Available online at www.sciencedirect.com
SCIENCE DIRECT®

Polymer 45 (2004) 699–705

polymer

www.elsevier.com/locate/polymer

Designability and thermal stability of protein structures

Ned S. Wingreen[a], Hao Li[b], Chao Tang[a,*]

[a]NEC Laboratories America, Inc., 4 Independence Way, Princeton, NJ 08540, USA
[b]Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA 94143, USA

Received 23 April 2003; received in revised form 25 June 2003; accepted 25 June 2003

Abstract

Only about 1000 qualitatively different protein folds are believed to exist in nature. Here, we review theoretical studies which suggest that some folds are intrinsically more designable than others, i.e. are lowest energy states of an unusually large number of sequences. The sequences associated with these folds are also found to be unusually thermally stable. The connection between highly designable structures and highly stable sequences is generally known as the 'designability principle'. The designability principle may help explain the small number of natural folds, and may also guide the design of new folds.

Keywords: Protein folding; Lattice models; Off-lattice models

1. Introduction

Two remarkable features of natural proteins are the simple fact that they fold and the limited number of distinct folds they adopt. Random amino-acid sequences typically do not fold to a unique structure. Rather they have many competing configurations of similar minimum free energy. Nature has evolved sequences that do fold stably, but it is estimated that the total number of qualitatively different folds is only about 1000 [1–3].

To attempt to explain these remarkable features of natural proteins, we have proposed a principle of designability [4–8]. The designability of a structure is the number of sequences which have that structure as their unique lowest-energy configuration [8]. In a wide range of models, structures are found to differ dramatically in designability, and sequences associated with highly designable structures have unusually high thermal stability [8–12]. We refer to this connection between the designability of a structure and the stability of its associated sequences as the designability principle. In model studies, highly designable structures are rare. As a result, thermally stable sequences are also rare. We conjecture that the designability principle also applies to real proteins, and that natural protein folds are exceptional, highly designable structures.

In this article, we review the designability principle. We start from a minimal model of protein structure in which the designability of a structure can be understood geometrically as the size of its basin of attraction in sequence space. More detailed models, including all 20 amino-acid types and off-lattice backbone configurations, reinforce the basic principle and provide a framework for the design of qualitatively new protein folds.

2. Purely hydrophobic (PH) model

Generally, the folding of proteins relies on the formation of a hydrophobic core of amino acids. Consideration of hydrophobicity alone leads to a very simple description of proteins—the 'purely hydrophobic' (PH) model [9]. In this model, sequences consist of only two types of amino acids, hydrophobic and polar [13]. Structures are compact walks on a cubic or square lattice. An example of a 6 × 6 square structure is shown in Fig. 1(a). As indicated, each structure consists of core sites surrounded by surface sites. The energy of a particular sequence folded into a particular structure is the number of hydrophobic amino acids occupying core sites, multiplied by − 1,

$$E = -\sum_{i=1}^{N} s_i h_i \quad (1)$$

A binary string $\{s_i\}$ represents each folded structure: $s_i = 1$

[*] Corresponding author. Tel.: + 1-609-951-2644; fax: + 1-609-951-2496.
E-mail address: tang@research.nj.nec.com (C. Tang).

0032-3861/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2003.10.062

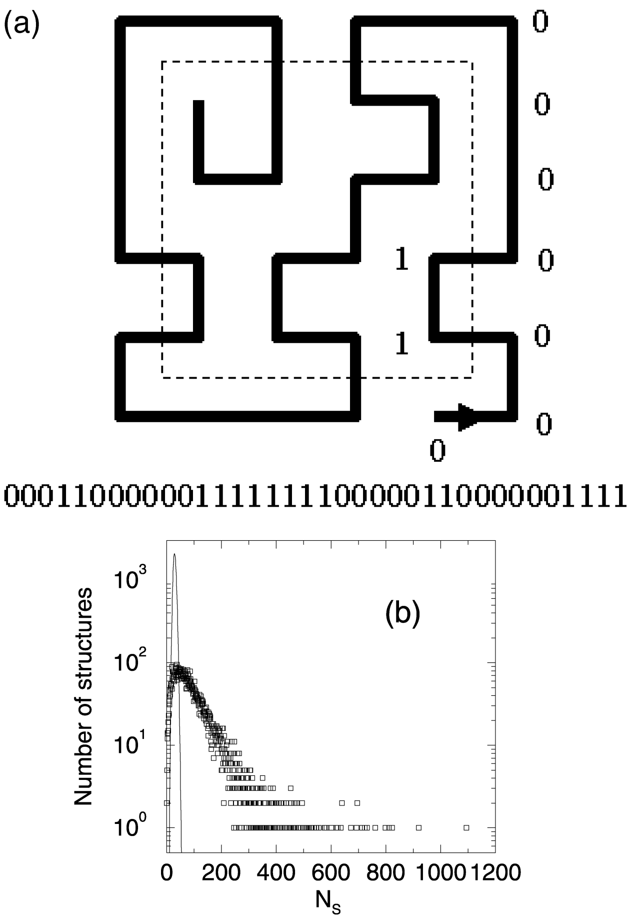

Fig. 1. (a) A 6 × 6 compact structure and its corresponding string. In the ‘purely hydrophobic’ (PH) model, only two types of sites are considered, surface sites and core sites. The core is shown enclosed by a dotted line. Each structure is represented by a binary string s_i ($i = 1, \ldots, 36$) of 0s and 1s representing surface and core sites, respectively. (b) Histogram of designability N_S for the 6 × 6 PH model, obtained using 19,492,200 randomly chosen sequences. A Poisson distribution with the same mean is shown for comparison.

if site i along the chain is in the core, and $s_i = 0$ if the site is on the surface. Similarly, a binary string h_i represents each sequence: $h_i = 1$ if the ith amino acid in the sequence is hydrophobic, and $h_i = 0$ if the amino acid is polar.

Within the PH model, structures differ dramatically in their designability. In practice, the designability is obtained by sampling a large number of binary sequences, and, for each sequence, recording the unique lowest-energy structure if there is one. Finally, the number of sequences which map to, i.e. ‘design’, each structure is summed to give the designability N_S of the structure. Fig. 1(b) shows a histogram of designability N_S for compact 6 × 6 structures. There are 57,337 structures, which map to the same number of structure binary strings $\{s_i\}$. However, since two or more structures can map into the same binary string (these ‘degenerate’ structures have zero designability), the number of distinct structure binary strings is smaller, being 30,408. Most structures have a designability around 50, but a small number of structures have designabilities more than 10 times this high. If sequences were randomly assigned to structures, the result would be the Poisson distribution which is shown for comparison, and there would be no structures with such high designability.

Importantly, the PH model has a simple geometrical representation that explains both the wide range of designabilities and the close connection between designability and thermal stability. To find the relative energies of different structures for a given sequence, the energy in Eq. (1) can be replaced by

$$E = \sum_{i=1}^{N} (h_i - s_i)^2 \qquad (2)$$

This replacement is allowed because the extra term $\sum h_i^2$ is a constant for a given sequence, and the other extra term $\sum s_i^2$ is also a constant, equal to the number of core sites, for all compact structures. Eq. (2) indicates that the energy of a sequence folded into a particular structure is simply the *Hamming distance* [14] between their respective binary strings. So, for a given sequence, the lowest-energy structure is simply the closest structure. The designability of a structure is therefore the exclusive volume of binary strings (sequences) that lie closer to it than to any other structure, as shown schematically in Fig. 2.

The wide range of structure designabilities can be traced to the nonuniform density of structures in the space of

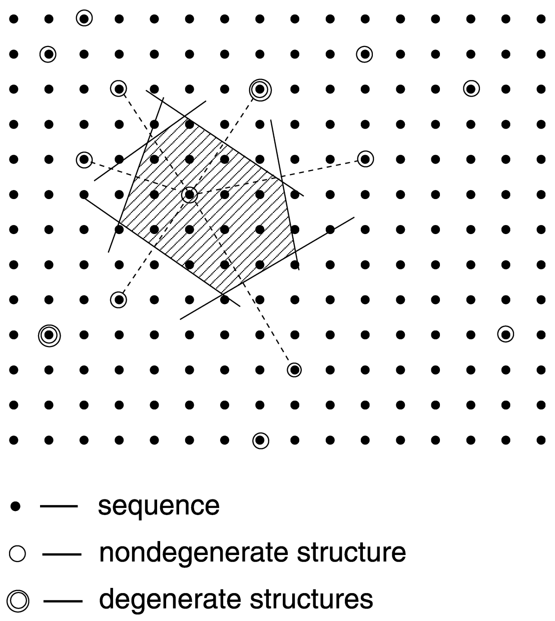

Fig. 2. Schematic representation of sequences and structures in the purely hydrophobic (PH) model. Dots represent sequences, i.e. all binary strings. Dots with circles represent binary strings associated with compact structures. Multiple circles indicate degenerate strings, i.e. strings associated with more than one compact structure. In the PH model, the energy of a sequence folded into a particular structure is the Hamming distance between their binary strings. Hence the number of sequences which fold uniquely to a particular structure—the designability of the structure—is the set of vertices lying closer to that structure than to any other, as indicated for one particular structure by the shaded region.

binary strings. Most structures are found in dense regions, i.e. in clusters of structures with similar patterns of surface and core sites. Structures found in these crowded regions have small exclusive volumes, and so, by definition, have small designabilities. In fact, many groups of distinct structures share an identical surface-core pattern (binary string) and therefore have zero designability. In contrast, a few structures fall in low-density regions, that is they have unusual surface-core patterns, and so have large exclusive volumes. These are the highly designable structures. In Fig. 3(a), we plot the number of structures $n(d)$ at a Hamming distance d from a structure with low, intermediate, and high designability, respectively. It shows that both low- and high-density neighborhoods typically have a large spatial extent, reaching nearly halfway across the space of binary strings.

The geometrical representation of the PH model makes clear the connection between thermal stability and designability. A sequence is considered to be thermally unstable if it has a small or vanishing energy gap between its lowest-energy structure and all other structures, and if there are many such competing structures. In the PH model, the energy of a sequence folded into a particular structure is the distance between their binary strings. A sequence which folds to a structure in a dense region (Fig. 2) will necessarily lie close to many other structures, and will therefore have many competing low-energy conformations. Even if the sequence perfectly matches the structure, with hydrophobic amino acids at all core sites and polar amino acids at all surface sites, the high surrounding density of structures with similar surface-core patterns implies a large number of competing folds. This is the hallmark of thermal instability. Therefore, the low-designability structures, which are found in high-density regions, will have associated sequences which are thermally unstable.

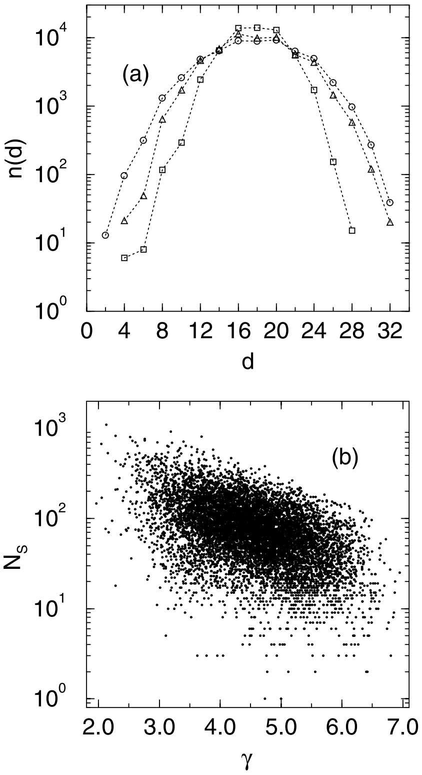

Fig. 3. (a) Number of neighboring structures $n(d)$ versus distance d to neighbors for three representative 6×6 structures, with low (circles), intermediate (triangles), and high (squares) designability. The distance between structures is defined as the Hamming distance between their binary strings. (b) Designability versus γ, the variance of $n(d)$, for all 6×6 structures.

In contrast, if a sequence folds to a structure in a low-density region, there will be relatively few nearby structures, and so relatively few competing folds. These sequences will be thermally stable. Therefore, the highly designable structures, from low-density regions, will have associated sequences of high thermal stability. This is the designability principle in a nutshell—high designability and thermal stability are connected because both arise from low-density regions in the space of binary strings which represent folded structures.

A measure of the 'neighborhood density' of structures around a particular structure is the variance γ of the quantity $n(d)$ shown in Fig. 3(a). The variance γ is directly related to the thermal stability—smaller γ implies lower neighhood density and hence higher thermal stability. In Fig. 3(b) we plot this variance as a function of designability. It shows a strong correlation between the designability and the thermal stability.

Since low-energy competing structures also represent kinetic traps, one expects the thermally stable sequences associated with highly designable structures to be fast folders as well. This has been tested for a lattice model closely related to the PH model [15].

3. Miyazawa–Jernigan (MJ) matrix model

Natural proteins contain 20 amino acids, not two, and their interactions are more complicated than simple hydrophobic solvation. Some of these real-world features are captured in the Miyazawa–Jernigan (MJ) matrix model. The MJ matrix is a set of amino-acid interaction energies inferred from the propensities of different types of amino acids to be neighbors in natural folded structures [16]. The model assigns the appropriate energy from the MJ matrix to every pair of amino acids that are on neighboring lattice sites, but are not adjacent (covalently bonded) on the chain, as indicated in Fig. 4(a) [11]. Thus, in the model, a sequence of length N is specified by the residue type μ_i, ($i = 1, 2, \ldots, N$) along the chain, where μ is one of the 20 natural amino acids. A structure is specified by the position $\mathbf{r}_i$, ($i = 1, 2, \ldots, N$) of each residue along the chain. The energy for a sequence folded into a structure is taken to be the sum of the contact energies, that is

$$E = \sum_{i<j} e_{\mu_i \mu_j} \Delta(\mathbf{r}_i - \mathbf{r}_j), \quad (3)$$

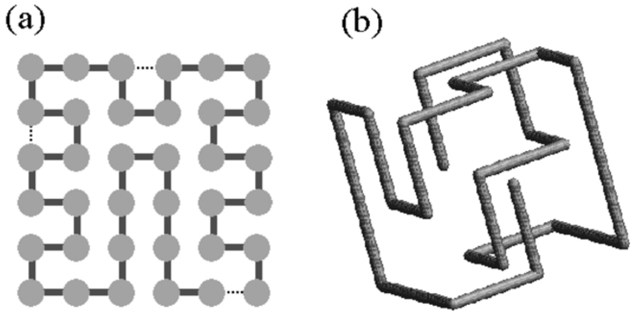

Fig. 4. (a) Most designable 6 × 6 structure using 20 amino-acid types. Only noncovalent nearest-neighbor interactions contribute to the energy, as indicated by dashed lines for a few pairs. Interaction energies are taken from the Miyazawa–Jernigan (MJ) matrix. (b) Most designable 3 × 3 × 3 structure using the same MJ-matrix energies.

where $e_{\mu_i\mu_j}$ is the contact energy (MJ matrix) between residue types μ_i and μ_j, and $\Delta(\mathbf{r}_i - \mathbf{r}_j) = 1$ if $\mathbf{r}_i$ and $\mathbf{r}_j$ are adjoining lattice sites with i and j not adjacent along the chain, and $\Delta(\mathbf{r}_i - \mathbf{r}_j) = 0$ otherwise. In studies using the MJ-matrix model, there are generally too many possible sequences (20^N) to enumerate, but the relative designabilities of structures can be obtained accurately by random sampling of sequences.

Fig. 5(a) shows a histogram of designability for compact 6 × 6 structures obtained using the MJ matrix of interaction energies. The form of the distribution is very similar to the PH-model histogram, including the tail of highly designable structures (Fig. 1(b)). There is also a strong correlation between thermal stability and designability in the MJ-matrix model [11]. For thermal stability, one can use some measure of 'neighborhood' density of states. We find that in the MJ-matrix model, the thermal stability of a sequence folded into a structure is well correlated with the local energy gap [17] between the lowest-energy structure and the next lowest. Fig. 5(b) shows the energy gap averaged over sequences which fold to structures of a given designability N_S. With increasing designability, there is a clear increase in the average gap, and hence in the thermal stability of associated sequences. The results of the MJ-matrix model are very similar to those obtained with the PH-model. Indeed, the same structures are found to be highly designable in both models, including the same top structure shown in Fig. 4(a). The most designable 3 × 3 × 3 structure is shown in Fig. 4(b). Qualitatively, the results of the MJ-matrix model are the same for three-dimensional structures (Fig. 6) as for two-dimensional ones.

Why are the results of the purely hydrophobic model and the Miyazawa–Jernigan-matrix model so similar? In fact, both models are dominated by hydrophobic solvation energies. The interaction energy between any two amino acids i and j in the MJ matrix can be well approximated by $-(h_i + h_j)$, where h_i is an effective hydrophobicity for each

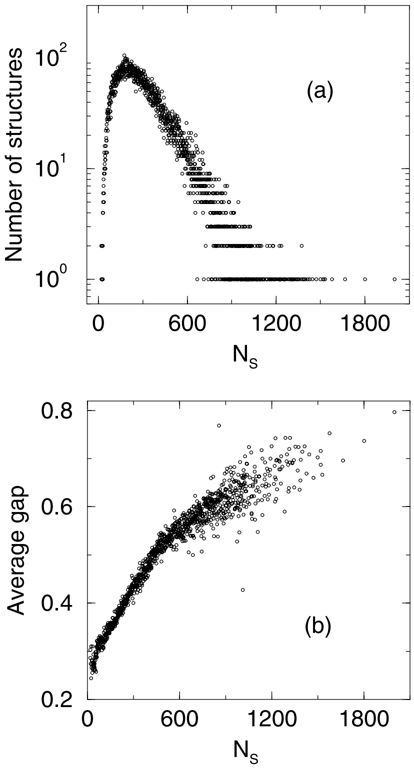

Fig. 5. (a) Histogram of designability N_S for the 6 × 6 MJ-matrix model. (b) Average gap versus designability for the 6 × 6 MJ-matrix model. Data obtained using 9,095,000 randomly chosen sequences.

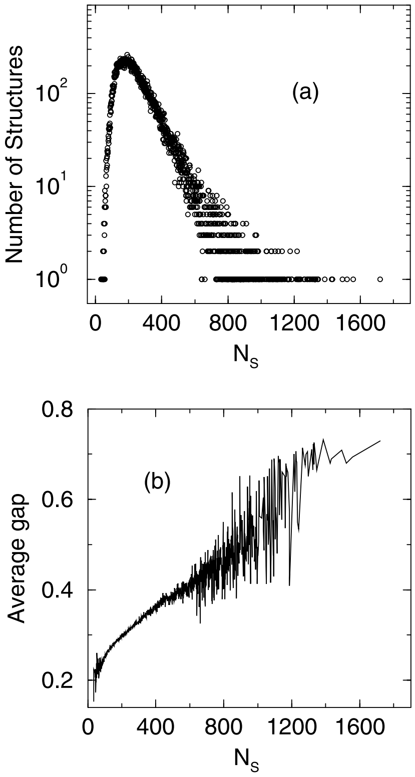

Fig. 6. (a) Histogram of designability N_S for the 3 × 3 × 3 MJ-matrix model. (b) Average gap versus designability for the 3 × 3 × 3 MJ-matrix model. Data obtained using 13,550,000 randomly chosen sequences.

amino acid [18]. This implies that the energy of formation of a noncovalent nearest-neighbor pair is simply the desolvation energy of shielding one face of each amino acid from the surrounding water. As a result, the MJ-matrix model can be viewed as a variant of the PH-model in which there are 20 possible values of hydrophobicity instead of just two. An additional distinction is that the MJ-matrix model has a range of different site types (core, edge, and corner in two dimensions; core, edge, face, and corner in three dimensions) rather than just surface and core as in the PH model. Overall, these differences are not enough to alter the basic designability principle, or even to change the set of highly designable 6 × 6 structures.

For the MJ-matrix model, one can still construct a space of structure strings, now including several levels of solvent exposure between 0 (most exposed) and 1 (most buried). As in the PH model, some regions of this space are dense with structures and some have few structures. Structures with similar surface-exposure strings compete for sequences. As a result, structures in high-density regions have small basins of attraction for sequences, and structures in low-density regions have large basins. Moreover, sequences associated with structures in low-density regions have few competing conformations and are the most thermally stable. Therefore, the designability principle holds in the MJ-matrix model for the same reason it does in the PH model: high designability and high thermal stability are connected because both arise in low-density regions in the space of strings, i.e. the space of surface-exposure patterns of structures.

Lattice–protein models in which hydrophobic solvation does not dominate may show different behavior. For example, Buchler and Goldstein reported results of a variant of the MJ-matrix model in which the dominant hydrophobic term $-(h_i + h_j)$ had been subtracted out [19]. They found a set of highly designable structures different from that obtained with the full MJ matrix, and similar to the set obtained for a random pairing potential between amino acids.

4. Off-lattice models

Natural proteins fold in three dimensions, and their main degrees of freedom are bond rotations. Does the designability principle extend to off-lattice models with more realistic degrees of freedom? One model for which designability has been studied off-lattice is a 3-state discrete-angle model, of the type introduced by Park and Levitt [20]. The results strongly confirm the designability principle, and suggest the possibility of creating new, highly designable folds in the laboratory [10].

The main degrees of freedom of a protein backbone are the dihedral angles ϕ and ψ. Certain pairs of ϕ–ψ angles are preferred in natural structures, since they lead to conserved secondary structures such as α-helices and β-strands [21]. Discrete-angle models for protein structure take advantage of these preferences by allowing only certain combinations of angles. For an m-pair model, the total number of backbone structures grows as m^N. With $m = 3$, it is possible to computationally enumerate all structures up to roughly $N = 30$ amino acids.

Fig. 7(a) shows an example of a protein backbone of length $N = 23$ generated using a 3-state model with $(\phi, \psi) = (-95, 135)$, $(-75, -25)$, and $(-55, -55)$, where the first pair of angles corresponds to a β-strand and the second two correspond to variants of α-helices. Structures are decorated with spheres representing sidechains, as shown in Fig. 7(b). Only compact self-avoiding structures are considered as possible protein folds.

To assess designability of these off-lattice structures, the solvent-exposed area of each sidechain sphere is evaluated. An energy of hydrophobic solvation is defined as in Eq. (1) by $E = -\sum_{i=1}^{N} s_i h_i$, where now s_i is the fractional exposure to solvent of the ith sidechain, and the h_i are amino-acid hydrophobicities. Fig. 8(a) shows a histogram of designability for the 3-state model. There is a wide range of designabilities, with a tail of very highly designable structures. A strong correlation exists between the designability of a structure and the thermal stability of its associated sequences, as shown in Fig. 8(b).

The designability principle evidently applies to the 3-state model as well as to the lattice models discussed earlier. This is not surprising, because folding in the 3-state model is also driven by hydrophobic solvation. Each structure in the model is represented by a string of sidechain solvent exposures, represented by real numbers between 0 and 1. Again, the space of these strings has high- and low-density regions, with the, by now familiar, relation between low density and high designability and thermal stability leading to the designability principle.

A major advantage of the 3-state model is that it addresses structures that a real polypeptide chain can adopt. Among the highly designable folds, one recovers several recognizable natural structures, including an α-turn-α fold and a zincless zinc-finger. In addition, some of the highly designable folds, including a β–α–β structure, have not been observed in nature as independent domains. Results of our effort to create this fold in the laboratory are encouraging [22].

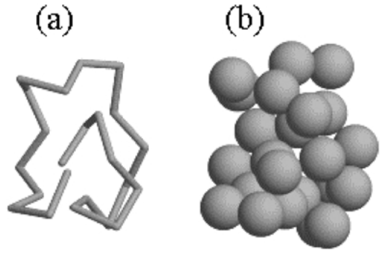

Fig. 7. (a) Example of a compact, self-avoiding 23-mer backbone generated using three dihedral-angle pairs. (b) Backbone with generic sidechain spheres centered on C_α positions.

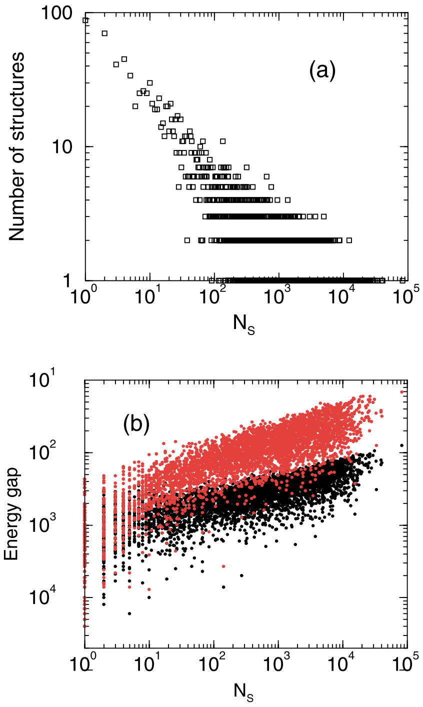

Fig. 8. (a) Histogram of designability for 23-mer off-lattice structures of the type shown in Fig. (7). (b) Average energy gap (black dots) and largest energy gap (red dots) versus designability. Data generated by enumeration of all binary sequences.

5. Discussion and conclusion

The designability principle has been explored in a number of models for proteins, including all 20 amino acids and realistic backbone conformations. In these models, the strong link between designability and thermal stability can be traced to the dominance of the hydrophobic solvation energy. Whenever hydrophobicity is dominant, each structure can be reduced to its pattern of solvent exposure. In the same vein, each sequence can be reduced to its pattern of hydrophobicity. Sequences will fold so as to best match their hydrophobic residues to the buried sites of structures. Both designability (number of sequences per structure) and thermal stability depend on a competition among structures with similar patterns of solvent exposure. Highly designable structures are those with unusual patterns of surface exposure, and therefore with few competitors. This lack of competitors also implies that the sequences folding to highly designable structures are thermally stable.

Since hydrophobicity is generally accepted to be the dominant force for folding of real proteins, the designability principle may provide a guide to understanding the selection of natural protein structures. Of course, real proteins are held together by forces other than hydrophobicity. Next to hydrophobicity, the formation of hydrogen bonds is the most important factor in determining how a typical protein

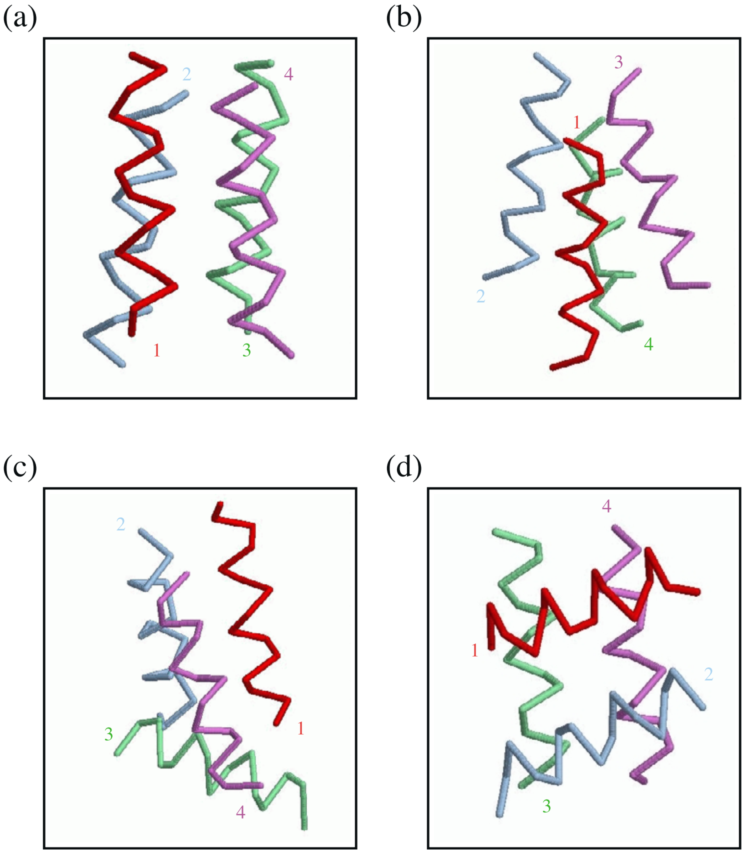

Fig. 9. Four most designable four-helix bundles generated by packing tethered 15-residue α-helices. The helices are numbered at their N-terminals.

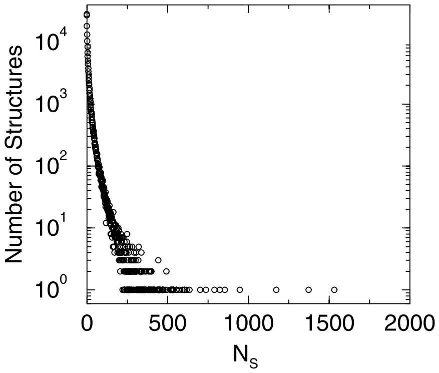

Fig. 10. Histogram of designability N_S for four-helix bundles. Data obtained using 2,000,000 randomly chosen binary sequences.

folds. The backbone hydrogen bonds of α-helices and β-sheets help stabilize particular folds. These secondary structures can be incorporated within the framework of designability as a favorable energy bias for formation of α-helices and β-sheets.

One way to incorporate hydrogen bonding in the design of new protein folds is to specify in advance the secondary structure of the protein. This approach has the added advantage of greatly reducing the number of degrees of freedom. The desired secondary structures can be designed into the sequence via the propensities of particular amino acids to form α-helices and β-strands.

This approach to design was recently carried out for four-helix bundles [12]. Compact, self-avoiding structures consisting of four tethered 15-residue α-helices were generated and assessed for designability. Fig. 9 shows the four most designable distinct folds, which closely correspond to natural four-helix bundles. As shown in Fig. 10, the histogram of designability for the four-helix model has the characteristic long tail of highly designable structures.

The principle of designability has been motivated here in terms of hydrophobic solvation. More generally, the dependence of both designability and thermal stability on a competition among structures broadens the application of the principle. For example, designability and thermal stability have been found to correlate in nonsolvation models including random-interaction models [19] and folding of two-letter RNA [23]. In the future, we hope that designability will provide a guide to the design of new structures both for polymers other than proteins and for solvents other than water.

Acknowledgements

We gratefully acknowledge the contributions of many coworkers in developing the notion of designability, in particular Eldon Emberly, Robert Helling, Régis Mélin, Jonathan Miller, Tairan Wang, and Chen Zeng.

References

[1] Chothia C. Nature 1992;357:543–4.

[2] Orengo CA, Jones DT, Thornton JM. Nature 1994;372:631–4.

[3] Brenner SE, Chothia C, Hubbard TJP. Curr Opin Struct Biol 1997;7: 369–76.

[4] For a broad review of designability, seeHelling R, Li H, Mélin R, Miller J, Wingreen N, Zeng C, Tang C. J Mol Graph Model 2001;19: 157–67.

[5] Finkelstein AV, Ptitsyn OB. Prog Biophys Mol Biol 1987;50:171–90.

[6] Yue K, Dill KA. Proc Natl Acad Sci USA 1995;92:146–50.

[7] Govindarajan S, Goldstein RA. Biopolymers 1995;36:43–51.

[8] Li H, Helling R, Tang C, Wingreen NS. Science 1996;273:666–9.

[9] Li H, Tang C, Wingreen NS. Proc Natl Acad Sci USA 1998;95: 4987–90.

[10] Miller J, Zeng C, Wingreen NS, Tang C. Proteins 2002;47:506–12.

[11] Li H, Tang C, Wingreen NS. Proteins 2002;49:403–12.

[12] Emberly E, Wingreen NS, Tang C. Proc Natl Acad Sci USA 2002;99: 11163–8.

[13] Lau KF, Dill KA. Macromolecules 1989;22:3986–97.

[14] The Hamming distance between two binary strings, s_i and t_i, where $i = 1, 2, \ldots, N$, is defined as $\sum_{i=1}^{N} |s_i - t_i|$.

[15] Mélin R, Li H, Wingreen NS, Tang C. J Chem Phys 1999;110: 1252–62.

[16] Miyazawa S, Jernigan RL. Macromolecules 1985;18:534–52.

[17] Sali A, Shakhnovich E, Karplus M. Nature 1994;369:248–51.

[18] Li H, Tang C, Wingreen NS. Phys Rev Lett 1997;79:765–8.

[19] Buchler NEG, Goldstein RA. J Chem Phys 2000;112:2533–47.

[20] Park BH, Levitt M. J Mol Biol 1995;(249):493–507.

[21] Ramachandran GN, Sasisekharan V. Adv Protein Chem 1968;23: 283–438.

[22] Fan K, Zeng C, Jing C, Wingreen NS, Lai L, Tang C. To be published.

[23] Mukhopadhyay R, Emberly E, Tang C, Wingreen NS. Phys Rev E 2003;68:041904.