PROTEINS: Structure, Function, and Genetics 34:113–124 (1999)

Effect of Alphabet Size and Foldability Requirements on Protein Structure Designability

Nicolas E.G. Buchler[1] and Richard A. Goldstein[1,2]*
[1]*Biophysics Research Division, Ann Arbor, Michigan*
[2]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan*

ABSTRACT **A number of investigators have addressed the issue of why certain protein structures are especially common by considering structure *designability*, defined as the number of sequences that would successfully fold into any particular native structure. One such approach, based on *foldability*, suggested that structures could be classified according to their maximum possible foldability and that this optimal foldability would be highly correlated with structure designability. Other approaches have focused on computing the designability of lattice proteins written with reduced two-letter amino acid alphabets. These different approaches suggested contrasting characteristics of the most designable structures. This report compares the designability of lattice proteins over a wide range of amino acid alphabets and foldability requirements. While all alphabets have a wide distribution of protein designabilities, the form of the distribution depends on how protein "viability" is defined. Furthermore, under increasing foldability requirements, the change in designabilities for all alphabets are in good agreement with the previous conclusions of the foldability approach. Most importantly, it was noticed that those structures that were highly designable for the two-letter amino acid alphabets are not especially designable with higher-letter alphabets. Proteins 1999;34:113–124.**

Key words: protein folding; designability; foldability; alphabet; lattice models; spin-glass theory

INTRODUCTION

It has been noted by a number of investigators that certain structures are more commonly observed among proteins than others.[1–4] A variety of models have been developed to explain this phenomenon by considering protein structure *designability,* that is, the number of sequences that would successfully form one structure or another. Highly designable structures would be more likely to have been found through the process of evolution, as well as be more robust to random mutational changes.[5,6] These structures might also represent attractive targets for protein design.[7,8] Several designability approaches have focused on protein energetics and kinetic issues by considering the number of sequences that would be both foldable and thermodynamically stable. For instance, Finkelstein and colleagues[9–11] used energetic arguments to explain why particular local motifs might be easier to stabilize and thus more common in the protein database. Govindarajan and Goldstein[12] developed a model for sequence *foldability,* a thermodynamic measure characterizing how amenable the free-energy landscape is to successful protein folding, and showed that different structures would have different maximum possible foldabilities. It was demonstrated that those structures with the largest optimal foldabilities would be the most designable, as there would be many possible sequences far from the optimum, and yet still be adequately foldable.[12,13] Conversely, a protein structure that had a low optimal foldability would be poorly designable, as these structures could only be formed by the relatively rare sequences with close-to-optimal interactions.

Other groups of investigators have used two- and three-dimensional lattice models to explicitly enumerate those sequences that have a nondegenerate ground-state conformation in one structure or another.[14–17] In order to reduce the total number of possible sequences, they constructed their sequences with a two-letter amino acid alphabet with all residues belonging to one of two types, either hydrophobic or polar. The presence of a nondegenerate ground state was assumed to be adequate to ensure that the protein could successfully fold into that structure; those sequences with ground-state degeneracy were considered to represent "unviable" proteins. In agreement with the theoretical and analytical approaches described above, these groups observed that a greater proportion of viable sequences folded into some structures compared with others; that is, some structures were more designable than others.

There are a number of reasons to question the biological relevance of these latter models. First, it is not clear how dependent the results are on the use of a reduced two-letter amino acid alphabet. A sequence is considered viable if and only if there is a nondegenerate ground state. The effect of sequence degeneracies is, however, strongly alphabet dependent. For instance, Shakhnovich noted that when the entropy of the amino acid alphabet drops below

Grant sponsor: National Institutes of Health; Grant numbers: LM05770 and GM08270; Grant sponsor: National Science Foundation; Grant number: BIR9512955.
*Correspondence to: Richard A. Goldstein, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055. E-mail: richardg@umich.edu
Received 14 May 1998; Accepted 31 August 1998

that of the conformational entropy per residue, as in the case of the two-letter code, a sizable number of degenerate sequences exist.[18,19] By contrast, exact ground-state degeneracies are less expected in protein models that use larger alphabets, and should be virtually nonexistent in real proteins, given the continuous nature of the interaction strengths. Discarding large numbers of degenerate sequences might skew designability results, making conclusions based on two-letter codes inapplicable to natural proteins with a larger alphabet entropy.

There are additional problems in considering the absence of degeneracies as a necessary and sufficient condition for protein viability. While the presence of one dominant state is generally a requirement for folding, naturally occurring proteins often exist in a number of conformational substates with similar free energies.[20] Such proteins would be considered nonviable according to these latter models. Furthermore, as has been demonstrated by lattice simulations, lack of an exact ground-state degeneracy is not adequate to ensure that a protein sequence can actually fold or be stable in its conformation of lowest free energy.[21–24] The concept of foldability was introduced to address such issues. On the basis of these questions, it is important to examine whether the designability results obtained with these models can be extrapolated to natural proteins made up of 20 amino acids and where foldability and stability are important.

This report addresses some of these questions by looking at the designability of lattice proteins both as a function of alphabet size and required foldability. We find that the observation that various structures have a disproportional number of foldable sequences is a universal aspect for all alphabets. The form of the distribution of designabilities, as well as which structures are most designable, however, are strongly dependent on the size of the alphabet and on how sequence viability is defined. In particular, those structures that are highly designable for the two-amino acid alphabets are not the most designable structures for the larger alphabets. Overall, our results demonstrate the difficulty in extrapolating the results obtained with the reduced amino acid models to naturally occurring proteins.

MODELS AND METHODS

Our model consists of a 25-residue protein chain confined to a maximally compact 5 × 5 two-dimensional lattice, in which each residue is assigned to a lattice point. Aside from computational feasibility, our particular choice of lattice model is motivated by consideration of realistic solvation ratios of residues and the level of compactness found in globular-protein native states; two-dimensional lattice proteins have a more natural ratio of solvated versus buried residues as compared with three-dimensional lattice proteins of the same size. Our consideration of only compact conformations stems from previous lattice protein results demonstrating that hydrophobic collapse favors native states having nearly maximal compaction.[14] In addition, other recent work has shown that the interactions favoring compaction should be quite strong in optimally folding proteins.[25] While 25 residues is rather short for natural proteins, a *corresponding state* analysis suggests that these lattice models may be appropriate for natural proteins of longer length, with a number of amino acids in the protein represented by each lattice-model residue.[26]

For the maximally compact 5 × 5 two-dimensional lattice protein, there are a total of 1,081 possible self-avoiding walks on this lattice, excluding rotations and reflections, which represents the 1,081 different possible conformations for the protein chain. The energy function for a sequence S in a particular conformation k is given by a simple pair-contact form:

$$E_S^k = \sum_{i<j} \gamma_{ij}^S \Delta_{ij}^k = \boldsymbol{\gamma}_{\mathbf{ij}}^{\mathbf{S}} \cdot \boldsymbol{\Delta}_{\mathbf{ij}}^{\mathbf{k}} \qquad (1)$$

where $\boldsymbol{\gamma}_{\mathbf{ij}}^{\mathbf{S}}$ specifies the residue contact energies of all possible contacts that can be formed for a particular sequence S, and Δ_{ij}^k is equal to 1 if nonsequential residues i and j are on adjacent lattice sites in conformation k and zero otherwise. The structure vector $\boldsymbol{\Delta}_{\mathbf{ij}}^{\mathbf{k}}$ is unique for each conformation k, as no two conformations have identical pair contacts. Owing to the nature of the lattice, the only contacts possible are between odd and even residues, making the total number of possible contacts equal to 132. Of these 132 possible contacts, a subset of only 16 contacts are actually made in each compact conformation.

The free-energy landscape of the protein is completely determined by the interaction vector $\boldsymbol{\gamma}_{\mathbf{ij}}^{\mathbf{S}}$. Each contact energy $\gamma_{ij}^S = \gamma(\mathscr{A}_i^S, \mathscr{A}_j^S)$ is a function of the amino acids $\mathscr{A}_i^S$ and $\mathscr{A}_j^S$, at sequence positions i and j as specified in the definition of the amino acid alphabet. It is the size of the alphabet, the details of its amino acid pair-contact energies, and the requirements for sequence viability that determines the relative representation of structures in the sequence database. This report explores the alphabet dependence of structure designability by comparing results achieved with six different amino acid alphabets: the HP, Li, and π (PI) two-letter alphabets, the hHYX four-letter alphabet, the Miyazawa-Jernigan (MJ) 20-letter alphabet, and the independent interaction model (IIM) infinite amino acid alphabet, where each possible contact potential is independent from other possible contacts. This latter model is achieved by randomly drawing all 132 γ_{ij}^S interactions from a gaussian distribution. The energetic details of these various alphabets are summarized in Table I.

For each alphabet with its specific amino acid pair-contact potential and a given set of possible conformations, we synthesize an ensemble of sequences and generate their corresponding energy landscapes. It is assumed that the native state of each sequence is its lowest-energy conformation, an assumption known as the *thermodynamic hypothesis*.[32,33] We first adopt the standard approach of Lipman, Li, Bornberg-Bauer, and their respective co-workers and presume that a sequence would fold if and only if the lowest-energy structure is nondegenerate. As mentioned in the introduction and shown in Table II,

TABLE I. Energy Parameters for the Various Alphabets†

HP	H	P
H	−1.0	0.0
P	0.0	0.0

Li	H	P
H	−2.3	−1.0
P	−1.0	0.0

π	H	P
H	−3.14	−1.00
P	−1.00	0.00

hHYX	h	H	Y	X
h	−2.0	−4.0	−1.0	2.0
H	−4.0	−3.0	−1.0	0.0
Y	−1.0	−1.0	0.0	2.0
X	2.0	0.0	2.0	0.0

†Contact energies $\gamma(\mathscr{A}_i, \mathscr{A}_j)$ for the two- and four-letter alphabets. Dill et al.[27] and Lipman et al.[15] have used the HP two-letter alphabet, which mimics the effect of hydrophobic collapse in its tendency to bury hydrophobes in the core and segregate polar residues to the protein surface. A refinement of the HP model, the Li two-letter alphabet is based on dominant eigenvalue analysis of the Miyazawa-Jernigan statistical potential and was constructed so that two like-contacts (HH and PP) are energetically favored over two unlike contacts (HP and HP).[16,28] We also constructed a π (PI) alphabet, which represents a compromise between the HP alphabet and the Li alphabet. The use of a transcendental number in the potential prohibits the possibility of "accidental degeneracies" (i.e., structures with the same energy without identical numbers of the same types of contacts) for any size protein in any lattice. The hHYX four-letter alphabet was taken from the Crippen empirical potential,[29] as modified by Bornberg-Bauer.[30] The 20-letter MJ alphabet, not shown here, was based on the Miyazawa-Jernigan statistical potential.[31]

TABLE II. Statistics Describing the Ensemble of Sequences Constructed Using the Various Amino Acid Alphabets†

Alphabet	Degeneracy (%)	$\langle \mathscr{F} \rangle_{\text{non-deg}}$	$\sigma_{\mathscr{F}_{\text{non-deg}}}$	$\langle \mathscr{F} \rangle_{\text{deg}}$	$\sigma_{\mathscr{F}_{\text{deg}}}$
HP two-letter	81.58	3.4931	0.4180	2.9237	0.4221
Li two-letter	63.09	2.9831	0.3058	2.7539	0.3056
PI two-letter	61.65	3.1677	0.3619	2.8917	0.3485
hHYX four-letter	41.63	3.1385	0.3474	2.8836	0.3125
MJ 20-letter	4.39	3.1341	0.3547	2.9667	0.3248
IIM infinite-letter	0.01	3.1217	0.3373	2.8943	0.2878

†For each alphabet, we list the percentage of the constructed sequences that had degenerate ground states, the average foldability of the sequences ($\langle \mathscr{F} \rangle$), and the standard deviation of the foldabilities ($\sigma_{\mathscr{F}}$), both for sequences with nondegenerate (non-deg) and degenerate (deg) ground states. Conformations were considered degenerate if their energies differed by less than 10^{-4}.

smaller alphabets have a higher percentage of sequences with degenerate ground states compared with larger alphabets.

The approach described above ignores the energetic and kinetic considerations at the heart of recent protein folding models. For this reason, we also consider that a nondegenerate ground state does not guarantee a protein's ability to fold into its native state. There has been extensive theoretical, computational, and experimental work elucidating the relationship between the thermodynamic properties of proteins that characterize their energy landscape and the ability of a protein to fold. For example, using concepts borrowed from the physics of spin glasses, Bryngelson and Wolynes[21,22] consider that two thermodynamic transitions are possible in a protein: one to the folded state at a temperature T_f, and the other to a glassy state at a temperature T_g. For temperatures below T_g the density of conformational states suffers an entropy crisis, folding kinetics become slow, and it becomes difficult for the protein to transit from misfolded local minima to other stable states. T_f defines the temperature at which the global, energy minimum is deep enough to be preferentially populated over other possible conformations and stable with respect to thermal fluctuations. Foldability requires a temperature regime that is both adequately below T_f for the folded state to be stable yet sufficiently above T_g for the folded state to be accessible. This demands that the folding temperature T_f be substantially higher than that of the glassy transition temperature T_g. Thus, the ratio T_f/T_g is a measure of how easily a given sequence can fold into its native structure by escaping misfolded, metastable kinetic traps and freely exploring the energy landscape, while also having a stable native fold at its global energy minimum.

Using the random energy model (REM) to describe the landscape, one can analytically relate T_f/T_g to the protein sequence foldability $\mathscr{F} \equiv \Delta/\Gamma$, where Δ measures the depth of the free energy of the native state with respect to the average of the ensemble of random states, and Γ is the standard deviation of free energies of the random ensemble.[34,35] The REM is one of the simplest possible approaches and ignores all correlations in the free-energy landscape. Maximizing the foldability, however, results in the stabilization of the native state, a reduction in the depth of metastable traps, a destabilization of competing dissimilar conformations, and the stabilization of conformations similar to the native state producing the funnel-like energy landscapes central to a number of more recent models.[36] This foldability approach was supported by Monte Carlo simulations with lattice proteins, which showed that foldable proteins were characterized by a large value of $\mathscr{F}$.[23,37,38] We assume, based on the results of these simulations, that a sequence should be foldable if its foldability $\mathscr{F}$ exceeds a critical value $\mathscr{F}_{\text{crit}}$. More sophisticated models have been developed by a number of other researchers.[39–48]

For any sequence with a specified alphabet and associated matrix of amino acid pair-contact potentials $\gamma(\mathscr{A}_i, \mathscr{A}_j)$ we can calculate the energy of all possible compact conformations, find the native conformation, and measure the sequence foldability $\mathscr{F}$. We can determine the sequence viability by requiring this foldability to be larger than the critical foldability $\mathscr{F}_{\text{crit}}$. For the two-letter alphabets, we did an exhaustive enumeration of all 2^{25}–33 million sequences, whereas the other alphabets were examined by randomly sampling ~20 million sequences. In all cases, we verified that the number of sequences is large enough for the averages to be well defined. Table II lists the sequence foldability statistics for all the alphabets, both for sequences with nondegenerate and degenerate ground states. With this extensive set of data, we can examine the

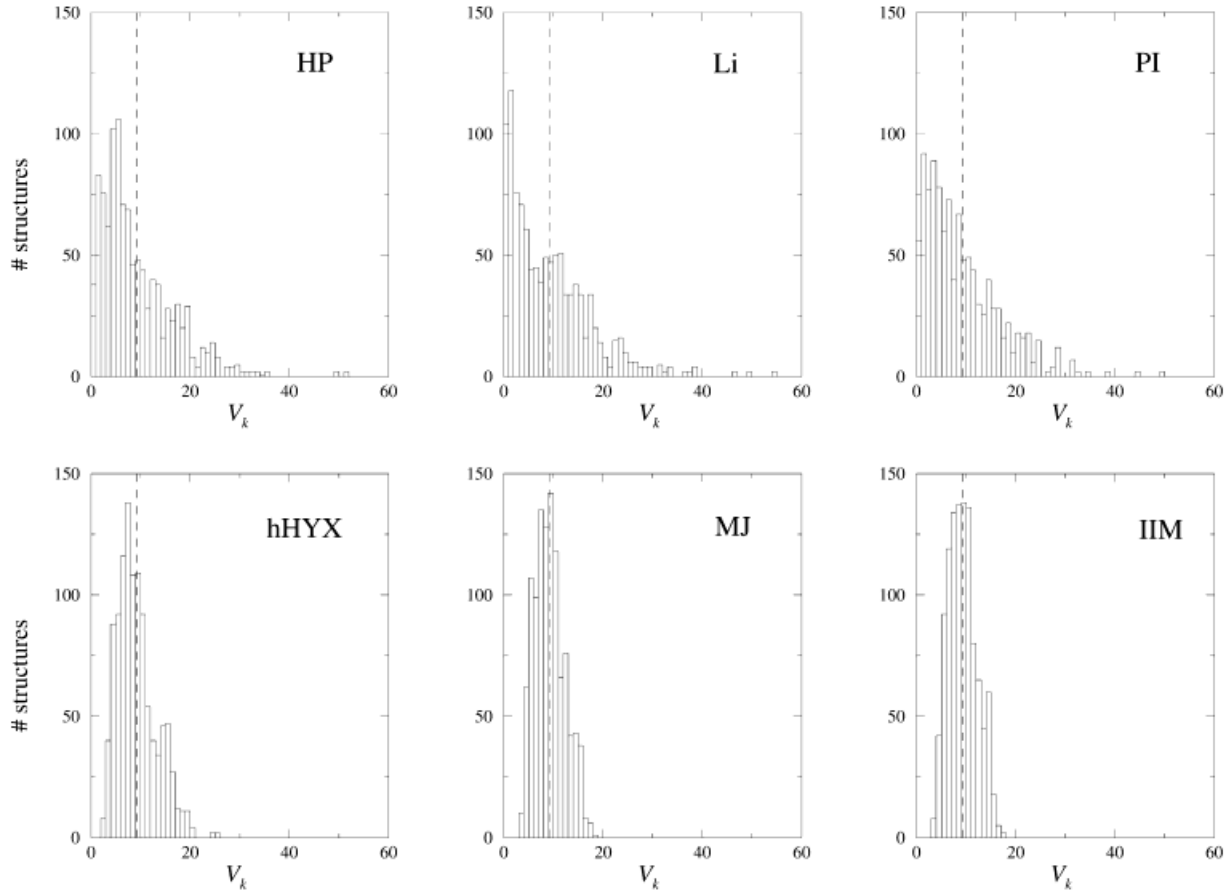

Fig. 1. Histogram of the designability distributions V_k for the various alphabets. The V_k for all alphabets have been normalized to 10,000. The two-letter alphabets have a broad exponential distribution, whereas the higher-letter alphabets have a narrower, gaussian-like distribution. The dashed line represents the expected value of V_k if all structures were equally designable, as assumed by Wang.[49]

resulting distribution of sequences over native state conformations and see how structure designability is affected by both alphabet size and foldability requirements.

RESULTS

The designability, defined as the fraction of viable sequences that have conformation k as their unique native state (i.e., the volume of foldable sequence space folding into native structure k) is written as V_k. For each alphabet, there are 1,081 possible native states and, thus, 1,081 different V_k values. The V_k for each respective alphabet were normalized so that the sum over all structures equals 10,000. Figure 1 shows a histogram of the designabilities for different alphabets. The observation that some structures are more designable than others seems to be a universal feature of all alphabets. This highlights the fact that structural designability cannot be described by a uniform probability, where all motifs are equally designable in the sequence database, as suggested by Wang,[49] and fits recent statistical studies of the distribution of various fold types among proteins of known structure.[50]

The two-letter alphabets all have a broad, exponential distribution of designabilities where there are a large number of poorly designable structures with small V_k and a very few "super-designable" structures. This is similar to the designability distributions found by Lipman and Wilbur,[15] Li et al.,[16] Bornberg-Bauer,[17] Renner and Bornberg-Bauer,[53] and even the (four-letter alphabet) RNA structural designability results of Schuster and co-workers.[51,52] This form of the distribution, however, does not correspond to the situation for larger alphabets where the designability distribution becomes narrower and more gaussian in form. These same data are portrayed in a Zipf's law plot in Figure 2, which shows the designability V_k plotted against the relative rank of the designability. Again, it is clear how similar the distributions are for the various two-letter alphabets and how different they are relative to the higher-letter alphabets.

Tang and co-workers[16] noticed that the most designable lattice structures for the two-letter alphabet tended to be highly symmetric and reasoned that this might explain the high degree of symmetry observed among naturally occurring proteins. This connection between protein symmetry and design, mirrors the results of earlier HP lattice protein work by Yue and Dill,[54] where structures with tertiary symmetry such as four-helix bundles, α/β-barrels, and

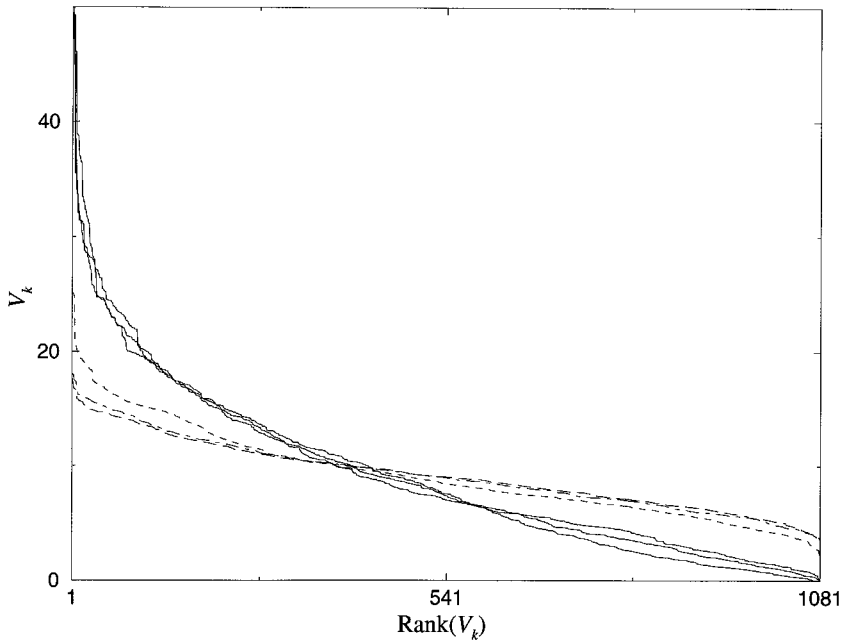

Fig. 2. Designability distributions for the HP, Li, and PI two-letter alphabets (—), four-letter hHYX alphabet (----), 20-letter MJ alphabet (– - – -), and infinite-letter IIM alphabet (– – –), displayed in a Zipf's plot. V_k values are plotted against the index of the structure ranked by V_k value (i.e., the structure with the largest number of foldable sequences has a rank of 1). All two-letter alphabets have similar designability distributions, whereas the higher alphabets tend to cluster around a different distribution.

parallel β-helices were found to be globally optimal structures of HP sequences with minimal degeneracy. At the other end of the alphabet spectrum, Govindarajan and Goldstein[12,13] concluded, on the basis of their foldability model with the infinite-letter IIM alphabet, that highly designable structures would have many long-range contacts. In light of these structural conclusions for both alphabet extremes, one should ask how much does the relative ordering of structures ranked by their designabilities depends on the alphabet size. Are the structures that are highly designable for the more realistic MJ 20-letter alphabet closer to those of the two-letter alphabets or the infinite-letter IIM alphabet? Figure 3 shows the correlation in the relative values of V_k for the same structure between different alphabets. Correlation coefficients of these data are presented in Table III. As shown, the values of V_k are highly correlated across the various two-letter alphabets, indicating that it is the size of the alphabet, rather than the specific details of the amino acid pair-contact potentials, that is important in determining the designability of particular structures. Similarly, structure designabilities are highly correlated between the 20-letter MJ and infinite-letter IIM alphabet, supporting the structural conclusions of the foldability model concerning the presence of long-range contacts in highly designable proteins. By contrast, however, the relative designabilities of particular structures between the two-letter alphabets and the higher 20-letter and infinite-letter IIM alphabets are negligibly correlated with each other; the highly designable structures for the two-letter alphabets are not overly designable for the larger alphabets, and vice-versa. The four-letter hHXY alphabet represents an intermediate case with some degree of correlation with both smaller and larger alphabets. Thus, it appears that the structural designability results for two-letter alphabets may contain artifacts because of the small alphabet size. This indicates that caution should be exercised when extrapolating correlations between high designability and particular structural features as observed with two-letter codes to natural proteins.

One of the most striking characteristics of the smaller-letter alphabets is the relative abundance of sequences with degenerate ground states, an abundance that disappears for the larger alphabets. As mentioned in the introduction, this is of suspect biological significance as exact degeneracies are unlikely to be observed in real proteins. How much are the differences in relative designabilities due to the large number of degenerate sequences with the two-letter codes? In order to address this question, we relaxed the requirement for sequences to have nondegenerate ground states. We re-examined the exhaustive sequence enumeration for the Li two-letter alphabet, where we now considered all sequences with native-state degeneracy <108 (10%) to be viable. If a sequence has a ground-state degeneracy of m, we assign to each degenerate native structure a sequence volume of $1/m$. The results of these calculations are summarized in Figure 4. This new, "relaxed" Li* two-letter alphabet now exhibits a gaussian, rather than a broad, exponential designability distribution, similar to what was observed for the higher-letter (and less degeneracy-prone) alphabets. Interestingly, as shown in Figure 5, including sequences with degenerate ground states did not change the identity of the most designable structures: the Li* two-letter alphabet is still very much a two-letter alphabet. Thus, whereas the *form* of the two-letter alphabet designability distribution seems to depend on the requirement of ground state nondegeneracy, the relative ordering of which structures are most designable depends more directly on the size of the alphabet.

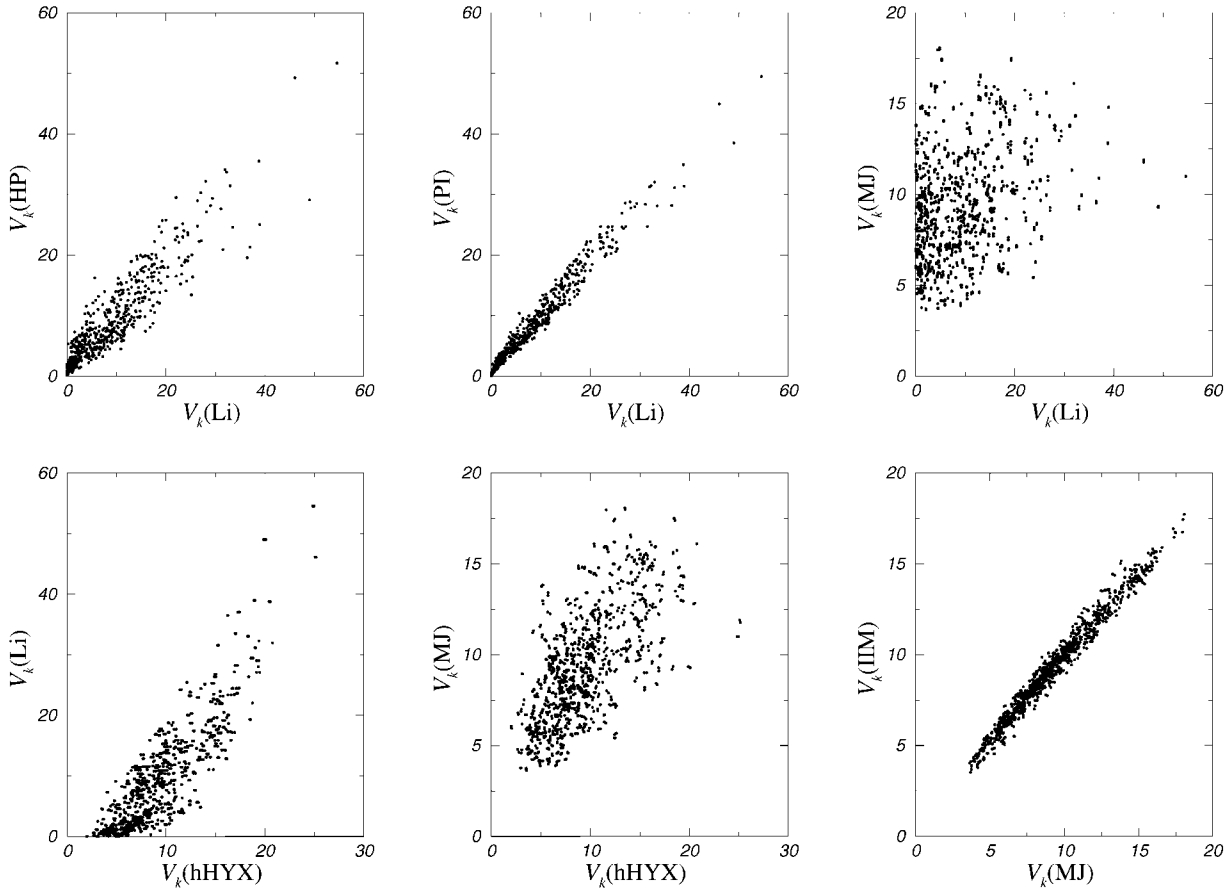

Fig. 3. Scatter plots displaying the V_k values of identical structures as computed for different pairs of alphabets. As shown, the relative V_k values are highly correlated between the various two-letter alphabets as well as between the 20-letter MJ alphabet and the infinite-letter IIM alphabet. There is little correlation, however, between the V_k values for the two-letter and the 20-letter alphabets. The four-letter hHYX alphabet represents an intermediate case between the two-letter and higher-letter alphabets, with V_k values correlated to both.

TABLE III. Correlation Coefficients Comparing the Relative V_k of Corresponding Structures for Different Alphabets[†]

Alphabet	Li	hHYX	IIM
HP	0.9131	0.9510	0.4602
Li	—	0.8609	0.2654
PI	0.9833	0.9229	0.3533
hHYX	0.8609	—	0.6136
MJ	0.3236	0.6764	0.9877
IIM	0.2654	0.6136	—

[†]The various two-letter alphabets have highly correlated V_k values, as do the two largest alphabets. There is little correlation in the V_k values between the structures of the two-letter and the larger alphabets. The four-letter hHYX represents an intermediate case, correlated with both the two-letter and higher-letter alphabets.

As discussed in the introduction, the existence of a nondegenerate ground-state does not necessarily guarantee a viable, foldable protein sequence. The question arises how the need to be adequately foldable, that is, to have foldability $\mathscr{F}$ greater than some critical foldability $\mathscr{F}_{\text{crit}}$, affects protein structure designability. One of the predictions of the foldability model was that at higher $\mathscr{F}_{\text{crit}}$ those structures which were already highly designable would become even more overrepresented, while poorly designable structures would become underrepresented and eventually impossible. This effect is demonstrated both in Figure 6, which shows Zipf's plots of the designability distributions of the six different alphabets under increasing foldability pressure where sequences had to both be nondegenerate and have a foldability $\mathscr{F} > \mathscr{F}_{\text{crit}}$ to be considered viable, and in Figure 7, which shows the resulting effect on the designability of specific structures. As shown, highly designable structures are relatively overrepresented (larger V_k) for increasing $\mathscr{F}_{\text{crit}}$, whereas poorly designable structures become underrepresented and eventually "extinct" and can no longer be formed by any viable sequence. In particular, as the critical foldability increases, the form of the designability distribution of higher-letter alphabets begins to resemble the exponential distribution characteristic of the simpler, two-letter alphabets.

While this prediction of the foldability model holds across all alphabets, the extinction of the less designable structures with increasing $\mathscr{F}_{\text{crit}}$ is especially pronounced for

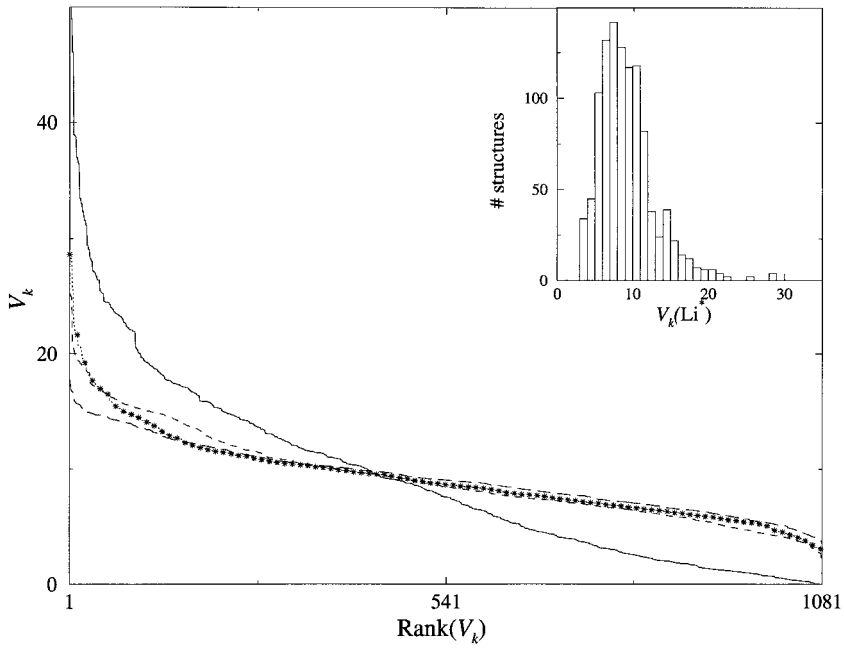

Fig. 4. Zipf's plot showing the distribution of designabilities with the Li* alphabet where ground-state degeneracies are allowed (*–*) compared with the original Li alphabet (—), hHYX alphabet (----), and IIM alphabet (– – –). Relaxing the non-degeneracy requirement causes the observed distribution of the (Li*) two-letter alphabet to more closely match that of larger alphabets. Insert: Histogram of the Li* designability distribution. The shape of this distribution closely matches the hHYX four-letter alphabet shown in Figure 1.

the Li and PI two-letter alphabets compared to the more robust HP two-letter alphabet. It is interesting that, while the form of the distribution of designabilities for the various two-letter alphabets was originally quite similar, the particular effect of a foldability requirement on individual two-letter alphabets is highly dependent on the specifics of the amino acid pair-contact potentials. This can be explained based on the statistics of sequence foldabilities $\mathscr{F}$ for the different two-letter alphabets: the nondegenerate sequences for the HP alphabet have higher $\mathscr{F}$ on average, whereas the nondegenerate sequences for the Li alphabet have lower average $\mathscr{F}$ (Table II). The reason for this foldability difference between the two-letter alphabets resides in the discrete nature of the two-letter energy landscape and the specifics of the amino acid pair-contact potentials. As an illustration, for an HP sequence where every conformation k has a total of 16 contacts and each contact can only have two possible pair-contact energies (-1 or 0), the conformational energy density of any HP sequence is distributed over only 17 possible total energy values $E = \{-16, -15, \ldots, -1, 0\}$. Because of the selection of only the few sequences with nondegenerate ground states in this sparse energy landscape, the foldability is selectively sampled and poorly averaged over the interaction space. As shown in Table II, better averaging over the interaction space is achieved when one uses higher-letter alphabets. Thus, it appears that while the two-letter alphabet foldability statistics are highly sensitive to the energetic details of the amino acid pair-contact potentials, these statistics become more robust when one increases the size of the alphabet.[18] We speculate that the details of foldability statistics, the levels of degeneracy, and the differences in designability for these small alphabets might be due to the nonisotropic sampling of interaction space by two-letter alphabet sequences and the large differences in interaction space between nearly identical sequences.

As discussed earlier, removing the need for the native state to be nondegenerate changed the form of the distribution of designabilities for the two-letter alphabets so as to more closely match that of the large alphabet designabilities. Yet, the ordering of relative designabilities of different structures remained mostly unchanged. Likewise, as seen in Figure 6, enforcing the need for a sufficient foldability causes the distribution of designabilities of the larger alphabets to more closely resemble that of the two-letter alphabets. So, what happens to the relative ordering of which structures are more designable under this more stringent criterion? Will the structures that are highly designable in the two-letter alphabets and four-letter alphabets emerge to be dominant when one increases the foldability pressure on the IIM infinite-letter alphabet or MJ 20-letter alphabet? The answer to this question is displayed both in Figures 7 and 8. Similar to the phenomenon of allowing degeneracies for the two-letter Li alphabet, it appears that change in the form of the designability distribution of larger alphabets with increasing $\mathscr{F}_{\text{crit}}$ does not mean a change in the relative designability of different structures; those structures that are highly designable for the larger alphabets remain highly designable without becoming more similar to the ordering produced with the two-letter alphabets. This point is particularly emphasized in Figure 8, which compares the relative values of V_k of the same structures for the two-letter Li alphabet and the 20-letter MJ alphabet under different conditions of high $\mathscr{F}_{\text{crit}}$. The V_k values for the smaller and larger alphabets remain uncorrelated; one cannot transform the structure designability results of two-letter alphabet to that of larger alphabet by simply varying selective pressure, either by relaxing the assumption of nondegeneracy or enforcing higher sequence foldabilities. It appears that determining which structures are designable is a property inherent in the size of the alphabet.

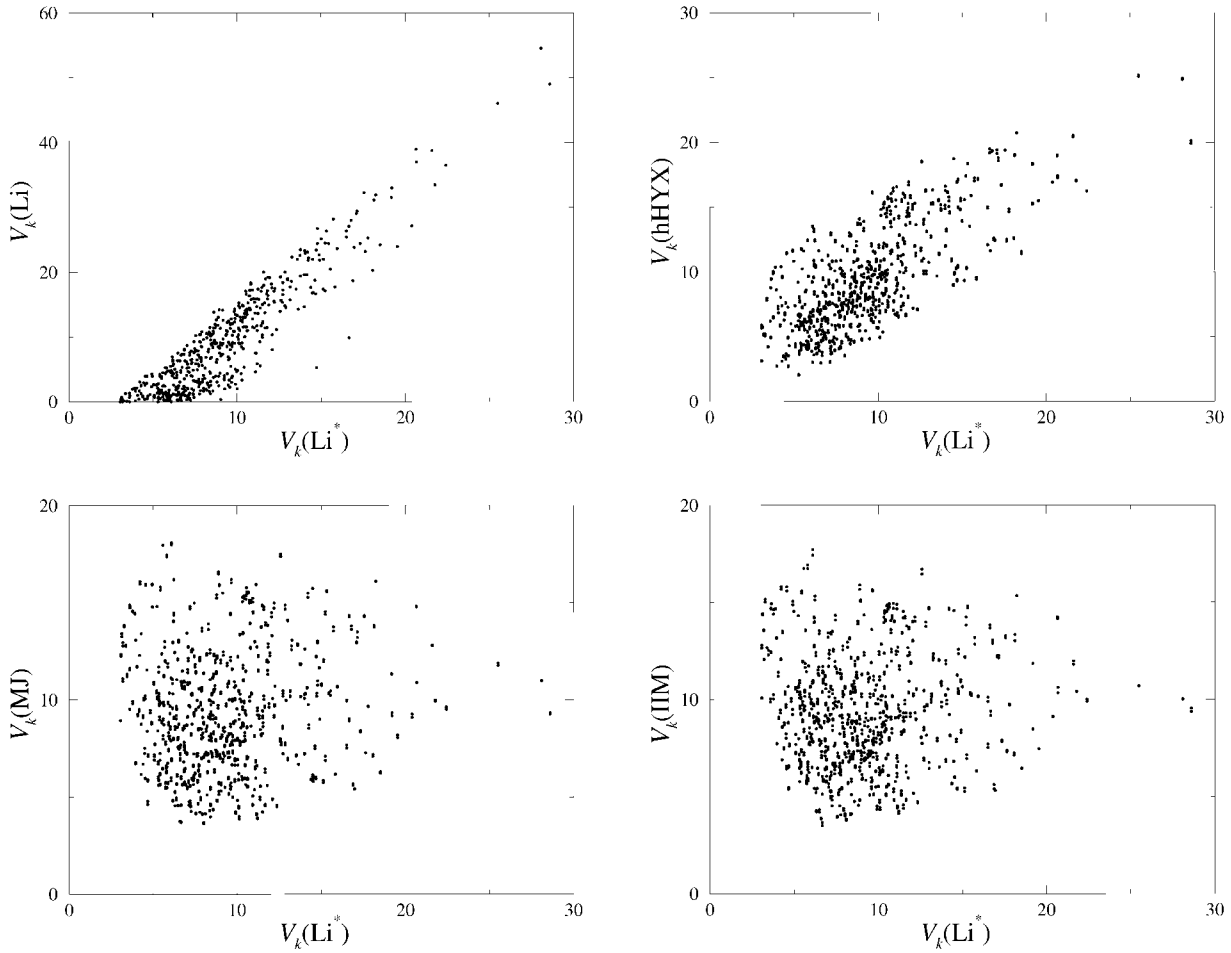

Fig. 5. Scatter plots displaying the V_k values computed with the Li* alphabet, which allows degenerate ground states, against the V_k values of the same structure for selected other alphabets. Even though the distribution of V_k values for the Li* alphabet becomes similar to that of the larger alphabets, the relative V_k values for different structures still maintains a strong correlation with the original Li and other two-letter alphabets.

DISCUSSION

The form of the distribution of designabilities seems to most strongly dependent on the way that protein viability is defined, changing from a gaussian-like distribution to a more exponential distribution as the requirements for viability are increased. For instance, with the larger-alphabet codes, the vast majority of all sequences have nondegenerate ground states and the resulting distribution of designabilities is close to gaussian. If we require that these sequences have a foldability greater than $\mathcal{F}_{\text{crit}}$, the proportion of sequences that are viable decreases, and the distribution becomes more closely exponential. Conversely, for the two-letter codes, the vast majority of sequences have degenerate ground states and are thus considered unviable. The distribution of designabilities is consequentially roughly exponential. Removing the requirement of nondegeneracy of the ground state allows most sequences to be viable, and the designability distribution shifts to more closely gaussian. So in this way, requiring the ground state to be nondegenerate for two-letter alphabets, although of suspect biological relevance, has an effect similar to the requirement for a minimum foldability. This is because, as shown in Table II, sequences with degenerate ground states are more likely to have smaller foldabilities.

By contrast, the relative ordering of the structures by designability depends on the size of the alphabet and is relatively insensitive to how viability is defined. While the differences in the distribution of designabilities can be understood in the context of foldability models, it is more difficult to explain why there is a significant difference in the relative ordering of designabilities of different structures between the two-letter alphabets and the higher-letter alphabets. This problem can be addressed through considering the *interaction landscape* introduced by Govindarajan and Goldstein,[12,13] consisting of the the continuous space of all possible values for the interaction vector, $\boldsymbol{\gamma}_{\mathbf{ij}}^{\mathbf{S}}$ as defined in the methods section. For the 5×5 lattice protein, this interaction landscape is in a 132-dimensional space, where each dimension corresponds to the pair-contact energy γ_{ij} of a pair of residues i and j that can possibly come into contact. Specific sequences correspond

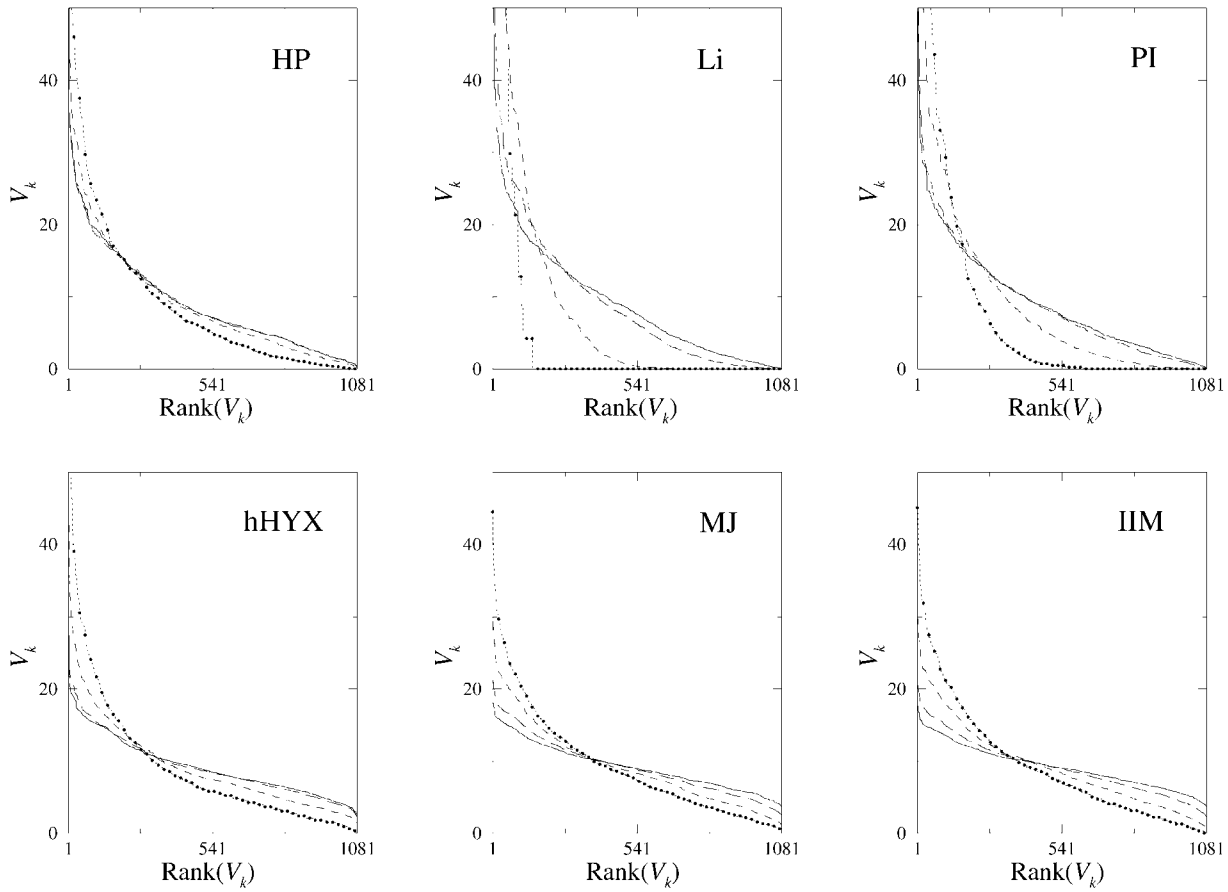

Fig. 6. Zipf's plots showing the effect of foldability requirements on the distribution of designabilities, where only sequences with foldabilities $\mathcal{F}$ larger than a minimum foldability $\mathcal{F}_{\text{crit}}$ are considered viable, for $\mathcal{F}_{\text{crit}} = 0.00$ (—), 3.13 (– – –), 3.77 (----), and 4.30(●–●). This minimum foldability requirement makes the distribution of designabilities more extreme by preferentially eliminating sequences that would otherwise fold into the lesser-designable structures.

to discrete points (unique $\boldsymbol{\gamma}_{\mathbf{ij}}^{\mathbf{S}}$) in this interaction landscape. It is assumed that the $\boldsymbol{\gamma}_{\mathbf{ij}}^{\mathbf{S}}$ corresponding to different possible protein sequences were randomly distributed throughout this interaction landscape.[13] Thus, based on the foldability model, one expects the designability of different structures to be given by relative volumes of the interaction space as sampled by the IIM alphabet. In this report, the infinite-letter IIM alphabet is such a random and unbiased distribution of points in this interaction space. The random distribution of MJ 20-letter sequences throughout interaction space is supported by comparing the pair-correlation function for random pairs of sequences with a random distribution.[55] Additional evidence that MJ sequences are randomly distributed in interaction space is presented in Figures 1, 2, and 3, where both the distribution of V_k and the relative V_k values for particular structures are well correlated between the 20-letter MJ alphabet and the IIM infinite-letter alphabet, giving us confidence in extrapolating the results of the foldability model to the designability of 20-letter MJ proteins.

Conversely, it appears that sequences in the two-letter and four-letter alphabets are likely not distributed in a random way throughout the interaction landscape. This is possibly attributable to correlations between the energies of the possible contacts and the relatively sparse number of possible amino acid pair-contact energies. For the two-letter alphabets, there is a maximum of three possible unique energies (HH, HP, PP). Thus, each dimension of the interaction vector $\boldsymbol{\gamma}_{\mathbf{ij}}^{\mathbf{S}}$ can only have three possible values. As already mentioned, these effects lead both to higher levels of ground-state degeneracy and possible nonisotropic sampling of interaction space. This discrepancy between the interaction-landscape picture and the results from the two-letter alphabet may be exacerbated by the relatively small number of residues that actually interact in any structure. If so, the two-letter alphabet results on longer proteins on larger lattices may be more correlated with the results in interaction space and those of higher alphabets.

CONCLUSION

There have been two classes of attempts to understand why certain proteins are overrepresented among biological proteins. The first class focused on energetic and kinetic considerations and considered the number of sequences that should be able to successfully fold and be stable in a native conformation. One such approach led Govindarajan and Goldstein to the conclusion that those structures with

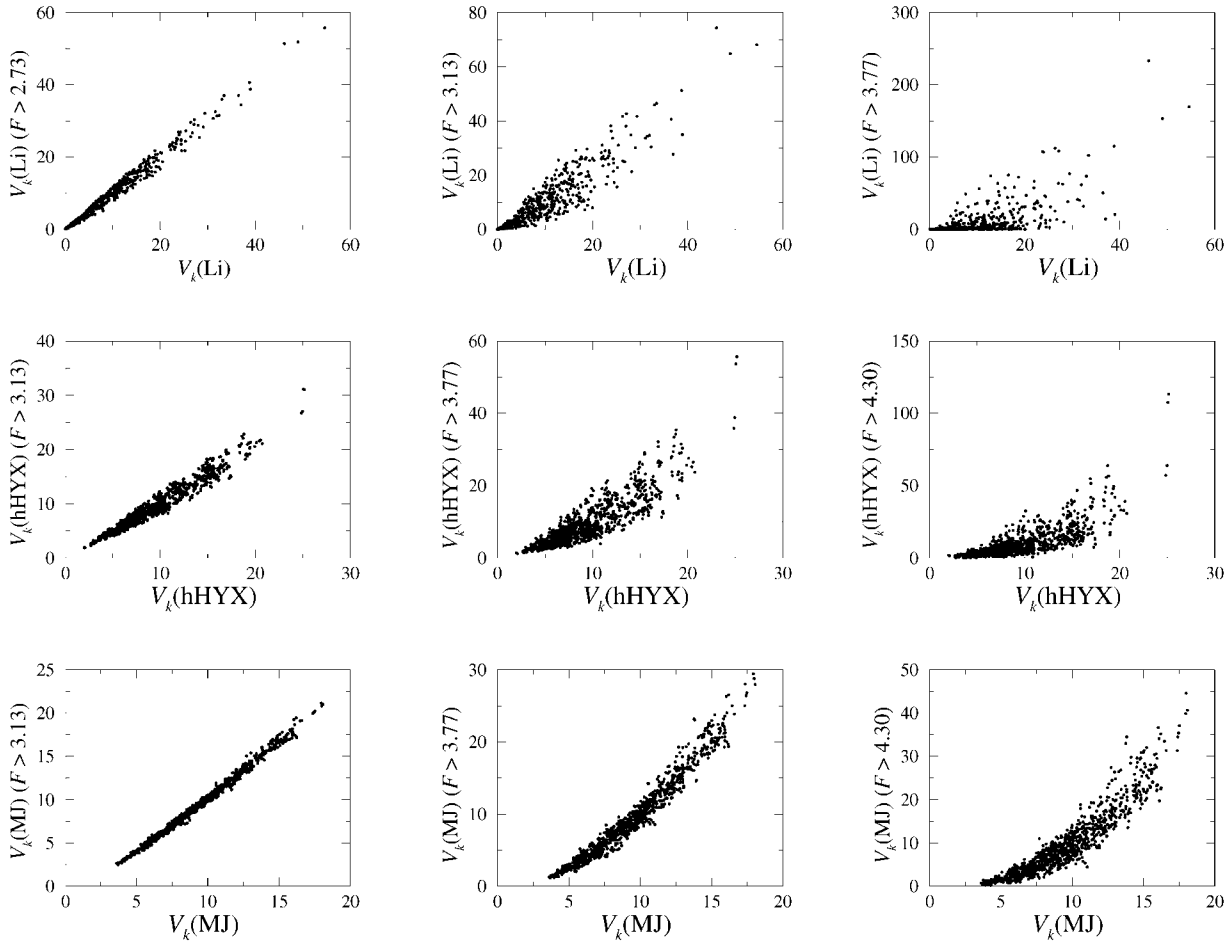

Fig. 7. Scatter plots showing how requiring a foldability value greater than $\mathcal{F}_{\text{crit}}$ affects the relative designability of various structures, for a variety of alphabets. Even with such selective pressure on protein foldability, the relative ordering of structures by designability remains strongly correlated with the original ordering ignoring foldability requirements.

many long-range interactions, and thus larger optimal foldabilities, should be highly designable and overrepresented in the sequence database. These results were based on certain assumptions concerning how real sequences were distributed throughout interaction space. The second class used reduced alphabets and extensive computations with lattice proteins to determine how many sequences had one native-state conformation or another and to look at the properties of both sequences and designable structures. These lattice models were based on the assumption that the presence of a nondegenerate ground state was adequate for ensuring protein foldability and stability. Computational results with these reduced alphabets lead to the conclusion that highly symmetric conformations should be common among naturally occurring proteins.

The question arises: how much do the different assumptions made in these two classes of approaches lead to dissimilar designability results? What is the consequence of using a reduced two-letter alphabet or an interaction space framework (infinite-letter IIM alphabet), as compared with the 20-letter alphabet that "real" proteins work with? How many of the designability conclusions depend on what is considered necessary for adequate stability and foldability? These are the questions that we have tried to address in this report, where we examined how the properties of ensembles of viable proteins depend on the size of the amino acid alphabet and how viability is defined.

Our results indicate that the form of the distribution of designabilities and the relative rank of different structures as ordered by designability are highly dependent on the details of the model. Specifically, the form of the distribution is highly dependent on the proportion of sequences that would be considered viable, shifting from somewhat gaussian to more exponential as the percentage of viable sequences is reduced on the basis of either degeneracy or foldability. Conversely, the relative ordering of which structures are highly designable is mostly dependent on the size of the alphabet, and less dependent on the definition of viability. Most importantly, those structures that are highly designable for the two-letter alphabets are not the same highly designable structures found for the

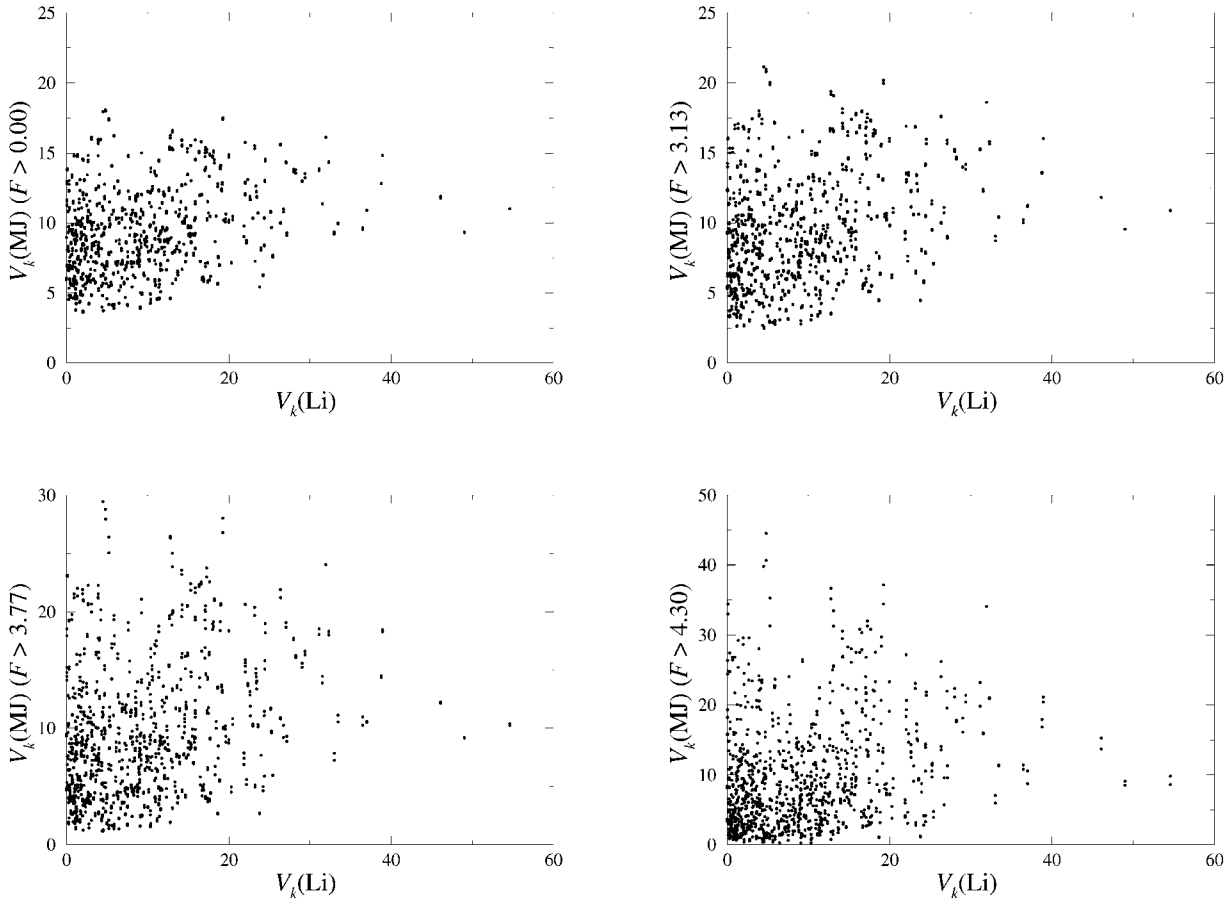

Fig. 8. Scatter plots showing correlations between the relative designabilities of different structures when foldability requirements are imposed on the MJ alphabet and the original Li alphabet, not considering foldability. The presence of selective foldability pressure does not cause the designabilities of proteins with the larger-letter alphabets to become more correlated with those written in the smaller-letter alphabets.

20-letter MJ and infinite-letter IIM alphabet. This suggests that extrapolating structural conclusions from the relative designability of lattice proteins calculated with two-letter alphabets to the properties of real 20-amino acid proteins may be problematic.

In all respects, the designability results of the 20-letter MJ alphabet are highly similar to those of the infinite-letter IIM alphabet, supporting the applicability of the results of the foldability model to natural proteins written in a 20-letter code. (As the interactions between amino acids in the protein can also depend on sidechain conformations, local context, multibody effects, and post-translational modifications, the effective size of the alphabet for natural proteins may actually be significantly larger than 20.) Thus, in light of the recent controversy surrounding the proposal that the *designability principle* is an alternative to the foldability model,[56] these results indicate that foldability provides a framework for understanding the overall distribution of designabilities and identifying which structures would be expected to be particularly designable. In addition, the foldability model can be used to examine how proteins will evolve given the need to fold, how the resulting evolutionarily derived proteins will fold, and what properties characterize the resultant proteins.[33,55,57–61] Thus, the real controversy lies not in the incompatibility of foldability and designability, but rather in understanding why the highly designable structures for two-letter alphabets are so different from those of higher-letter, more realistic alphabets.

ACKNOWLEDGMENTS

We thank Sridhar Govindarajan, Erich Bornberg-Bauer, and Darin Taverna for helpful comments.

REFERENCES

1. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–557.
2. Chothia C. One thousand protein families for the molecular biologist. Nature 1992;357:543–544.
3. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature 1994;372:631–634.
4. Murzin AG, Brenner SE, Hubbard TJP, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
5. Schuster P, Stadler PF. Landscapes: complex optimization prob-

lems and biopolymer structures. Computers Chem 1994;3:295–324.
6. Taverna DM, Goldstein RA. The distribution of structures in evolving protein populations. Folding Design (submitted).
7. Jones DT. Theoretical approaches to designing novel sequences to fit a given fold. Curr Opin Biotechnol 1995;6:452–459.
8. Hellinga HW. Rational protein design: combining theory and experiment. Proc Natl Acad Sci USA 1997;94:10015–10017.
9. Finkelstein AV, Ptitsyn OB. Why do globular proteins fit the limited set of folding patterns? Prog Biophys Mol Biol 1987;50:171–190.
10. Finkelstein AV, Gutin AM, Badretdinov AY. Why are the same protein folds used to perform different functions? FEBS Lett 1993;325:23–28.
11. Finkelstein AV, Gutin AM, Badretdinov AY. Boltzmann-like statistics of protein architectures. Subcell Biochem 1995;24:1–26.
12. Govindarajan S, Goldstein RA. Searching for foldable protein structures using optimized energy functions. Biopolymers 1995;36: 43–51.
13. Govindarajan S, Goldstein RA. Why are some protein structures so common? Proc Natl Acad Sci USA 1996;93:3341–3345.
14. Chan HS, Dill KA. Sequence space soup of proteins and copolymers. J Chem Phys 1991;95:3775–3787.
15. Lipman DJ, Wilbur WJ. Modelling neutral and selective evolution of protein folding. Proc R Soc Lond Biol 1991;245:7–11.
16. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science 1996;273: 666–669.
17. Bornberg-Bauer E. How are model protein structures distributed in sequence space? Biophys J 1997;73:2393–2403.
18. Gutin AM, Shakhnovich EI. Ground state of random copolymers and the discrete random energy model. J Chem Phys 1993;98:8174–8177.
19. Shakhnovich EI. Protein design: a perspective from simple tractable models. Folding Design 1998;3:R45–R58.
20. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscape and motions of proteins. Science 1991;254:1598–1603.
21. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 1987;84: 7524–7528.
22. Bryngelson JD, Wolynes PG. A simple statistical field theory of heteropolymer collapse with application to protein folding. Biopolymers 1990;30:171–188.
23. Šali A, Shakhnovich EI, Karplus MJ. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. J Mol Biol 1994;235:1614–1636.
24. Shakhnovich EI. Proteins with selected sequences fold into unique native conformation. Phys Rev Lett 1994;72:3907–3910.
25. Chiu TL, Goldstein RA. Compaction and folding in model proteins. J Chem Phys 1997;107:4408–4415.
26. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. Toward an outline of the topography of a realistic protein-folding funnel. Proc Natl Acad Sci USA 1995;92:3626–3630.
27. Chan HS, Dill KA. The protein folding problem. Phys Today 1993;46:24–32.
28. Li H, Tang C, Wingreen N. Nature of driving force for protein folding: a result from analyzing the statistical potential. Phys Rev Lett 1997;79:765–768.
29. Crippen GM. Prediction of protein folding from amino acid sequence over discrete conformation spaces. Biochemistry 1991;30: 4232–4237.
30. Bornberg-Bauer E. Chain growth algorithms for HP-type lattice proteins. In: Proceedings of RECOMB97. New York: ACM, 1997; p 47–55.
31. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
32. Anfinsen C. Principles that govern the folding of a protein chain. Science 1973;181:223–230.
33. Govindarajan S, Goldstein RA. On the thermodynamic hypothesis of protein folding. Proc Natl Acad Sci USA 1998;95:5545–5549.
34. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Optimal protein folding codes from spin glass theory. Proc Natl Acad Sci USA 1992;89:4918–4922.
35. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc Natl Acad Sci USA 1992;89:9029–9033.
36. Leopold PE, Montal M, Onuchic JN. Protein folding funnels: a kinetic approach to the sequence–structure relationship. Proc Natl Acad Sci USA 1992;89:8721–8725.
37. Šali A, Shakhnovich EI, Karplus MJ. How does a protein fold? Nature 1994;369:248–251.
38. Betancourt MR, Onuchic JN. Kinetics of proteinlike models: the energy landscape factors that determine folding. J Chem Phys 1995;103:773–787.
39. Shoemaker BA, Wang J, Wolynes PG. Structural correlations in protein folding funnels. Proc Natl Acad Sci USA 1997;94:777–782.
40. Socci ND, Onuchic JN, Wolynes PG. Diffusive dynamics of the reaction coordinate for protein folding funnels. J Chem Phys 1996;104:5860–5868.
41. Pande VS, Grosberg AY, Tanaka T. Statistical mechanics of simple models of protein folding and design. Biophys J 1997;73:3192–3210.
42. Pande VS, Grosberg AY, Tanaka T. On the theory of folding kinetics for short proteins. Folding Design 1997;2:109–114.
43. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.
44. Plotkin SS, Wang J, Wolynes PG. Statistical mechanics of a correlated energy landscape model for protein folding funnels. J Chem Phys 1997;106:2932–2948.
45. Wolynes PG. Folding funnels and energy landscapes of larger proteins within the capillarity approximation. Proc Natl Acad Sci USA 1997;94:6170–6175.
46. Mirny LA, Shakhnovich EI. How evolution makes proteins fold quickly. Proc Natl Acad Sci USA 1998;95:4976–4981.
47. Hao MH, Scheraga HA. Molecular mechanisms for cooperative folding of proteins. J Mol Biol 1998;277:973–983.
48. Chan HS, Dill KA. Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. Proteins 1998;30:2–33.
49. Wang Z-X. How many fold types of protein are there in nature? Proteins 1996;26:186–191.
50. Govindarajan S, Recabarren R, Goldstein RA. Estimating the total number of protein folds. Proteins (Submitted).
51. Fontana W, Stadler PF, Tarazona P, Weinberger ED, Schuster P. RNA folding and combinatory landscape. Phys Rev E 1993;47: 2083–2099.
52. Fontana W, Konings DAM, Stadler PF, Schuster P. Statistics of RNA secondary structures. Biopolymers 1993;33:1389–1404.
53. Renner A, Bornberg-Bauer E. Exploring the fitness landscapes of lattice proteins. In: Pacific Symposium on Biocomputing '97. Altman RB, Dunker AK, Hunter L, Klein TE, eds. World Scientific Singapore 1996, p 361–372.
54. Yue K, Dill KA. Forces of tertiary structural organization in globular proteins. Proc Natl Acad Sci USA 1995;92:146–150.
55. Govindarajan S, Goldstein RA. The foldability landscape of model proteins. Biopolymers 1997;42:427–438.
56. Borman S. Protein folding model focuses on "designability." Chem Eng News 1996;74:36.
57. Shakhnovich EI, Gutin AM. Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 1990; 346:773–775.
58. Gutin AM, Abkevich VL, Shakhnovich EI. Evolution-like selection of fast-folding model proteins. Proc Natl Acad Sci USA 1995;92: 1282–1286.
59. Abkevich VI, Gutin AM, Shakhnovich EI. How the first biopolymers could have evolved. Proc Natl Acad Sci USA 1996;93:839–844.
60. Govindarajan S, Goldstein RA. Evolution of model proteins on a foldability landscape. Proteins. In press.
61. Govindarajan S, Goldstein RA. Site mutations in model proteins. Mathematical Modelling and Scientific Computing. In press.