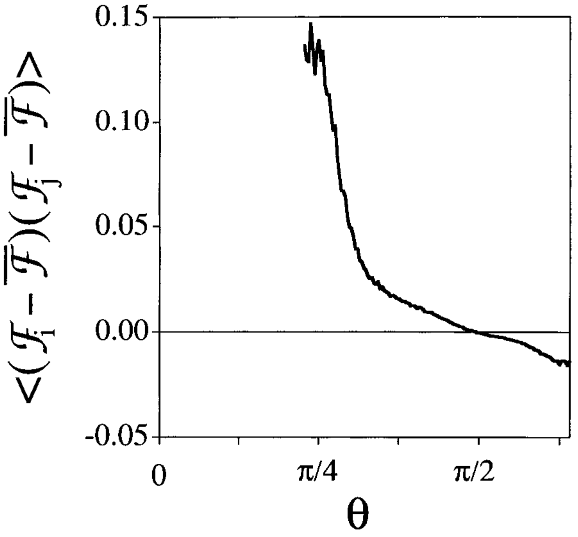

FIGURE 3 Correlation function $\langle(\mathcal{F}_i - \overline{\mathcal{F}})(\mathcal{F}_j - \overline{\mathcal{F}})\rangle_\theta$, representing the correlation of the peak heights for structures whose optima are separated by a distance θ in interaction space. $\overline{\mathcal{F}}$ represents the optimal value of the foldability averaged over all possible structures.

We consider site mutations as our elemental evolutionary step; we randomly change one amino acid into another, keeping the rest of the sequence constant. We can then calculate the new point in interaction space Γ' and the new foldability $\mathcal{F}'$, as well as the possibly new native structure $\mathcal{N}'$. Results of lattice simulations indicate that there is a minimum value of the foldability $\mathcal{F}_{\text{crit}}$ required for the protein to be able to fold sufficiently rapidly.[31,32] For modeling evolution in the presence of such selective pressure, we accept only those mutations that result in a value of $\mathcal{F}'$ larger than $\mathcal{F}_{\text{crit}}$. This allows us to include an adjustable selective pressure in our model; increasing the selective pressure corresponds to increasing $\mathcal{F}_{\text{crit}}$. A ''generation'' is considered to occur whether or not the attempted mutation is accepted. Evolution of sequences are carried out for a period of 10,000 generations on 5 different sequences selected at random, under varying degrees of selective pressure. When the initial selected sequence has a foldability less than $\mathcal{F}_{\text{crit}}$, a simple hill-climbing algorithm is used to find a similar sequence with an adequately high foldability as a starting point for the simulation. In all cases, the first 100 generations are omitted from the analysis.

One complication imposed by the use of our interaction landscape is that there is not a simple relationship between changes in the sequence and movement in the interaction space, due to the heterogeneous nature of the amino acids. As a result, a site mutation may correspond to a smaller or larger change in $\mathcal{I}$. Figure 4 shows $P(\delta_\theta)$, the distribution of step sizes δ_θ in $\mathcal{I}$ caused by site mutations during the simulated evolution, for various values of $\mathcal{F}_{\text{crit}}$. As the selective pressure increases, the distribution shifts to smaller values of δ_θ as more conservative mutations have a higher probability of being accepted than less conservative mutations. Figure 5 shows the off-diagonal elements of the mutation matrix observed for various values of $\mathcal{F}_{\text{crit}}$. At low values of $\mathcal{F}_{\text{crit}}$, all possible mutations are equally likely. As the fitness criterion increases and the mutations become increasingly conservative, the mutation matrix starts to more closely resemble the mutation matrix observed for biological proteins, also shown in Figure 5. The fact that different degrees of selective pressure effects both the absolute amino acid substitution rate as well as the relative substitution rates suggests that mutation matrices should be evolutionary-rate dependent, as was demonstrated by Koshi and Goldstein, who derived separate mutation matrices for the framework and hypervariable region of antibody molecules.[7] The general agree-

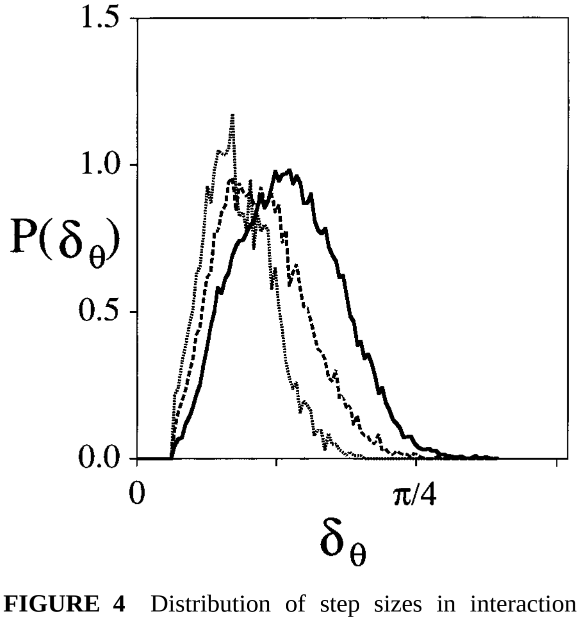

FIGURE 4 Distribution of step sizes in interaction space δ_θ corresponding to single site mutations under different degrees of selective pressure, represented by various values of $\mathcal{F}_{\text{crit}}$: weak selective pressure ($\mathcal{F}_{\text{crit}}$ = 3.0, ——), medium selective pressure ($\mathcal{F}_{\text{crit}}$ = 5.0, – – –), and strong selective pressure ($\mathcal{F}_{\text{crit}}$ = 6.0, • • •). As the selective pressure increases, the mutations become more conservative, resulting in smaller steps in the interaction space.

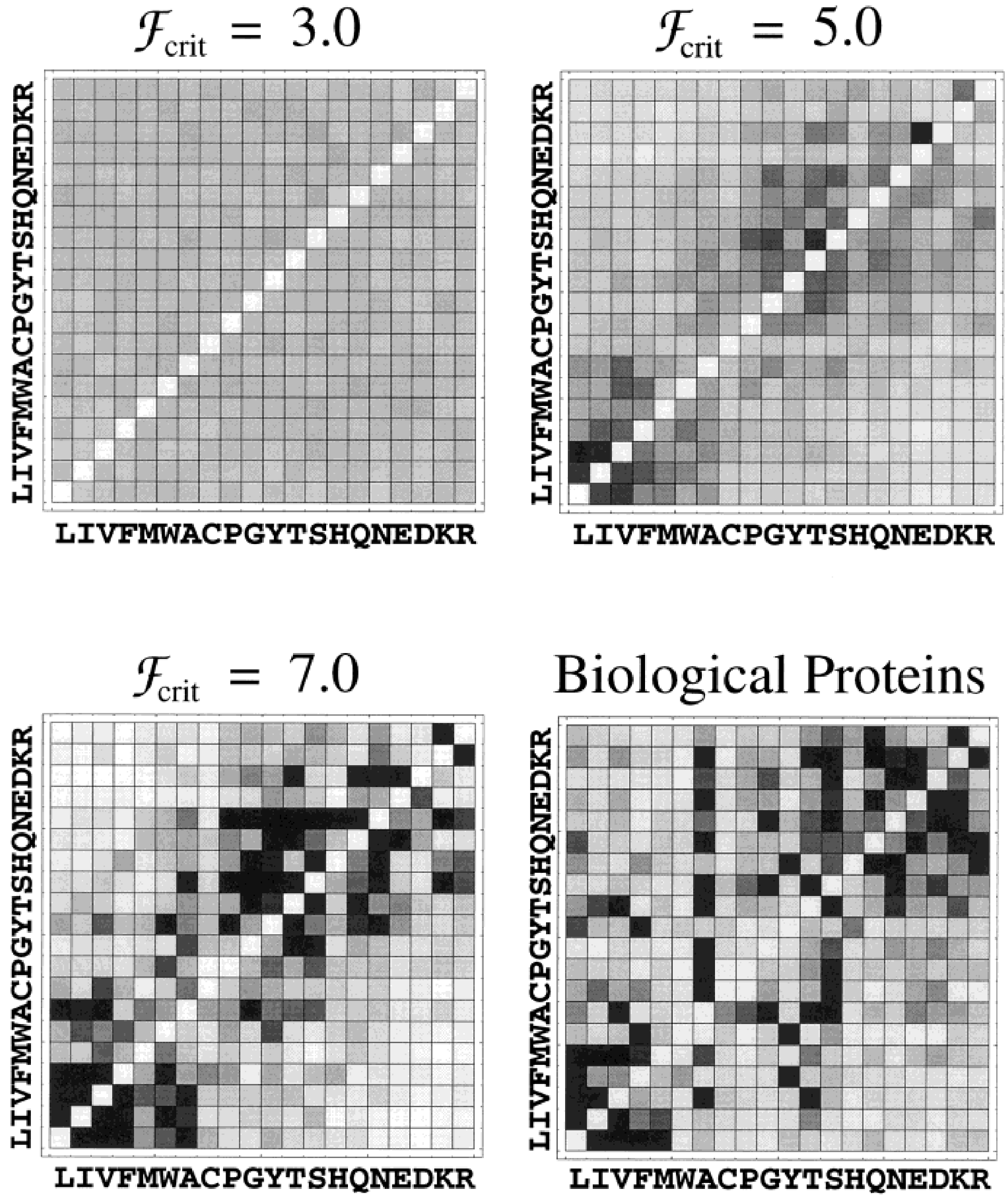

FIGURE 5 Mutation matrices corresponding to the relative probability of various mutations, as a function of selective pressure, compared with the mutation matrix derived from biological proteins.[54] The density of the off-diagonal elements represent the relative probability of one residue mutating to another in a given period of evolutionary time. Rather strong selective pressure is required in order to give the distribution of relative mutation rates seen in biological proteins.

ment between the mutation matrices calculated with our simple model and that actually observed in biological proteins is relatively good, given the overly simple form of the foldability criterion and the energy function, the absence of any criterion besides foldability in the fitness function, and the use of a

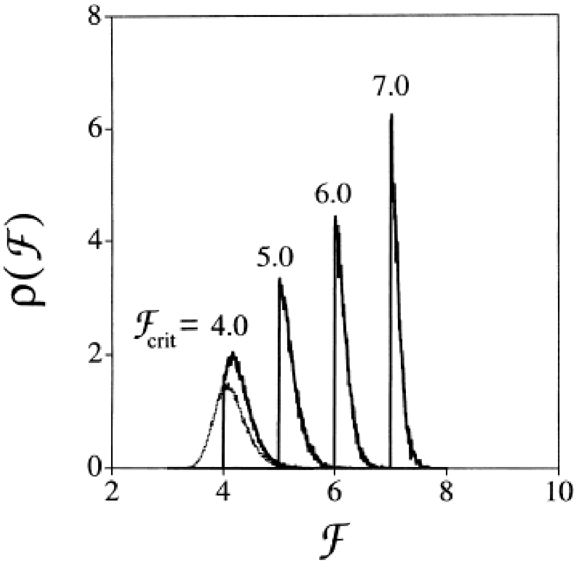

FIGURE 6 Distribution of $\mathcal{F}$ values during the evolutionary trajectories for various values of $\mathcal{F}_{\text{crit}}$ (——), compared with the distribution for trajectories under no selective pressure ($\mathcal{F}_{\text{crit}} = 0$, · · ·). Most of the trajectory is confined to regions where the foldability is only slightly greater than the critical foldability. Under such circumstances, proteins would be expected to be only marginally foldable. This tendency is increased with increasing selective pressure.

model for the protein that is highly simplified where almost all of the residues are exposed. It is interesting to note that the selective pressure necessary to produce a heterogeneity of mutation rates approaching biological proteins correspond to a relatively high selective-pressure regime.

As has not always been sufficiently appreciated, evolution does not necessarily correspond to optimization. Because of the high dimensionality of the interaction landscape, the volume of interaction space corresponding to different $\mathcal{F}$ values increases sharply with decreasing $\mathcal{F}$, so that most of the sequences with $\mathcal{F} > \mathcal{F}_{\text{crit}}$ have foldabilities only marginally larger than $\mathcal{F}_{\text{crit}}$. This is shown in Figure 6, which represents the probability that a sequence has a particular value of $\mathcal{F}$ during the evolutionary trajectory. If $\mathcal{F}_{\text{crit}}$ represents the dividing line between marginally foldable proteins and marginally unfoldable proteins, then most biological proteins would be only marginally foldable.[6] Our criterion for foldability is closely related to thermal stability. Our results suggest that proteins should be only stable enough to ensure sufficient foldability (as well as adequate protection against proteolysis and aggregation). Marginal stability has long been observed in proteins, often rationalized by the need to ensure some degree of conformal flexibility. We show that, while such arguments may be correct, it is not necessary to postulate additional reasons for marginal stability—they would naturally arise given the distribution of fitness parameters in interaction space.

As the selective pressure is increased, the evolutionary walks on the landscape became increasingly confined to ''neutral networks,'' paths through the sequence space where structure is preserved. This tendency is emphasized in Figure 7, which shows a density plot of the probability that two sequences representing different points on an evolutionary trajectory separated by a distance θ in the interaction space would have native structures separated by q in configuration space. The role of these neutral networks and their influence on the evolutionary process have been emphasized by Fontana and coworkers, who found that it is possible for the sequence of RNA to vary considerably for a fixed structure.[27,50] It is possible, given the simple Boolean code of RNA base pairing, for the sequence to change considerably with the interactions changing but little. (Exchanging C for G and T for A would likely not change the final structure, although the Hamming distance between the two structures would be a maximum.) Here we demonstrate, at least for proteins, that the neutral networks occupy large regions of *interaction* space, not just sequence space. The fact that the sequences *and* the interactions can vary within a fixed structure, in good agreement with experimental observations.[51,52]

CONCLUSION

In order to understand and model protein evolution, it is necessary to first characterize the underlying fitness landscape. We have developed a simple model for this landscape, based upon the fact that proteins need to fold. In previous work, we have shown how selective pressure can explain how the plasticity of protein sequences can exist with a tremendous robustness of the resulting structures, why certain native structures are overrepresented among biological proteins, and why certain interactions dominate the folding process.[2,3,6] That work was largely based on a static picture of evolution—that the size of the various regions in the interaction space corresponding to different native structures represent how likely that structure is to result from evolution. This would represent a time-independent equilibrium picture, where each sequence corresponding to a foldable protein was a priori equally likely.

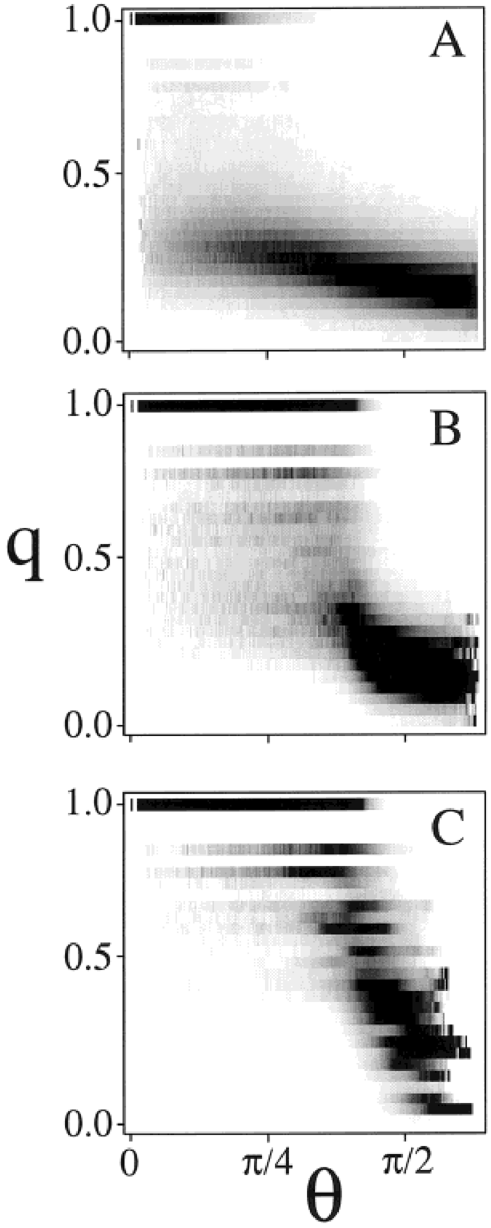

FIGURE 7 Density plot of $P(q|\theta)$, the conditional probability that two sequences during an evolutionary trajectory separated in interaction space by an amount θ will have dissimilar native structures, as measured by q, for various values of $\mathcal{F}_{\text{crit}}$: (A) $\mathcal{F}_{\text{crit}} = 3$, (B) $\mathcal{F}_{\text{crit}} = 5$, and (C) $\mathcal{F}_{\text{crit}} = 6$. As the selective pressure is increased, the trajectories become increasingly confined to neutral networks where there is relatively little change in the resulting structure, even for large changes in interactions.

Evolution, however, represents a dynamic situation, and it is not obvious how the static picture will relate to the results of dynamics. For instance, topological features of the fitness landscape, such as plateaus and ridges, can have critical importance in the evolutionary trajectories. The first step toward understanding this process is the characterization of the fitness landscape. By taking advantage of our ability to calculate a rigorously defined fitness function by using foldability, we are able not only to characterize the space, but characterize how movement on the space is affected by the degree of selective pressure.

Our studies of the topology of the landscape indicates that models that assume a random distribution of structures in interaction space may be overly simplified. Similar structures are closer in this space than would be expected at random, while dissimilar structures are actually further in space, reflecting the need for non-native structures to be destabilized in order for the protein to fold rapidly. In general, structures with similar optimal foldabilities tend to cluster together. This is especially true of proteins with rather low optima.

Our dynamical simulations largely conform to our expectations based on our static evolutionary model. As expected, the dynamic trajectories spend a disproportionate percentage of their time as protein sequences that are marginally foldable. This tendency becomes especially strong as the selective pressure is increased. If foldability represents the dominance of the folding transition over the glassy transition, this means that glassy behavior might be exhibited in proteins cooled below physiological temperatures, as is in fact observed,[53] although these results are somewhat controversial.

It has long been observed that protein structures remain fixed during evolution, while the sequence and even the stabilizing interactions change rather quickly.[51,52] It is natural to think that the selective pressure to preserve function would confine evolution to such neutral networks, explaining the observed robustness of structure. The fact that most mutations are rather conservative can explain the plasticity of the sequence. Two interesting observations arise in the evolutionary trajectories of our simple lattice proteins. The first is that conservation of structure is completely compatible with strongly varying stabilizing interactions, even at strong degrees of selective pressure. Compensatory but nonconservative mutations occur that preserve foldability. The second is that it is not necessary to postulate an additional evolutionary constraint on maintaining a fixed structure to explain why structures are fixed—it is a natural result of the nature of the fitness landscape at high degrees of selective pressure. This is even true when the movement in interaction space is large relative to the correlation length between interactions and structure. Such a natural confinement to constant structure may have played a significant role in allowing rapid evolution of proteins to fulfill different functions. The fact that the neutral networks of proteins are strongly dependent on the value of the optimal foldability,[6] combined with the fact that highly optimizable structures are further apart in interaction space, can help us understand the dominance of certain structural motifs among biological proteins in terms of this confinement.

We would like to thank Kurt Hillig and James Raines for computational assistance, and Erich Bornberg-Bauer and Michael Thompson for helpful discussions. Financial support was provided by the College of Literature, Science, and the Arts; the Program in Protein Structure and Design; the Horace H. Rackham School of Graduate Studies; NIH grant LM0577; and NSF equipment grant BIR9512955.

REFERENCES

1. Shakhnovich, E. I. & Gutin, A. M. (1990) *Nature (London)* **346,** 773–775.
2. Govindarajan, S. & Goldstein, R. A. (1995) *Biopolymers* **36,** 43–51.
3. Govindarajan, S. & Goldstein, R. A. (1995) *Proteins* **22,** 413–418.
4. Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995) *Protein Sci.* **4,** 561–602.
5. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 1282–1286.
6. Govindarajan, S. & Goldstein, R. A. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 3341–3345.
7. Koshi, J. M. & Goldstein, R. A. (1996) *Proteins,* **27,** 336–344.
8. Li, H., Helling, R., Tang, C. & Wingreen, N. (1996) *Science* **273,** 666–669.
9. Eisenhaber, F., Persson, P. & Argos, P. (1995) *Crit. Rev. Bioch. Mol. Biol.* **30,** 1–94.
10. Thompson, M. J. & Goldstein, R. A. (1996) *Proteins* **25,** 28–37.
11. Benner, S. A., Chelvanayagam, G. & Turcotte, M. (1997) *Chem. Rev.,* in press.
12. Pauling, L. & Zuckerkandl, E. (1963) *Acta Chem. Scand.* **17,** S9–S16.
13. Yang, Z., Kumar, S. & Nei, M. (1995) *Genetics* **141,** 1641–1650.
14. Jermann, T. M., Optiz, J. G., Stackhouse, J. & Benner, S. A. (1995) *Nature (London)* **374,** 57–59.
15. Koshi, J. M. & Goldstein, R. A. (1996) *J. Mol. Evol.* **42,** 413–420.
16. Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F. & Wilson, A. C. (1990) *Nature (London)* **345,** 86–88.
17. Stackhouse, J., Presnell, S. R., McGeehan, G. M., Nambiar, K. P. & Benner, S. A. (1990) *FEBS Lett.* **262,** 104–106.
18. Shih, P., Malcolm, B. A., Rosenberg, S., Kirsch, J. F. & Wilson, A. C. (1993) *Methods Enzymol.* **224,** 576–590.
19. Wright, S. (1932) *Int. Proceed. Sixth Intl. Cong. Genet.* **1,** 356–366.
20. Macken, C. A. & Perelson, A. S. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 6191–6195.
21. Derrida, B. & Peliti, L. (1991) *Bull. Math. Biol.* **53,** 355–382.
22. Bak, P., Flyvbjerg, H. & Lautrup, B. (1992) *Phys. Rev. A* **46,** 6724–6730.
23. Kauffman, S. A. (1993) *The Origins of Order,* Oxford University Press, New York.
24. Lipman, D. J. & Wilbur, W. J. (1991) *Proc. R. Soc. Lond. (Biol.)* **245,** 7–11.
25. Fontana, W., Stadler, P. F., Tarazona, P., Weinberger, E. D. & Schuster, P. (1993) *Phys. Rev. E* **47,** 2083–2099.
26. Fontana, W., Konings, D. A. M., Stadler, P. F. & Schuster, P. (1993) *Biopolymers* **33,** 1389–1404.
27. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. (1994) *Proc. R. Soc. Lond. B* **255,** 279–284.
28. Renner, A. & Bornberg-Bauer, E. (1996) in *Pacific Symposium on Biocomputing '97,* Altman, R. B., Dunker, A. K., Hunter, L. & Klein, T. E., Eds., World Scientific, Singapore, pp. 361–372.
29. Levinthal, C. (1969) in *Mossbauer Spectroscopy in Biological Systems,* Debrunner, P., Tsibris, J. C. M. & Munck, E., Eds., University of Illinois Press, Urbana, pp. 22–24.
30. Branden, C. & Tooze, J. (1991) *Introduction to Protein Structure,* Garland Publishing, New York.
31. Šali, A., Shakhnovich, E. I. & Karplus, M. J. (1994) *J. Mol. Biol.* **235,** 1614–1636.
32. Šali, A., Shakhnovich, E. I. & Karplus, M. J. (1994) *Nature (London)* **369,** 248–251.
33. Go, N. (1983) *Ann. Rev. Biophys. Bioeng.* **12,** 183–210.
34. Karplus, M. & Shakhnovich, E. (1992) in *Protein Folding,* Creighton, T., Ed., W. H. Freeman, New York, pp. 127–195.
35. Karplus, M. & Šali, A. (1995) *Curr. Biol.* **5,** 58–73.
36. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995) *Proteins* **21,** 167–195.

37. Wolynes, P. G., Luthey-Schulten, Z. & Onuchic, J. N. (1996) *Curr. Biol.* **3,** 425–432.
38. Maynard-Smith, J. (1970) *Nature (London)* **225,** 563–564.
39. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18,** 534–552.
40. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7524–7528.
41. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 4918–4922.
42. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 9029–9033.
43. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1993) in *Proceedings of the 26th Annual Hawaii International Conference on System Sciences*, Vol. 1, Mudge, T. N., Milutinovic, V. & Hunter, L., Eds., IEEE Computer Society Press, Los Alamitos, pp. 699–707.
44. Shakhnovich, E. I. & Gutin, A. M. (1993) *Protein Eng.* **6,** 793–800.
45. Shakhnovich, E. I. & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7195–7199.
46. Shakhnovich, E. I. (1994) *Phys. Rev. Lett.* **72,** 3907–3910.
47. Shakhnovich, E. I. (1994) in *Protein Structure by Distance Analysis,* Bohr, H. & Brunak, S. Eds., IOS Press, Amsterdam, pp. 201–212.
48. Chan, H. S. & Dill, K. A. (1994) *J. Chem. Phys.* **100,** 9238–9257.
49. Chiu, T. L. & Goldstein, R. A. (1997) *J. Chem. Phys.,* in press.
50. Schuster, P. & Stadler, P. F. (1994) *Comput. Chem.* **3,** 295–324.
51. Aronson, H. E. G., Royer, W. E., Jr. & Hendrickson, W. A. (1994) *Protein Sci.* **3,** 1706–1711.
52. Laurents, D. V., Subbiah, S. & Levitt, M. (1994) *Protein Sci.* **3,** 1938–1944.
53. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991) *Science* **254,** 1598–1603.
54. Koshi, J. M. & Goldstein, R. A. (1995) *Protein Eng.* **8,** 641–645.