ELSEVIER

The designability of protein structures

Robert Helling,[1] Hao Li,[2] Régis Mélin,[3] Jonathan Miller, Ned Wingreen, Chen Zeng,[4] and Chao Tang

NEC Research Institute, Princeton, NJ, USA

It has been noted that natural proteins adapt only a limited number of folds. Several researchers have investigated why and how nature has selected this small number of folds. Using simple models of protein folding, we demonstrate systematically that there is a "designability principle" behind nature's selection of protein folds. The designability of a structure (fold) is measured by the number of sequences that can design the structure—that is, sequences that possess the structure as their unique ground state. Structures differ drastically in terms of their designability. A small number of highly designable structures emerge with a number of associated sequences much larger than the average. These highly designable structures possess proteinlike secondary structures, motifs, and even tertiary symmetries. In addition, they are thermodynamically more stable and fold faster than other structures. These results suggest that protein structures are selected in nature because they are readily designed and stable against mutations, and that such a selection simultaneously leads to thermodynamic stability. © 2001 by Elsevier Science Inc.

Keywords: protein folding, lattice models, off-lattice models, enumeration, designability

INTRODUCTION

A protein consists of a chain of amino acids whose sequence is determined by the information in DNA/RNA. An open protein chain, under normal physiological conditions, will fold into a three-dimensional configuration (the native state) to perform its function. For single domain globular proteins, which are our focus here, the length of the chain ranges from ~30 to ~400 amino acids. For these proteins, the surface-to-core ratio (the number of amino acids on the surface of a protein over that in the core) is of the order of unity. A protein can be folded (to its native state) and unfolded (to a flexible open chain) reversibly by changing the temperature, pH, or the concentration of some denaturant in solution. The protein folding problem can be traced back at least 70 years when Wu[1,2] first pointed out that denaturation was in fact the unfolding of the protein from "the regular arrangement of a rigid structure to the irregular, diffuse arrangement of the flexible open chain." A remarkable turning point came about 40 years ago when Anfinsen[3] and coworkers established the so-called "thermodynamic hypothesis." That is, that for single domain proteins (1) the information coded in the amino acid sequence of a protein completely determines its folded structure, and (2) the native state is the global minimum of the free energy. These conclusions should be somewhat surprising to physicists. For the configurational "(free) energy landscape" of a heteropolymer of the size of a protein is typically "rough," in the sense that there are typically many metastable states, some of which have energies very close to the global minimum. How could a protein always fold into its unique native state with the lowest energy? The answer is evolution. Indeed, random sequences of amino acids are usually "glassy" and usually can not fold uniquely. But natural proteins are not random sequences. They are a small family of sequences, selected by nature via evolution, and each has a distinct global minimum that is well separated from other metastable states (Figure 1). One might ask: what are the unique and yet common properties of this special ensemble of proteinlike sequences? Can one distinguish them from other sequences without the arguably impossible task of constructing the entire energy landscape? The answer lies in the heart of the question we introduce in the next paragraph and is the focus of this discussion.

There are about 50,000–100,000 different proteins in the human body, with a much larger number of natural proteins in the biological world. Protein structures are classified into different folds. Proteins of the same fold have the same major secondary structures in the same arrangement with the same topological connections,[4] with some small variations typically in the loop region. In some sense, folds are distinct templates of protein structures. Proteins with a close evolutionary relation often have high sequence similarity and share a common fold. It is intriguing that common folds occur even for proteins with

Corresponding author: C. Tang, NEC Research Institute, 4 Independence Way, Princeton, NJ 08540, USA.

E-mail address: tang@research.nj.nec.com (C. Tang)

[1]Present address: Max Planck Institut für Gravitationsphysik, Albert-Einstein-Institut, Schlaatzweg 1, 14473 Potsdam, Germany.

[2]Present address: Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA 94143, USA.

[3]Present address: Centre de Recherches sur les Très basses températures (CRTBT), BP 166X, 38042 Grenoble Cédex, France.

[4]Present address: Department of Physics, George Washington University, Washington, D.C. 20052, USA.

Journal of Molecular Graphics and Modelling 19, 157–167, 2001

655 Avenue of the Americas, New York, NY 10010

1093-3263/01/$–see front matter
PII S1093-3263(00)00137-6

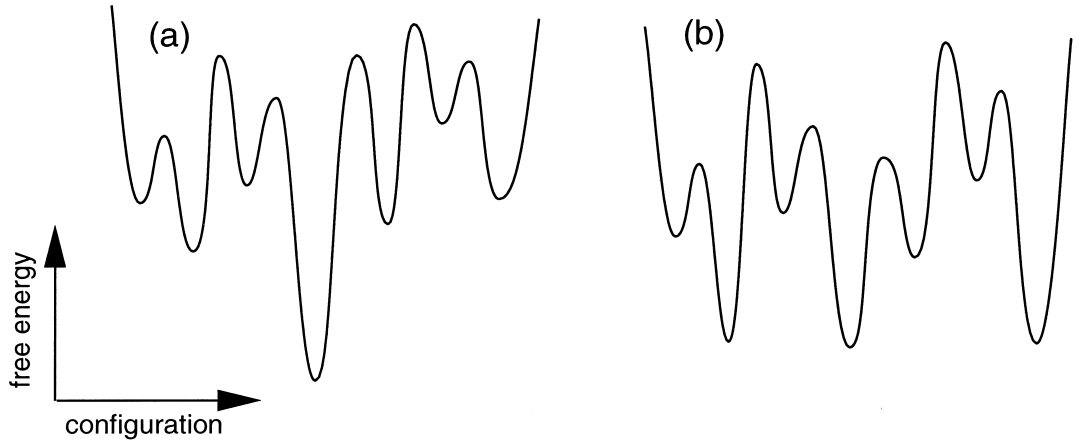

Figure 1. The schematic energy landscapes of (a) a protein sequence and (b) a random sequence.

different evolutionary origins and biological functions. The number of folds is therefore much lower than the number of proteins. Figure 2 shows the cumulative number of solved protein domains along with the cumulative number of folds as a function of the year. It is increasingly less likely that a newly solved protein structure would take a new fold. It is estimated that the total number of folds for all natural proteins is only about 1,000.[5,6] Some of the frequently observed folds, or "superfolds,"[7] are shown in Figure 3. Among apparent features of these folds are secondary structures (α helices and β strands), regularities, and symmetries. Therefore, as in the case of sequences, protein structures or folds are also a very special class.

Is there anything special about natural protein folds—are they merely an arbitrary outcome of evolution or is there some fundamental reason behind their selection? This question has been addressed by a number of authors from different viewpoints. One of the earliest attempts is by Finkelstein and coworkers.[8–10] They argued that certain motifs are easier to stabilize and thus more common, either because they have lower (e.g., bending) energies or because they have unusual energy spectra over random sequences. Yue and Dill[11] observed in a lattice HP model that protein-like folds are associated with sequences that have minimal number of degenerate low energy states. Govindarajan and Goldstein[12–14] suggested that a protein structure should fold fast. They studied the "foldability" of structures in a lattice model and found that the optimal foldability varies from structure to structure. They further argued that structures with larger optimal foldability should tolerate more sequences and be more robust to mutations.

More recently, a "designability principle" was proposed as nature's selection mechanism for protein structures.[15–17] The designability of a structure is defined as the number of sequences that can design the structure—that is, sequences that possess the structure as their unique ground state. With the use of simple models, we have demonstrated that structures differ drastically in their designability. A small number of structures are highly designable while the majority of structures have low designability. The highly designable structures also possess other proteinlike features: thermodynamic stability, mutational stability, fast folding, regular secondary structures, and tertiary symmetries. Our results suggest that protein structures are selected in nature because they are readily designed and stable against mutations, and that such a selection simultaneously leads to thermodynamic stability. In the rest of this study, we

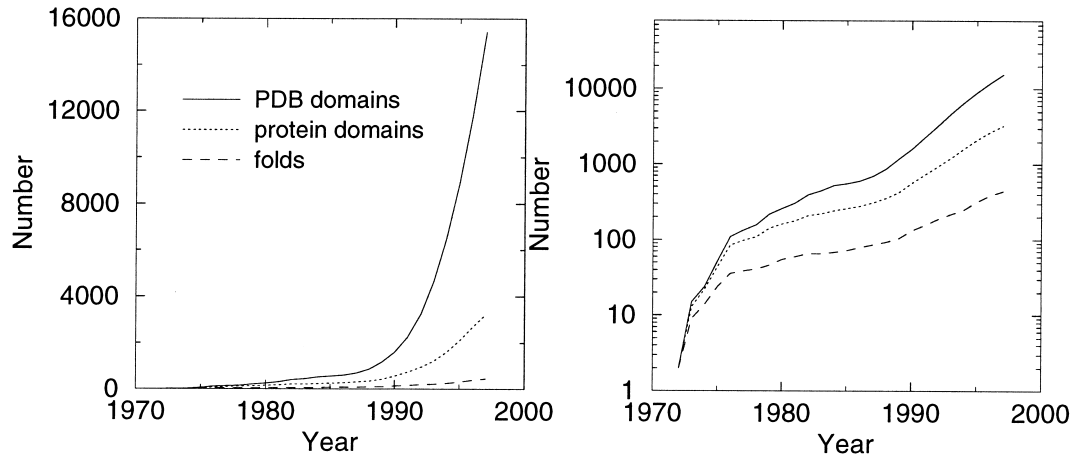

Figure 2. The cumulative numbers of PDB domains, (nonredundant) protein domains, and folds versus year. Source: SCOP[4] and Chothia.[6] Courtesy of Dr. Steven Brenner.

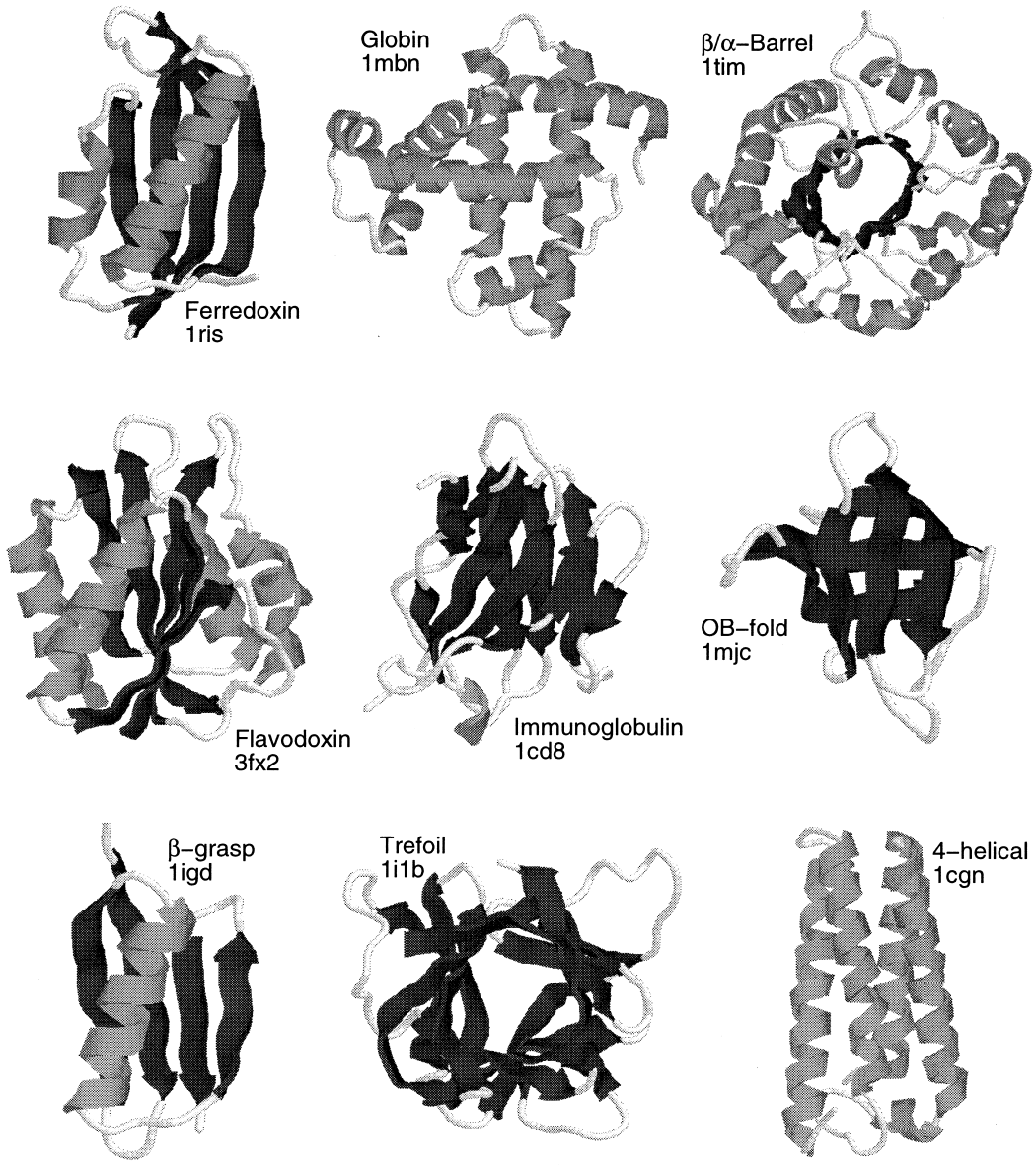

Figure 3. Representatives of some popular folds. β-strand is shown in dark, α-helix in grey, and turns and loops in white.

give a brief review and summary for some of the work on designability.

METHODS

We have used lattice and off-lattice models. For interacting potentials between amino acids, we have used HP, solvation, or Miyazawa-Jernigan matrix.[18,19]

HP Lattice Models

The simplest model of protein folding is the so-called "HP lattice model,"[20–22] whose structures are defined on a lattice and whose sequences take only two "amino acids": H (hydrophobic) and P (polar) (see Figure 4). The energy for a sequence folded into a structure is simply given by the short-range contact interactions

$$H = \sum_{i<j} e_{v_i v_j} \Delta(\mathbf{r}_i - \mathbf{r}_j), \qquad (1)$$

where $\Delta(\mathbf{r}_i - \mathbf{r}_j) = 1$ if $\mathbf{r}_i$ and $\mathbf{r}_j$ are adjoining lattice sites but i and j are not adjacent in position along the sequence, and $\Delta(\mathbf{r}_i - \mathbf{r}_j) = 0$ otherwise. Depending on the types of monomers in contact, the interaction energy $e_{v_i v_j}$ will be e_{HH}, e_{HP}, or e_{PP}, corresponding to H–H, H–P, or P–P contacts, respectively (see Figure 4).[a] We choose these interaction parameters[15] to satisfy the following physical constraints: (1) compact shapes have lower energies than any noncompact shapes; (2) H monomers are buried as much as possible, expressed by the relation $e_{\mathrm{PP}} >$

[a]The system is surrounded by water. The energy $e_{v\mu}$ is the relative energy of forming a $v - \mu$ contact in water. One can think of $e_{v\mu} = E_{v\mu} + E_{ww} - E_{vw} - E_{\mu w}$, where the Es are "absolute" energies and the subscript w denotes water molecules.

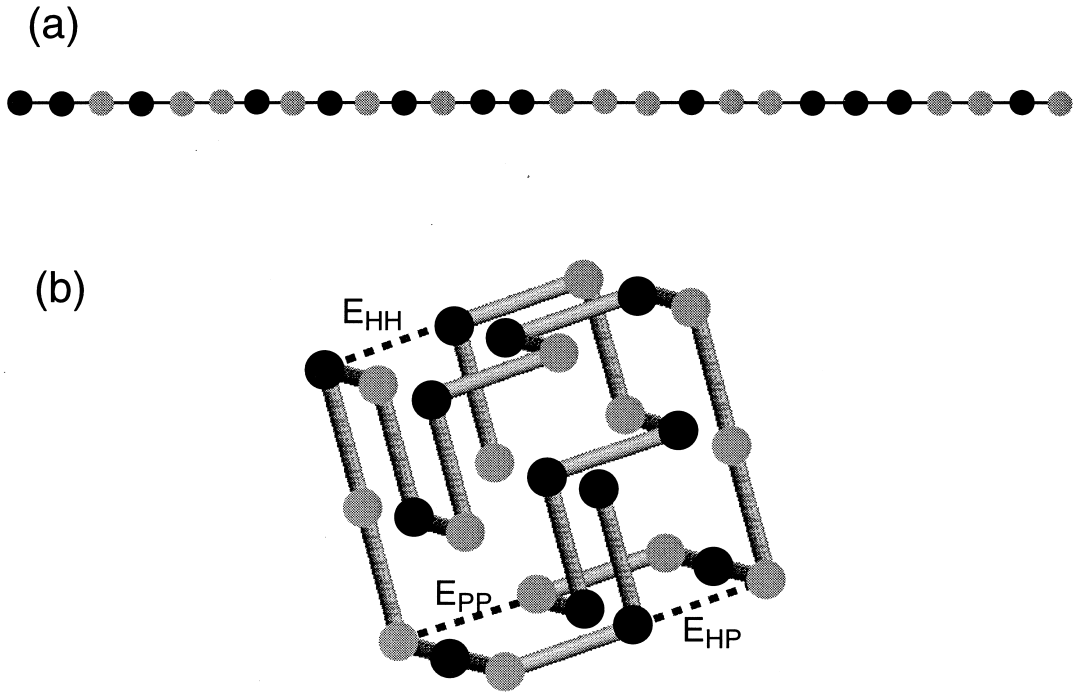

Figure 4. A 3D lattice HP model. A sequence of H (dark disc) and P (light disc) (a) is folded into a 3D structure (b).

$e_{HP} > e_{HH}$, which lowers the energy of configurations in which Hs are hidden from water; and (3) different types of monomers tends to segregate, expressed by $2e_{HP} > e_{PP} + e_{HH}$. Conditions 2 and 3 were derived from the analysis[23] of the real protein data contained in the Miyazawa-Jernigan matrix[18,19] of inter-residue contact energies between different types of amino acids. Since we consider only the compact structures, all of which have the same total number of contacts, we can freely shift and rescale the interaction energies, leaving only one free parameter. In our study, we choose $e_{HH} = -2.3$, $e_{HP} = -1$, and $e_{PP} = 0$ which satisfy conditions 2 and 3 above. The results are insensitive to the value of e_{HH} as long as both of these conditions are satisfied.[b]

Lattice Model with MJ Matrix

To ensure that our results are not an artifact of the HP model, we have studied the lattice model (1) with 20 amino acids.[24] In this case the interaction energies $e_{v_i v_j}$, where v_i can now be any one of the 20 amino acids, are taken from the Miyazawa-Jernigan matrix.[c]

[b]Li, Tang and Wingreen's[23] analysis of the interaction potential of amino acids arrived at a form $e_{\mu\nu} = h_\mu + h_\nu + c(\mu, \nu)$, where h_μ is a measure of hydrophobicity of the amino acid, μ, and c is a small mixing term. The additive term, i.e., the hydrophobic force, dominates the potential. The choice of $e_{HH} = -2.3$ in our study can be viewed as a result of a hydrophobic part -2 plus a small mixing part -0.3. Several authors have investigated the effect of the mixing contribution as a small perturbation to the additive potential.[32–35]

[c]Note that there are two or more matrices in studies by Miyazawa and Jernigan.[18,19] We use the matrix e_{ij}, which is the matrix containing all interactions including the hydrophobic interaction. Other matrices have removed, to various degrees, the hydrophobic contribution (e.g., the matrix e_{ij}' has removed the additive part and contains only the mixing term (see footnote [b]). Thus, using these modified MJ matrices without care may lead to very different and often unphysical results.

Off-Lattice Models

We have also studied the designability of structures for some off-lattice models.[25] Following Park and Levitt,[26] we use a discrete set of dihedral angles (ϕ_i, ψ_i), $i = 1, 2, \cdots, n$, to construct the structures. Side chains are represented by hard-core spheres centered around c_β atoms. We use a form of solvation energy as the energy function for a sequence folded onto a structure:

$$H = -\sum_{i=1}^{N} s_i h_{v_i}, \qquad (2)$$

where h_{v_i} is the hydrophobicity of the ith residue v_i along the chain and s_i is the degree of burial of the ith residue in the structure. Equation 2 is essentially a solvation model[27] at the residue level.[d]

RESULTS

HP Lattice Model

We have studied the HP lattice model (1) on a three-dimensional cubic lattice and on a two-dimensional square lattice.[15] For the three-dimensional case, we analyze a chain composed of 27 monomers. We consider all the structures that form a compact $3 \times 3 \times 3$ cube. There are a total of 51,704 such structures unrelated by rotational, reflection, or reverse labeling symmetries.[15,28] For a given sequence, the ground state structure is found by calculating the energies of all compact structures. We completely enumerate the ground states of all 2^{27} possible sequences. We find that only 4.75% of the se-

[d]Equation (2) can also be obtained by taking the mixing term of the equation in footnote [b] to zero.[16,32–35]

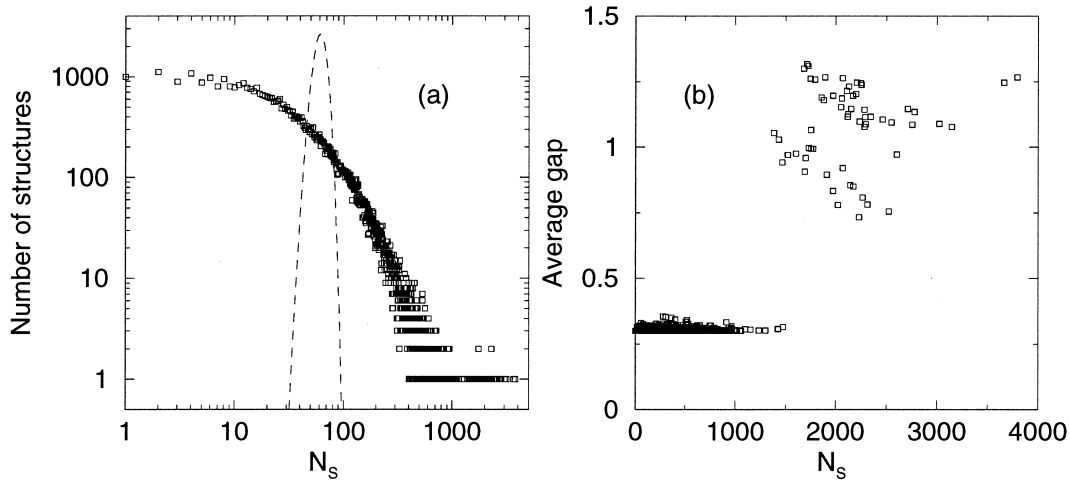

Figure 5. (a) Histogram of N_S *for the 3 × 3 × 3 system. (b) Average energy gap between the ground state and the first excited state versus* N_S *for the 3 × 3 × 3 system.*

quences have unique ground states and thus are potential proteinlike sequences. We then calculate the designability of each compact structure. Specifically, we count the number of sequences, N_S, that have a given compact structure S as their unique ground state. We find that compact structures differ drastically in terms of their designability, N_S. There structures can be designed by an enormous number of sequences, and there are "poor" structures that can only be designed by a few or even no sequences. For example, the top structure can be designed by 3,794 different sequences (N_S = 3,794), while there are 4,256 structures for which N_S = 0. The number of structures having a given N_S decreases monotonically (with small fluctuations) as N_S increases (Figure 5a). There is a long tail to the distribution. Structures contributing to the tail of the distribution have $N_S \gg \overline{N_S}$ = 61.7, where N_S is the average number. We call these structures "highly designable" structures. The distribution is very different from the Poisson distribution (also shown in Figure 5a) that would result if the compact structures were statistically equivalent. For a Poisson distribution with a mean $\overline{N_S}$ = 61.7, the probability of finding even one structure with $N_S > 120$ is 1.76×10^{-6}.

The highly designable structures are, on average, thermodynamically more stable than other structures. The stability of a structure can be characterized by the average energy gap $\overline{\delta_S}$, averaged over the N_S sequences that design the structure. For a given sequence, the energy gap δ_S is defined as the minimum energy difference between the ground state energy and the energy of a different compact structure. We find that there is a marked correlation between N_S and $\overline{\delta_S}$ (Figure 5b). Highly designable structures have average gaps much larger than those of structures with small N_S, and there is a sudden jump in $\overline{\delta_S}$ for structures with $N_S^c \approx 1{,}400$. This jump is a result of two possible different kinds of ground state excitations. One is to break an H–H bond and a P–P bond to form two H–P bonds, with an (mixing) energy cost of $2E_{HP} - E_{HH} - E_{PP} = 0.3$. The other is to change the position of an H-mer from relatively buried to relatively exposed so the number of H-water bonds (the lattice sites outside the 3 × 3 × 3 cube are occupied by water molecules) is increased. This kind of excitation has an energy ≥1. The jump in Figure 5b indicates that the lowest excitations are of the first kind for $N_S < N_S^c$, but are a mixture of the first and the second kind for $N_S > N_S^c$.

A striking feature of the highly designable structures is that they exhibit certain geometrical regularities that are absent

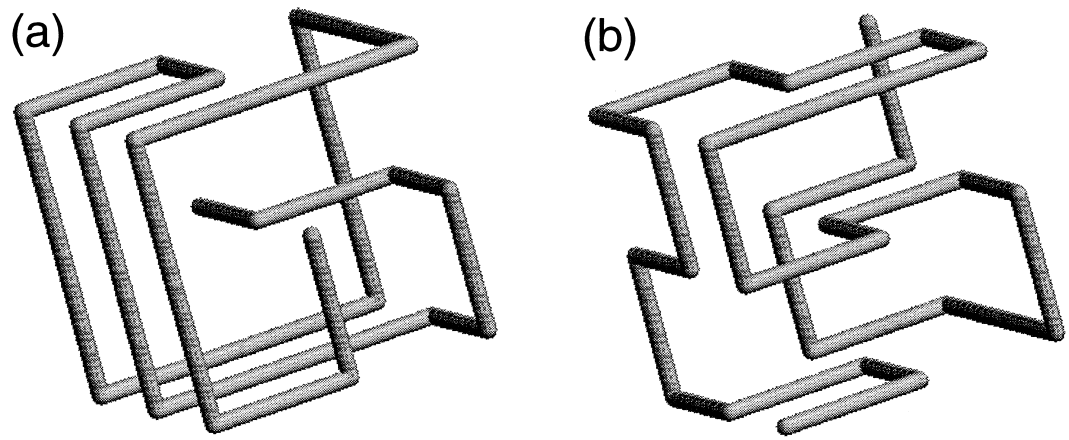

Figure 6. The top structure (a) and an ordinary structure with $N_S = 1$ *(b) for the 3 × 3 × 3 system.*

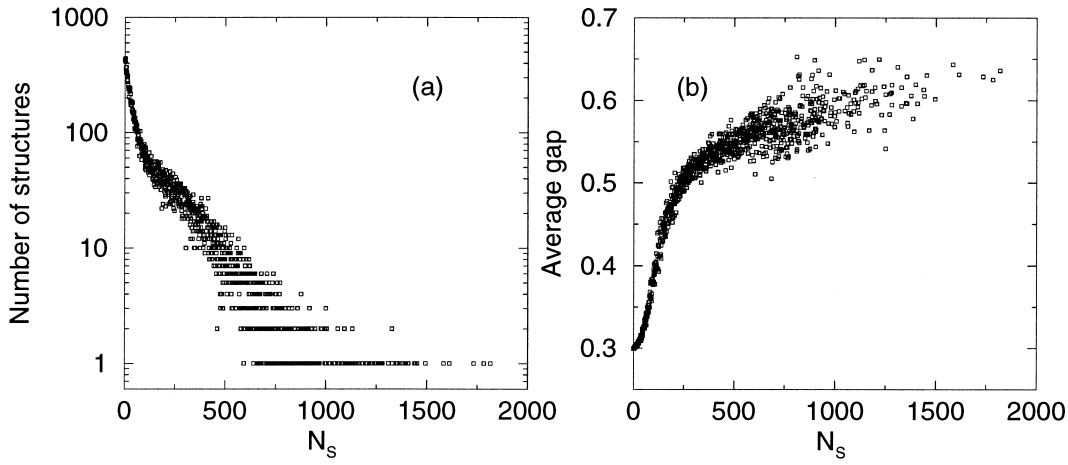

Figure 7. Histogram of N_S *(a), and the average energy gap between the ground state and the first excited state versus* N_S *(b), for the 2D 6 × 6 HP model.*

from random structures and are reminiscent of the secondary structures in natural proteins. Figure 6 shows the most designable structure along with a typical random structure. We examined the compact structures with the 10 largest N_S values and found that all have parallel running lines folded in a regular manner.

We have also studied the model on a 2D lattice. We take sequences of length 36 and fold them into compact 6 × 6 structures on the square lattice. There are 28,728 such structures unrelated by symmetries including the reverse-labeling symmetry. In this case, we did not enumerate all 2^{36} sequences but randomly sampled them to the extend where the histogram for N_Ss reached a reliable distribution. Similar to the 3D case, the N_Ss have a very broad distribution (Figure 7a). In this case the tail decays more like an exponential. The average gap also correlates positively with N_S (Figure 7b). Again, similar to the 3D case, we observe that the highly designable structures in 2D also exhibit secondary structures. In the 2D 6 × 6 case, as the surface-to-interior ratio approaches that of real proteins, the highly designable structures often have bundles of pleats and long strands, reminiscent of α helices and β strands in real proteins; in addition, some of the highly designable structures have tertiary symmetries (Figure 8).

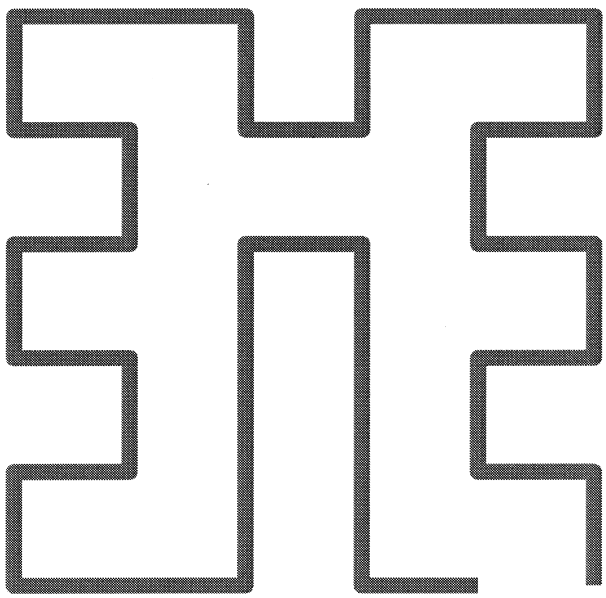

Figure 8. The top structure for the 2D 6 × 6 system.

Lattice Model with MJ Matrix

For the 3D × 3 × 3 system and the 2D 6 × 6 system, the total numbers of sequences are 20^{27} and 20^{36}, respectively, which are impossible to enumerate. So we randomly sampled the sequence space. Similar to the case of the HP model, N_Ss have a broad distribution in both 3D and 2D cases and the N_S has a positive correlation with the average gap. The data for the 6 × 6 system is shown in Figure 9. Furthermore, the N_Ss calculated with the MJ matrix correlate well with the ones obtained from the HP model (Figure 10). Thus the highly designable structures in the HP model are also highly designable in the 20-letter model.[e] With 20 amino acids, there are few sequences that will have exactly degenerate ground states. For example, in the case of 3 × 3 × 3 about 96.7% of the sequences have unique ground states. However, many of these ground states are almost degenerate, in the sense that there are compact structures other than the ground state with energies very close to the ground state energy. If we require that for a ground state to be truly unique there should be no other states of energies within g_c from the ground state energy, then the percentage of the sequences that have unique ground states is reduced to about 30% and 8% for $g_c = 0.4k_B$T and $g_c = 0.8k_B$T, respectively.

We find similar results as in the lattice models; structures differ drastically in their designability and that designability correlates well with thermodynamic stability. In Figure 11 we show the histogram of N_S for a 3-state model with a chain length of 23. One of the interesting results, which can not be seen in the simple lattice models we studied, is that some

[e]Recently, Buchler and Goldstein[36,37] studied the designability for structures on a 5 × 5 lattice, using various alphabet sizes. They obtained very poor or no correlation between the N_S's from our HP model and the "MJ" model. The reason for this discrepancy is that they have used a different MJ matrix (see footnote [c]).

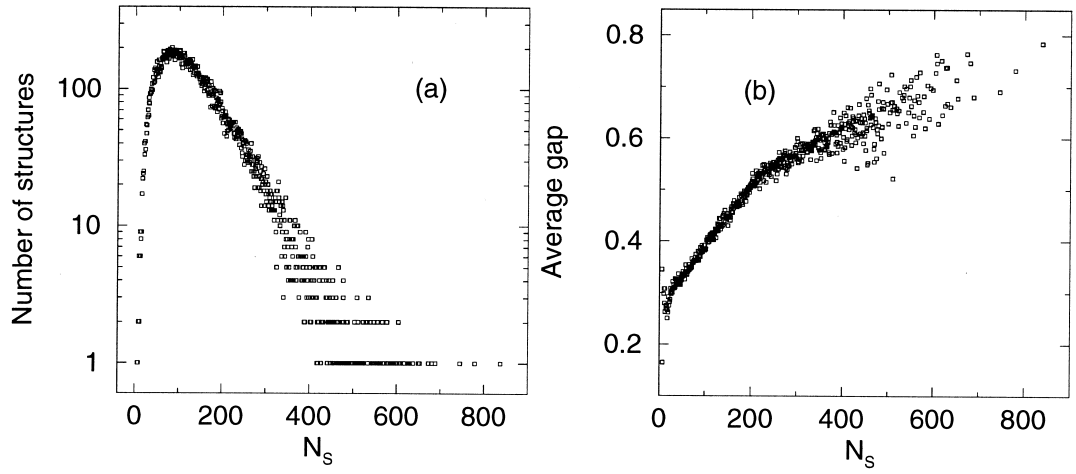

Figure 9. (a) Histogram of N_S*; (b) average energy gap between the ground state and the first excited state versus* N_S*, for the 2D 6 × 6 model with the MJ matrix. Data obtained with 3,995,000 random sequences.*

natural protein motifs like the zinc finger are among the top designable structures.

DISCUSSION

A number of questions arise: Among the large number of structures, why are some structures highly designable? Why does designability also guarantee thermodynamic stability? Why do highly designable structures have geometrical regularities and even symmetries? In this section we address these questions by using a geometrical formulation of the protein folding problem.[16]

Let us go back to Equation 2. To simplify the discussion, let us consider only compact structures and let s_i take only two values: 0 and 1, depending on whether the amino acid is on the surface or in the core of the structure, respectively. Therefore, each compact structure can be represented by a string $\{s_i\}$ of 0s and 1s: $s_i = 0$ if the ith amino acid is on the surface and $s_i = 1$ if it is in the core (see Figure 12a for an example on a lattice). Let us make further simplification by using only two amino acids: v_i = H or P, and let $h_H = 1$ and $h_P = 0$. Thus, a sequence $\{v_i\}$ is also mapped into a string $\{\sigma_i\}$ of 0s and 1s: $\sigma_i = 1$ if v_i = H, and $\sigma_i = 0$ if v_i = P. Let us call this model the PH (Purely Hydrophobic) model. Assuming every compact structure of a given size has the same numbers of surface and core sites and noting that the term $\sum_i \sigma_i^2$ is a constant for a fixed sequence of amino acids and does not play any role in determining the relative energies of structures folded by the sequence, Equation (2) is then equivalent to[16]:

$$H = \sum_{i=1}^{N} (\sigma_i - s_i)^2. \tag{3}$$

Therefore, the energy for a sequence $\vec{\sigma} = \{\sigma_i\}$ folded onto a structure $\vec{s} = \{s_i\}$ is simply the distance squared (or the Hamming distance in the case where both $\{\sigma_i\}$ and $\{s_i\}$ are strings of 0s and 1s) between the two vectors $\vec{\sigma}$ and $\vec{s}$.

We can now formulate the designability question geometrically. We have two ensembles or spaces: one being all the

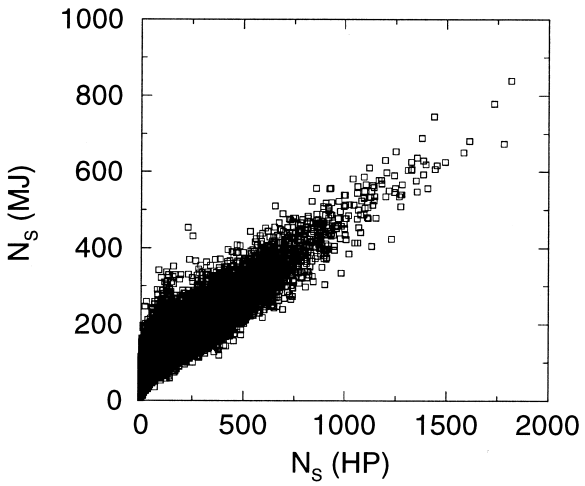

Figure 10. N_S *from the HP model versus* N_S *from the MJ matrix for 2D 6 × 6 structures.*

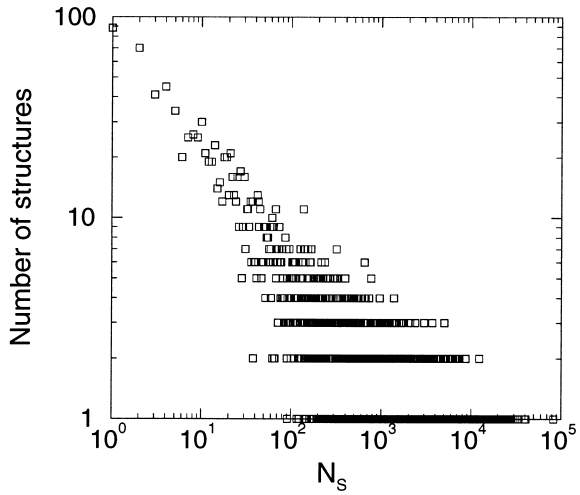

Figure 11. Histogram of N_S *for a 3-state off-lattice model with N = 23.*

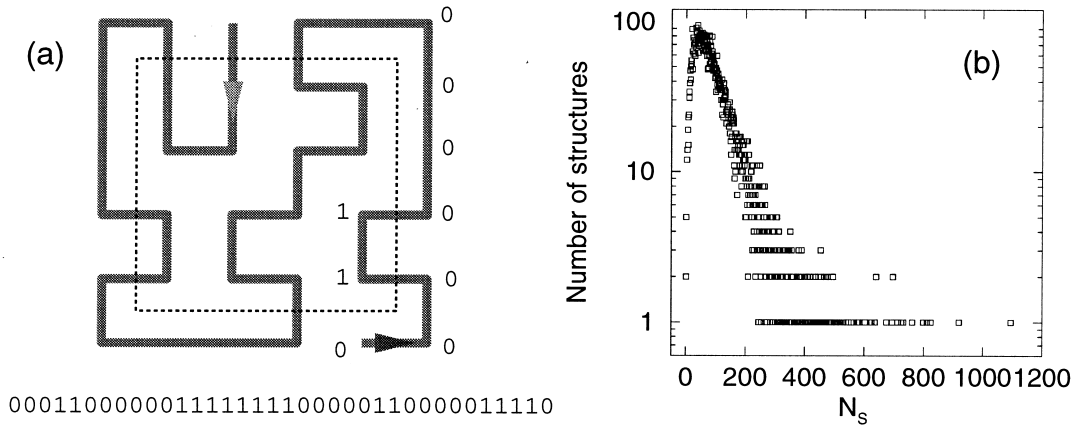

Figure 12. (a) A 6 × 6 compact structure and its corresponding string. A structure is represented by a string $\vec{s}$ *of 0s and 1s, according to whether a site is on the surface or in the core (which is enclosed by the dotted lines), respectively. Two structures, related by the reverse-labeling symmetry, are shown, corresponding to the two opposite paths indicated by the two arrows. The* $\vec{s}$ *of one structure is the reverse of the other. (b) The histogram of* N_S *for the 6 × 6 PH model obtained by using 19,492,200 randomly chosen sequences.*

sequences $\{\vec{\sigma}\}$ and the other all the structures $\{\vec{s}\}$. Both are represented by *N*-dimensional points or vectors where *N* is the length of the chain. The points of all the sequences are trivially distributed in the *N*-dimensional space. In the case of the PH model, the points representing sequences are all the vertices of an *N*-dimensional hypercube (all possible 2^N strings of 0s and 1s of length *N*). On the other hand, the points representing all the structures $\{\vec{s}\}$ have a very different distribution in the *N*-dimensional space. The $\vec{s}$s are constrainted and correlated. For example, in the case of the PH model where $s_i = 0$ or 1, not every string of 0s and 1s actually represents a structure. In fact, only a very small fraction of the 2^N strings of 0s and 1s correspond to structures. If we consider only compact structures where $\sum_i s_i = n_c$ with n_c the number of core sites, then the structure vectors $\{\vec{s}\}$ cover only a small fraction of the vertices of a hyperplane in the *N*-dimensional hypercube.

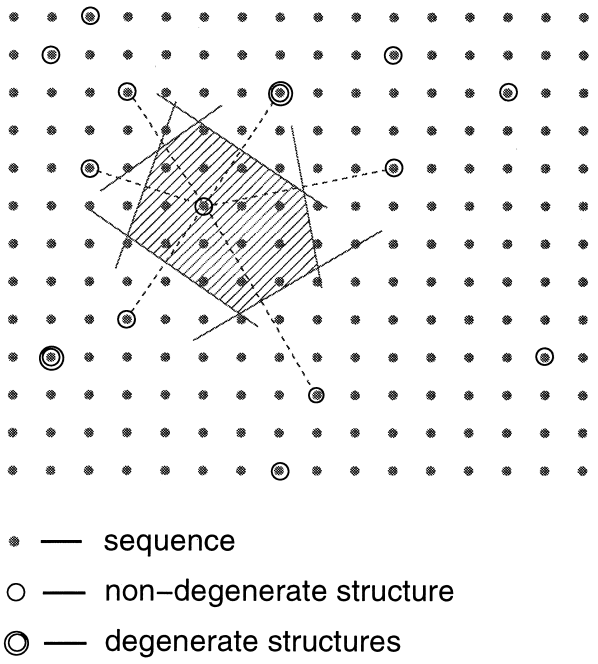

Figure 13. The sequence and structure ensembles in N*-dimension.*

Now imagine putting all the sequences $\{\vec{\sigma}\}$ and all the structures $\{\vec{s}\}$ together in the *N*-dimensional space (see Figure 13 for a schematic illustration). (In a more general case it would be simplest to picture if one normalizes $\{\vec{h}\}$ and $\{\vec{s}\}$ so that $0 \leq h_i, s_i \leq 1$.) From Equation 3, it is evident that a sequence will have a structure as its unique ground state only if the sequence is closer (measured by the distance defined by Equation 3) to the structure than to any other structures. Therefore, the set of all sequences $\{\vec{\sigma}(\vec{s})\}$ that uniquely design a structure $\vec{s}$ can be found by the following geometrical construction: draw bisector planes between $\vec{s}$ and all of its neighboring structures in the *N*-dimensional space (see Figure 13). The volume enclosed by these planes is called the Voronoi polytope around $\vec{s}$. $\{\vec{\sigma}(\vec{s})\}$ then consists of all sequences within the Voronoi polytope. Hence, the designabilities of structures are directly related to the distribution of the structures in the *N*-dimensional space. A structure closely surrounded by many neighbors will have a small Voronoi polytope and hence a low designability; while a structure far away from others will have a large Voronoi polytope and hence a high designability. Furthermore, the thermodynamic stability of a folded structure is directly related to the size of its Voronoi polytope. For a sequence $\vec{\sigma}$, the energy gap between the ground state and an excited state is the difference of the squared distances between $\vec{\sigma}$ and the two states (Equation 3). A larger Voronoi polytope implies, on average, a larger gap as excited states can only lie outside of the Voronoi polytope of the ground state. Thus, this geometrical representation of the problem naturally explains the positive correlation between the thermodynamic stability and the designability.

As a concrete example, we have studied a 2D PH model of 6×6.[16] For each compact structure, we divide the 36 sites into 16 core sites and 20 surface sites (see Figure 12a). Among the 28,728 compact structures unrelated by symmetries, there are 119 that are reverse-labeling symmetric (for a reverse-labeling

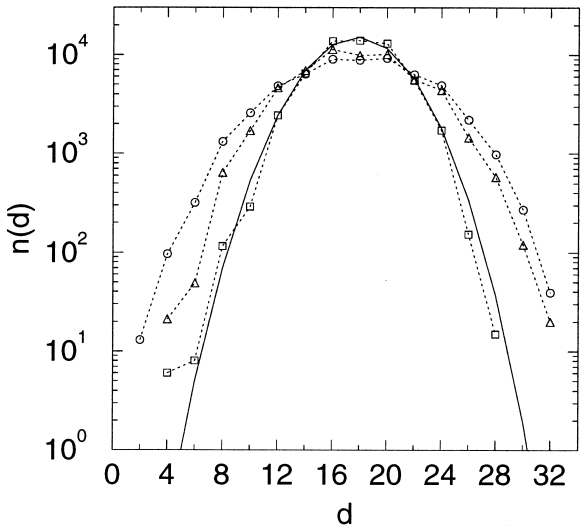

Figure 14. Number of structures versus the Hamming distance for three structures with low (circles), intermediate (triangles), and high (squares) designability. Also plotted is $n^0(d)$ *(solid line).*

symmetric structure, $s_i = s_{N+1-i}$). So the total number of structures a sequence can fold onto is $(28{,}728 - 119) \times 2 + 119 = 57{,}337$, which map into 30,408 distinct strings. There are cases in which two or more structures map into the same string. We call these structures degenerate structures, and a degenerate structure can not be the unique ground state for any sequence in the PH model. Out of the 28,728 structures, there are 9,141 nondegenerate structures (or 18,213 out of 57,337). A histogram for the designability of nondegenerate structures is obtained by sampling the sequence space using 19,492,200 randomly chosen sequences and is shown in Figure 12b. The set of highly designable structures is essentially the same as those obtained from the HP model discussed in the previous section. To further probe how structure vectors are distributed in the N-dimensional space, we measure the number of structures, $n_{\vec{s}}(d)$, at a Hamming distance, d, from a given structure $\vec{s}$. Note that all the 57,337 structures are distributed on the vertices of the hyperplane defined by $\sum_i s_i = 16$. There are a total of $C_{36}^{16} = 7{,}307{,}872{,}110$ vertices in the hyperplane. If the structure vectors were distributed uniformly on these vertices, $n_{\vec{s}}(d)$ would be the same for all structures and would be: $n^0(d) = \rho N(d)$, where $\rho = 57{,}337/7{,}307{,}872{,}110$ is the average density of structures on the hyperplane and $N(d) = C_{16}^{d/2} C_{20}^{d/2}$ is the number of vertices at distance, d, from a given vertex. In Figure 14, $n_{\vec{s}}(d)$ is plotted for three different structures with low, intermediate, and high designabilities, respectively, along with $n^0(d)$. We see that a highly designable structure typically has fewer neighbors than a less designable structure, not only at the smallest ds but out to ds of order 10–12. Also, $n_{\vec{s}}(d)$ is considerably larger than $n^0(d)$ for small d for structures with low designability. These results indicate that the structures are very nonuniformly distributed and are clustered—there are highly populated regions and lowly populated regions. A quantitative measure of the clustering environment around a structure is the second moment of $n_{\vec{s}}(d)$,

$$\gamma^2(\vec{s}) = \langle d^2 \rangle - \langle d \rangle^2 = 4 \sum_{ij} s_i s_j c_{ij}, \tag{4}$$

where

$$c_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle \tag{5}$$

and $\langle \cdot \rangle$ denotes average over all structures.

What are the geometrical characteristics of the structures in the highly populated regions and lowly populated regions, respectively? Naively, the structures in the highly populated regions are typical random structures that can be easily transformed from one to another by small local changes. On the other hand, structures in lowly populated regions are "atypical" structures, which tend to be more regular and "rigid." They have fewer neighbors so it is harder to transform them to other structures with only small rearrangements. One geometrical feature of highly designable structures is that they have more surface-to-core transitions along the backbone, i.e., there are more transitions between 0s and 1s in the structure string for a

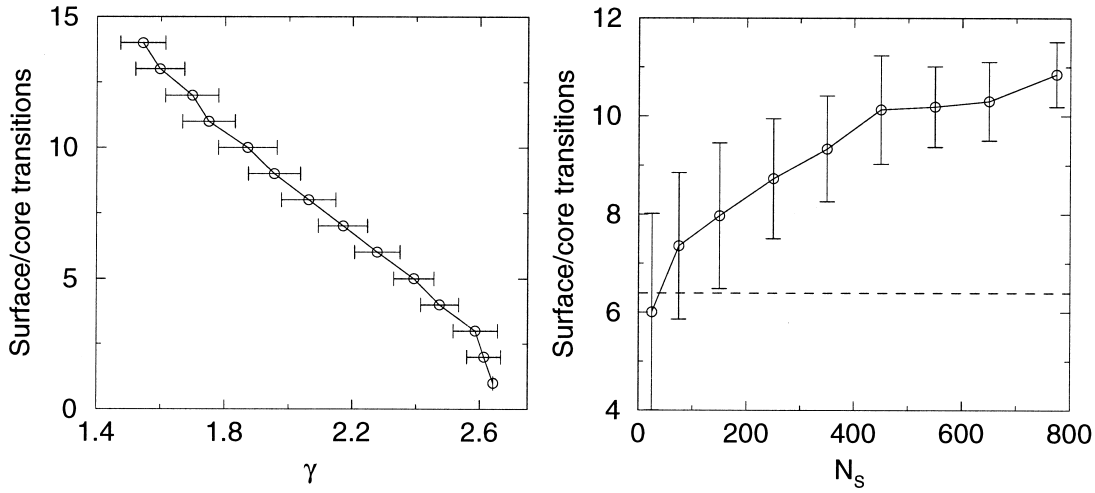

Figure 15. (a) The number of transitions between surface and core sites vs. γ *for all the* 6×6 *compact structures. (b) The number of transitions between surface and core site versus designability.*

highly designable structure than average (see Figure 15b).[16,29] We found that the number of surface-core transitions in a structure correlates well with γ (Figure 15a). Thus, a highly designable structure will have a small γ or a large number of surface-core transitions.

A great advantage of the PH model is that it is simple enough to test some ideas immediately. Two quantities often used to characterize structures are the energy spectra $N(E, \vec{s})$[9,30] and $N(E, \vec{s}, C)$.[30] The first one is the energy spectrum of a given structure, $\vec{s}$, over all sequences, $\{\vec{\sigma}\}$:

$$N(E, \vec{s}) = \sum_{\{\vec{\sigma}\}} \delta[H(\vec{\sigma}, \vec{s}) - E]. \quad (6)$$

The second one is over all sequences of a fixed composition C (e.g., fixed numbers of H-mers and P-mers in the case of two-letter code), $\{\vec{\sigma}\}_C$:

$$N(E, \vec{s}, C) = \sum_{\{\vec{\sigma}\}_C} \delta[H(\vec{\sigma}, \vec{s}) - E]. \quad (7)$$

It is easy to see that if two structure strings $\{s_i\}$ and $\{s_i'\}$ are related by permutation, i.e., $s_i = s_{k_i}'$, for $i = 1, 2, \cdots, N$, where $k_1, k_2, \cdots, k_N$ is a permutation of 1, 2, $\cdots$, N, then $N(E, \vec{s}) = N(E, \vec{s}')$ and $N(E, \vec{s}, C) = N(E, \vec{s}', C)$. Thus all maximally compact structures have the same energy spectra Equations 6 and 7. Therefore, in the case studied here structures differ in designability, not because they have different energy spectra Equations 6 and 7 as speculated elsewhere,[9,30] but because they have different neighborhoods in the structure space.

CONCLUSIONS

We have demonstrated with simple models that structures are very different in terms of their designability and that high designability leads to thermodynamic stability and "protein-like" structural motifs. Highly designable structures emerge because of an asymmetry between the sequence and the structure ensembles. A broad distribution of designability has also been found in RNA secondary structures.[31] However, the set of all sequences designing a good structure, instead of forming a compact Voronoi polytope like in proteins, forms a "neutral network" percolating the entire space.[31] It would be interesting to study the similarities and differences of the two systems. Finally, our picture indicates that the properties of the protein-like sequences are intimately coupled to those of the proteinlike (i.e., the highly designable) structures; the picture unifies various aspects of the two special ensembles. It also suggests that understanding the emergence and properties of the highly designable structures is a key to the protein folding problem.

REFERENCES

1 Wu, H. A theory of denaturation and coagulation of proteins. *Am. J. Physiol.* 1929, **90,** 562–563
2 Wu, H. Studies on denaturation of proteins XIII. A theory of denaturation. *Chinese J. Physiol.* 1931, **5,** 321–344
3 Anfinsen, C. Principles that govern the folding of protein chains. *Science* 1973, **181,** 223–230
4 Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. SCOP: a structural classification of protein database for the investigation of sequences and structures. *J. Mol. Biol.* 1995, **247,** 536–540. (http://scop.mrc-lmb.cam.ac.uk/scop/)
5 Brenner, S.E., Chothia, C., and Hubbard, T.J.P. *Curr. Opin. Struct. Biol.* 1997, **7,** 369–376
6 Chothia, C. One thousand families for the molecular biologist. *Nature* 1992, **357,** 543–544
7 Orengo, C.A., Jones, D.T., and Thornton, J.M. Protein superfamilies and domain superfolds. *Nature* 1994, **372,** 631–634
8 Finkelstein, A.V., and Ptitsyn, O.B. Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.* 1987, **50,** 171–190
9 Finkelstein, A.V., Gutin, A.M., and Badretdinov, A.Y. Why are the same protein folds used to perform different functions? *FEBS Lett.* 1993, **325,** 23–28
10 Finkelstein, A.V., Badretdinov, A.Y., and Gutin, A.M. Why do protein architectures have Boltzmann-like statistics. *Proteins* 1995, **23,** 142–150
11 Yue, K., and Dill, K.A. Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* 1995, **92,** 146–150
12 Govindarajan, S., and Goldstein, R.A. Searching for foldable protein structures using optimized energy functions. *Biopolymers* 1995, **36,** 43–51
13 Govindarajan, S., and Goldstein, R.A. Why are some protein structures so common? *Proc. Natl. Acad. Sci. U.S.A.* 1996, **93,** 3341–3345
14 Govindarajan, S., and Goldstein, R.A. The foldability landscape of model proteins. *Biopolymers* 1997, **42,** 427–438
15 Li, H., Helling, R., Tang, C., and Wingreen, N.S. Emergence of preferred structures in a simple model of protein folding. *Science* 1996, **273,** 666–669
16 Li, H., Tang, C., and Wingreen, N.S. Are protein folds atypical? *Proc. Natl. Acad. Sci. U.S.A.* 1998, **95,** 4987–4990
17 Mélin, R., Li, H., Wingreen, N.S., and Tang, C. Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. *J. Chem. Phys.* 1999, **110,** 1252–1262
18 Miyazawa, S., and Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* 1985, **18,** 534–552
19 Miyazawa, S., and Jernigan, R.L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 1996, **256,** 623–644
20 Lau, K.F., and Dill, K.A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 1989, **22,** 3986–3997
21 Lau, K.F., and Dill, K.A. Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 1990, **87,** 638–642
22 Chan, H.S., and Dill, K.A. Sequence space soup of proteins and copolymers. *J. Chem. Phys. 1991,* **95,** 3775–3787
23 Li, H., Tang, C., and Wingreen, N.S. Nature of driving force for protein folding: a result from analyzing the statistical potential. *Phys. Rev. Lett.* 1997, **79,** 765–768
24 Li, H., Wingreen, N., and Tang, C. to be published.

25 Miller, J., Zeng, C., Wingreen, N., and Tang, C. to be published.
26 Park, B.H., and Levitt, M. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* 1995, **249,** 493–507
27 Eisenberg, D., and McLachlan, A.D. Solvation energy in protein folding and binding. *Nature* 1986, **319,** 199–203
28 Chan, H.S., and Dill, K.A. *J. Chem. Phys.* 1990, **92,** 3118
29 Shih, C.T., Su, Z.Y., Gwan, J.F., Hao, B.L., Hsieh, C.H., and Lee, H.C. Mean-field HP model, designability and alpha-helices in protein structures. *Phys. Rev. Lett.* 2000, **84,** 386–389
30 Kussell, E.L., and Shakhnovich, E.I. Analytic approach to the protein design problem. *Phys. Rev. Lett.* 1999, **83,** 4437–4440
31 Schuster, P., Fontana, W., Stadler, P.F., and Hofacker, I.L. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. B* 1994, **255,** 279–284
32 Skorobogatiy, M., Guo, H., and Zuckermann, M.J. A deterministic approach to protein design problem. *Macromolecules* 1997, **30,** 3403–3410
33 Ejtehadi, M.R., Hamedani, N., Seyed-Allaei, H., Shahrezaei, V., and Yahyanejad, M. Stability of preferable structures for a hydrophobic-polar model of protein folding. *Phys. Rev. E* 1998, **57,** 3298–3301
34 Ejtehadi, M.R., Hamedani, N., Seyed-Allaei, H., Shahrezaei, V., and Yahyanejad, M. Highly designable protein structures and inter-monomer interactions. *J. Phys. A* 1998, **31,** 6141–6155
35 Ejtehadi, M.R., Hamedani, N., and Shahrezaei, V. Geometrically reduced number of protein ground state candidates. *Phys. Rev. Lett.* 1999, **82,** 4723–4726
36 Buchler, N.E.G., and Goldstein, R.A. Effect of alphabet size and foldability requirements on protein structure designability. *Proteins* 1999, **34,** 113–124
37 Buchler, N.E.G., and Goldstein, R.A. Surveying determinants of protein structure designability across different energy models and amino-acid alphabets: A consensus. *J. Chem. Phys.* 2000, **112,** 2533–2547