Emergence of Highly Designable Protein-Backbone Conformations in an Off-Lattice Model

Jonathan Miller, Chen Zeng, Ned S. Wingreen, and Chao Tang*
NEC Research Institute, Princeton, New Jersey

ABSTRACT **Despite the variety of protein sizes, shapes, and backbone configurations found in nature, the design of novel protein folds remains an open problem. Within simple lattice models it has been shown that all structures are not equally suitable for design. Rather, certain structures are distinguished by unusually high designability: the number of amino acid sequences for which they represent the unique lowest energy state; sequences associated with such structures possess both robustness to mutation and thermodynamic stability. Here we report that highly designable backbone conformations also emerge in a realistic off-lattice model. The highly designable conformations of a chain of 23 amino acids are identified and found to be remarkably insensitive to model parameters. Although some of these conformations correspond closely to known natural protein folds, such as the zinc finger and the helix-turn-helix motifs, others do not resemble known folds and may be candidates for novel fold design. Proteins 2002;47:506–512.**

Key words: protein folds; off-lattice model; designability; protein design; evolution

INTRODUCTION

The de novo design of proteins—an object of enormous activity in recent years—has so far dealt primarily with the *re*design of known protein folds.[1–8] Two major accomplishments in the direction of designing a fold that is distinct from known natural folds are the synthesis of a right-handed coiled coil[9] and the synthesis of a zinc finger without zinc.[10–12] To challenge the best efforts of de novo design, nature offers roughly 1000 qualitatively distinct protein folds.[13] Why has it proven difficult to design new protein folds? What program should we follow to achieve ab initio design of novel folds?

The principle of designability[14–19] offers an answer to both these questions for simple lattice models. The designability of a structure is measured by the number of sequences that design it, that is, the number of sequences that have the given structure as their unique lowest energy conformation. Structures can differ vastly in their designability,[14] and it has been shown that high designability entails other protein-like properties, such as mutational stability, thermodynamic stability,[14,15] and fast folding kinetics.[16,20] Design is hard in the sense that most structures have low designability and their associated sequences lack these protein-like properties. For successful de novo design, one should first identify the few highly designable structures.

It is an open question whether designability applies to real proteins as it does to lattice polymers. Real protein structures have a degree of complexity that cannot be effectively represented within a simple lattice model. For example, on a lattice the angles between bonds differ from those naturally adopted in real proteins. In addition, although in a cubic-lattice model the cube minimizes surface area for a given volume and is perfectly packed, no counterpart of the perfect cube exists once the lattice is removed. For designability to guide practical design of new folds it must apply to realistic descriptions of protein structure.

In this article we report the computation of designability within an off-lattice model that incorporates angles favored by natural proteins, for protein chains of up to $N = 23$ amino acids. We find that the essential qualitative features of designability survive the transition from lattice model to off-lattice model. In particular, it remains true that a small fraction of compact structures are highly designable: these are nondegenerate ground states for an enormous number of amino acid sequences. Most structures, on the other hand, are ground states for few, if any, amino acid sequences. Furthermore, the sequences that fold into highly designable structures typically have enhanced thermodynamic stability—the energy of the nearest excited state is separated from the ground-state energy by an appreciable gap.

MODELS AND METHODS

The model we adopt is closely related to the off-lattice, m-state discrete-angle model introduced by Park and Levitt.[21] Each configuration is defined by a sequence of C_α bonds of length 3.8 Å, and each pair of dihedral angles (ϕ, ψ) is restricted to one of only m alternatives; here we take $m = 3$. The set of m allowed angle pairs is chosen by fitting to the backbone coordinates of representative natural proteins,[21] as discussed below. To suppress self-intersections of the chain, we augment the model by introducing a

C. Zeng's present address is Department of Physics, George Washington University, Washington, DC 20052.

*Correspondence to: Chao Tang, NEC Research Institute, 4 Independence Way, Princeton, NJ 08540. E-mail: tang@research.nj.nec.com

Received 23 October 2001; Accepted 7 January 2002

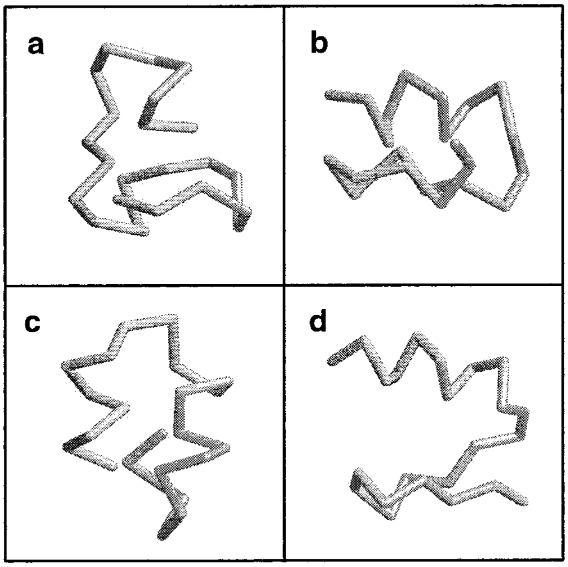

Fig. 1. **a–c**: Backbone configurations of 1st, 4th, and 15th most designable 23-mer structures. **d:** Backbone configuration of the zinc finger 1PSV,[12] truncated to 23 amino acids.

volume for the amino acid residues in the form of a sphere of radius r_β centered on C_β (the first carbon of the side-chain). The backbones of some configurations constructed in this fashion are shown in Fig. 1(a–c).

This off-lattice model incorporates properties of real polymers not well reproduced in simple lattice models. On the lattice, for example, allowed ground-state structures were limited to those maximally compact structures that fill the unique rectangle or box of minimum surface area. Off the lattice, every structure can be expected to have a distinct surface area. However, open or extended structures are not expected to be designable. We entertain as plausible ground-state structures only those with a surface area below some cutoff value A_c, which enters our computation as a parameter.*

Because a discrete angle set represents only a crude approximation to a continuum of angles, it is unrealistic to expect the surface area of a discrete-angle structure to faithfully reproduce the surface area of a structure built from more flexible angles. Importantly, using flexible angles would allow our more open structures (e.g., those just below the cutoff A_c) to contract and reduce their exposed surface areas. To achieve this equalizing effect of a continuum of angles within the limitations of a discrete-angle model, we normalize the vector of solvent-accessible surface areas $\mathbf{A} = (a_1, \ldots, a_N)$, where a_i is the solvent-accessible surface area of the i-th residue, in such a way as to preserve the pattern of surface exposure along a chain. A suitable procedure† is to normalize the vector $\mathbf{A}$ for each structure by the total exposed surface area of that structure: $\bar{\mathbf{A}} = \mathbf{A}/\Sigma_i a_i = (\tilde{a}_1, \ldots, \tilde{a}_N)$. This procedure treats all structures below the cutoff A_c as equally compact while preserving each structure's individual pattern of surface exposure along the chain.

As with real proteins, description and comparison of configurations off-lattice demands precision about what we mean by the term "structure." For example, a protein structure obtained by NMR represents an ensemble of configurations, no element of which necessarily provides a better fit to the data than any other. This ensemble presumably reproduces the temperature-induced fluctuations of a natural protein around its native state. On averaging over this ensemble for small stably folded polypeptides in the PDB database, one finds a typical center-of-mass root mean square (crms) of roughly 0.3–0.5 Å per residue. A similar range of crms can be inferred from the B values of protein crystals.[23] Accordingly, our off-lattice polymer configurations are grouped into clusters consisting of all configurations lying within a crms distance λ per residue of one another. Configurations within a cluster are to be thought of as variations of a single structure, and subsequently we will refer to clusters and structures interchangeably.

We define the designability of a structure as the sum of the designabilities of its included configurations. The designability of a configuration is simply the number of sequences with that configuration as a unique ground state.[14,15] To evaluate the energy of a sequence on each configuration, we associate a hydrophobicity h_i with each amino acid of the sequence. In practice, we assign a hydrophobicity which is either 0 (Polar) or 1 (Hydrophobic) to each monomer to create an HP-sequence[24]; that this is a reasonable simplification finds support in the work of Beasley and Hecht[1] [cf. Fig. 3(e) for the results of a more general choice]. The energy of a particular sequence folded into a particular configuration is obtained by taking the sum of the products of each amino acid's hydrophobicity h_i with its normalized surface exposure $\tilde{a}_i$,

$$E = \sum_i h_i \tilde{a}_i \tag{1}$$

We numerically evaluate the energy of all HP-sequences for all configurations.

Except as indicated explicitly in the text, we choose discrete angles and the amino acid radius to optimize the fit to the backbone of the zinc-less synthetic zinc finger[12] 1PSV [Fig. 1(d)]. We find that there are many angle sets that fit the backbone of 1PSV almost equally well. For example, the crms per residue between 1PSV and the structure obtained from each of our 10 best angle sets varies from 0.844 to 0.913 Å. The angle set we use for most

*We evaluate the area of each C_β sphere accessible to a probe sphere of radius 1.4 Å, by the methods used in the program SERF,[22] the slightly different values of surface area obtained by different methods do not in any way alter the outcome of the calculations.

†We have checked that certain alternative normalizations (e.g., normalizing by the total solvent-inaccessible surface area) do not alter the set of highly designable structures that emerge from our calculation. With no normalization, higher designability becomes closely correlated with lower solvent-accessible surface area.

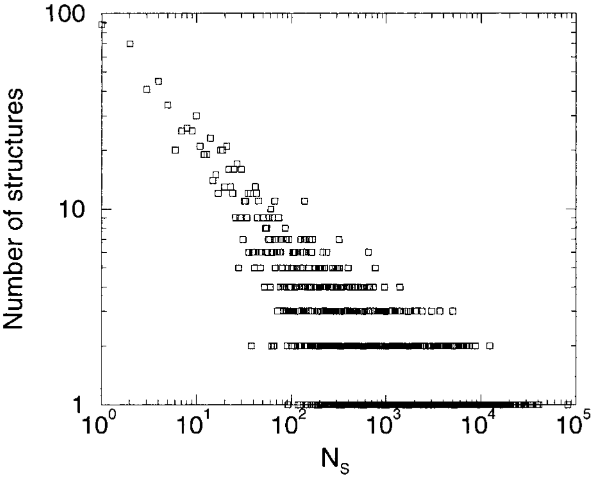

Fig. 2. Histogram of designabilities of 23-mer structures, using r_β = 1.9 Å. The surface area cutoff A_c is such that 10,000 configurations participate in the calculation, grouped into 4688 clusters with cluster radius λ = 0.4 Å.

of the calculations presented in this article is (ϕ, ψ) = (−95°, 135°), (−75°, −25°), and (−55°, −55°). The first pair lies in the β-region of the Ramachandran plot, and the other two pairs lie in the α-region. We take $r_\beta = 1.9$ Å, the radius above which the amino acids fit to the backbone of 1PSV would clash.

RESULTS

The designability of a structure denotes the number of distinct HP-sequences having that structure as their unique ground state. The distribution of designabilities for our model, displayed in Figure 2, reproduces a crucial feature first observed on the lattice: although most structures have very low designability, the trailing edge (or tail) of the distribution consists of a small number of structures of very high designability. Thus, designability distinguishes a small subset of structures from generic ones.

It turns out that the identities of these highly designable structures depend only weakly on the values of the parameters that enter our calculation: the surface area cutoff A_c, clustering radius λ, side-chain radius r_β, the set of allowed dihedral angles, and the range of amino acid hydrophobicities. More specifically, a significant fraction of structures identified as highly designable for one set of parameter values remains highly designable when these parameters are varied. We provide evidence for this important observation in the next five subsections.

Surface Area Cutoff

As discussed before, open structures are expected to exhibit low designability. We anticipate that the highly designable structures of interest to us will fall mainly within the class of compact structures; therefore, only these compact structures are needed in our calculation. The surface area cutoff A_c determines how compact a structure must be to qualify. We expect that, provided the choice of A_c is not too restrictive, its particular value ought not to be important.

A computationally practical choice of the surface-area cutoff eliminates most of the less compact configurations. A few of these might have proven highly designable if retained; however, our objective is not to find all highly designable structures, but only to identify some of them. Therefore, our major concern is not that we might incorrectly discard a few designable structures, but rather that we might produce false positives (structures that appear to be highly designable with a restrictive value of the cutoff but have low designability for a more relaxed cutoff). A larger cutoff admits previously disallowed configurations that "steal" some sequences from a configuration originally identified as highly designable, thereby reducing its designability.

In practice, as shown in Figure 3(a), highly designable structures tend to remain highly designable with increasing surface-area cutoff. For example, 9 of the 10 most designable structures remain within the 100 most designable even after the surface-area cutoff is relaxed sufficiently to admit a 10-fold increase in the number of participating structures.

Clustering Radius

As discussed in the previous section, structures whose backbones differ insignificantly from one another ought not to be considered distinct. This observation is embodied in our calculation by grouping into clusters those structures whose backbone configurations lie within a certain crms distance, λ, of one another. Varying the clustering radius, λ, leaves unchanged the set of configurations that participate in the calculation. For λ ≤ 0.1 Å, nearly every cluster consists of a unique configuration. To exhibit the dependence of the most designable structures on λ, we fix a configuration and follow the designability of the cluster to which that configuration belongs, as a function of λ. As shown in Figure 3(b), the most designable structures remain roughly the same as λ is varied over a wide range.

Side-Chain Radius

Excluded volume is incorporated by means of a hard sphere of radius r_β centered on the β-carbon of each amino acid. Increasing the side-chain radius r_β eliminates some configurations because of steric clashes, whereas decreasing r_β admits previously ineligible configurations. Starting at $r_\beta = 1.9$ Å, we identify the most designable structures and then count the fraction of these structures that remain highly designable as r_β is reduced. As shown in Figure 3(c), the identities of the most designable structures are well preserved.

Set of Dihedral Angles

Next, we address to what extent an outcome depends on a particular choice of the discrete set of dihedral angles. A discrete set of angles cannot sample the structure space fully and so cannot "hit" all possible structures. On the other hand, we know that the designability of a structure depends on the local density of solvent-exposure vectors $\tilde{\mathbf{A}}$ with highly designable structures occupying the lowest density regions.[15] If the subset of structures sampled by a discrete set of angles reasonably preserves density in the

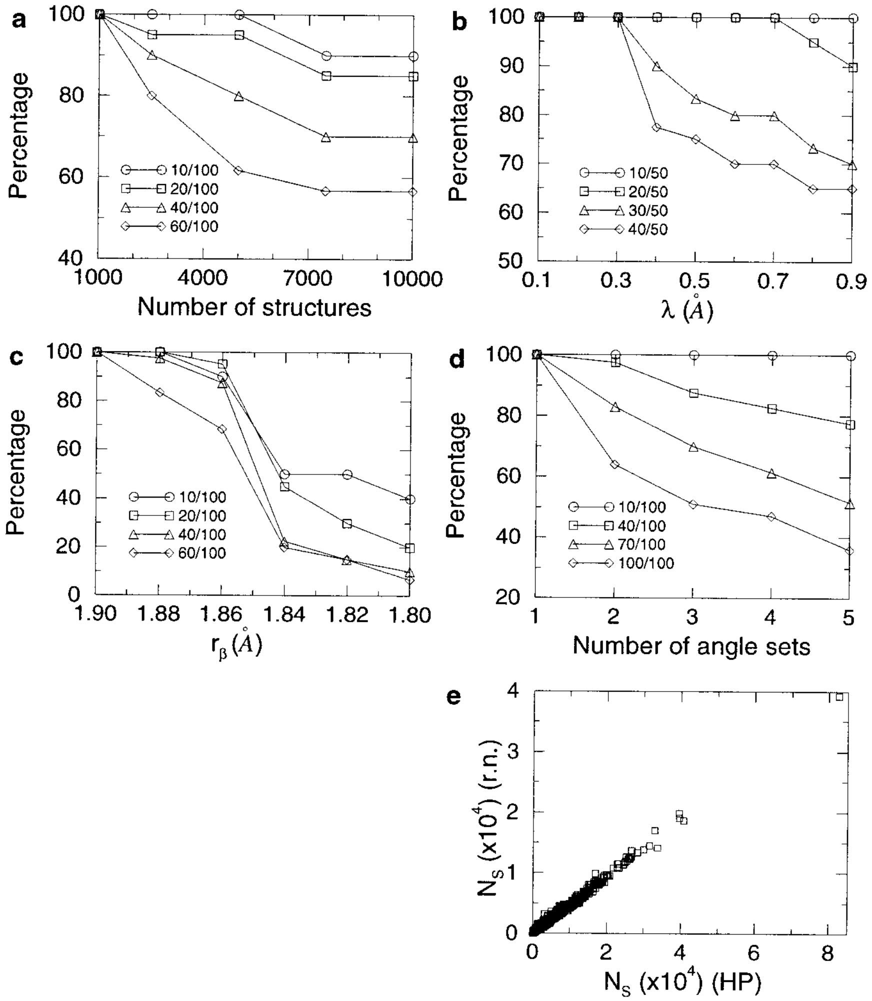

Fig. 3. Sensitivity to parameter changes of the most designable structures from Figure 2. **a:** Fraction of the 10, 20, 40, or 60 most designable structures that remain in the 100 most designable as the surface-area cutoff increases. The initial cutoff A_c is chosen so that only the 1000 most compact configurations participate and A_c increases until 10,000 configurations participate. **b:** Fraction of the 10, 20, 30, or 40 most designable structures that remain in the 50 most designable as the clustering radius λ is increased. The 5000 most compact configurations participate in the calculation and r_β = 1.9 Å. **c:** Fraction of the 10, 20, 40, or 60 most designable structures that remain in the 100 most designable as the side-chain radius r_β is changed. We have chosen the surface area cutoff so that 5000 structures participate in the designability calculation for r_β = 1.9 Å. If some configurations of the original most designable structures are not among the 5000 most compact configurations for some smaller r_β, we nevertheless retain them in the calculation. The clustering radius is λ = 0.4 Å. **d:** Fraction of the 10, 40, 70, or 100 most designable structures that remain in the 100 most designable as configurations from other angle sets are added. The values of the five angle sets are as follows set #1 = (−95°, 135°), (−75°, −25°), (−55°, −55°); set #2 = (−95°, 135°), (−85°, −55°), (−65°, −25°); set #3 = (−105°, 145°), (−85°, −15°), (−75°, −35°); set #4 = (−105°, 145°), (−85°, −35°), (−85°, −5°); set #5 = (−105°, 145°), (−85°, −35°), (−85°, −15°). **e:** Designability of structures obtained from 4,000,000 randomly generated sequences of real numbers in [0,1] versus designability from enumeration of HP-sequences. The 10000 most compact configurations participate in the calculation, λ = 0.4 Å, and r_β = 1.9 Å. (*Note:* the suppressed zeros in panels a, b, and d.)

space of structures, highly designable structures should remain highly designable as we improve our sampling of structure space.

To examine this possibility, we identify configurations generated by one angle set and follow their cluster designabilities as configurations from other angle sets are added. We take five different angle sets derived from fitting to 1PSV, and use the most compact configurations generated by each set. We calculate the designability of structures by using configurations from, respectively, one, two, three, four, and finally all five sets. We observe in Figure 3(d) that the most designable structures in set #1 remain highly designable even as configurations from sets #2, #3, #4, and #5 are added. This result is maintained under permutation of the five sets. Apparently, any reasonable choice of angle set covers the structure space sufficiently well that highly designable structures can be identified with high probability.

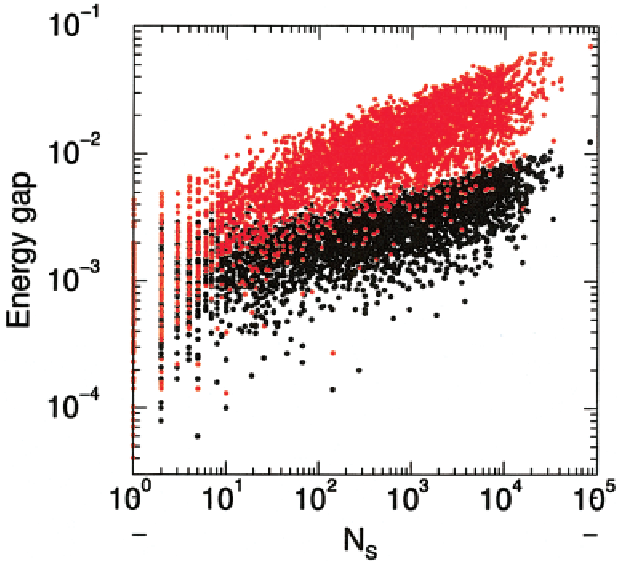

Fig. 4. Maximum energy gap (red dots) and average energy gap (black dots) for the HP-sequences that design a given structure, plotted versus structure designability. The 10,000 most compact configurations of the 23-mer participate in the calculation, with $\lambda = 0.4$ Å and $r_\beta = 1.9$ Å.

HP Sequences

To check whether the identification of designable structures depends on our use of HP (binary) sequences of amino acids, we recalculate designabilities by using amino acids with continuous real-valued hydrophobicities. We randomly choose 4,000,000 sequences $\mathbf{h} = (h_1, \ldots, h_N)$, where $h_i \in [0,1]$, and evaluate their energy for all configurations using Eq. (1). In Figure 3(e) we plot the designability calculated this way against that from the enumeration of HP sequences. As the figure shows, the highly designable structures computed by these two alternative methods are nearly identical.

Parameter Independence

In the preceding five subsections we have shown that the parameters can sustain a considerable degree of variation without significantly changing the outcome of the designability calculation. The weak dependence of the set of highly designable structures on parameters is illustrated in Figure 3. Because the identity of the highly designable structures is robust to parameter variation, we now examine their potential as candidates for design.

Gap

In particular, a prerequisite for design is believed to be the presence of a large separation between the ground-state energy and the energy of the lowest excited state. For each structure, we have identified the HP-sequence that makes this gap the largest. The value of this largest gap is shown in Figure 4, as a function of the designability of the structure. To convert the vertical scale of Figure 4 to real energies, we observe that one unit of energy corresponds to a sequence of exclusively hydrophobic amino acids ($h_i = 1$) folded into one of our typical compact structures. Our choice of surface area cutoff A_c guarantees that a typical compact configuration has around half of its maximal accessible surface exposed (about 25 Å^2 per residue). A conservative estimate for the energy of exposed surface,[23] 20 cal/Å^2/mol, then yields an energy on the order of 10 kcal/mol for a 23-mer. The highest gap energies achieved in Figure 4, of order 0.05, therefore correspond to a gap of 0.5 kcal/mol, around $k_\beta T$ for room temperature. This gap is roughly the energy to promote one hydrophobic amino acid from core to surface. Also plotted is the average gap for all HP-sequences that design a structure. It is evident that high designability correlates strongly with a large gap.

DISCUSSION AND CONCLUSION

The principle of designability is that some structures are intrinsically easier to design than others. However, up to now, designability has been shown only in highly restrictive lattice models. Our calculations indicate that the qualitative features of designability in lattice models are also exhibited off-lattice. Namely, a small minority of off-lattice structures are distinguished by high designability: these structures are lowest-energy states for many more than their share of sequences. Moreover, the sequences associated with these structures have enhanced thermodynamic stability. The work presented here, using an off-lattice model for protein-backbone configurations, makes it more plausible that designability applies to real proteins. Of course, the model used in the current study is highly simplified—it is a low-resolution discrete model of short chain with a very simple potential function. There is still a long way to go to show the designability principle in real proteins.

Nonetheless, the insensitivity to model parameters of the results presented suggests that our highly designable structures are possible candidates for real protein design. It is therefore worthwhile to study some of our best candidates in detail and to understand what architectural properties distinguish the most designable structures from the least designable ones and how the most designable ones compare with known natural structures.

Representative configurations of some of the most designable structures are shown in Figure 1(a–c). A striking characteristic of the highly designable structures is that each has a well-defined core consisting of a small subset of the amino acids of the chain. For example, in Figure 5 we have plotted the inaccessible surface area of each amino acid along the chain for the configuration appearing in Figure 1(b). Observe that 5 of the 23 amino acids are more than 70% buried. Also shown in Figure 5 is the probability that a hydrophobic amino acid occupies a particular site, averaged over all HP-sequences that design the structure, revealing the preference of hydrophobic amino acids for the core.

A quantitative measure of the core in a structure is the variance v_s of the exposure vector $\tilde{\mathbf{A}}$: $v_s = (1/N) \sum_i \tilde{a}_i^2 - (1/N^2) (\sum_i \tilde{a}_i)^2$. In Figure 6, we plot v_s versus the designability N_s. On average the two quantities correlate well; however, the scatter of the data is large in the region of low

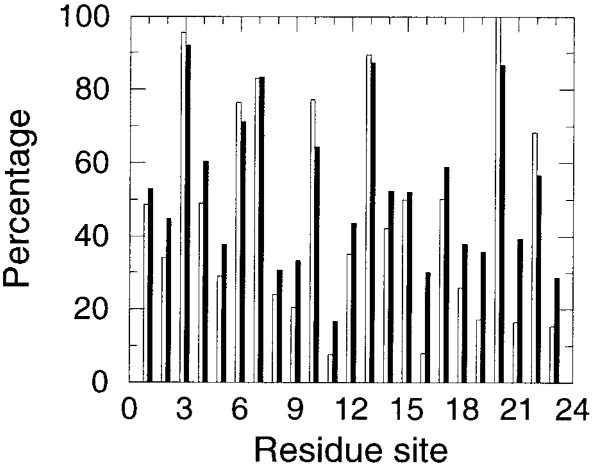

Fig. 5. Solid bars: Inaccessible surface for residues (C_β spheres) of the highly designable configuration shown in Figure 1(b). Hollow bars: Probability, averaged over all HP-sequences that design the configuration, that a particular site along the chain is occupied by a hydrophobic amino acid.

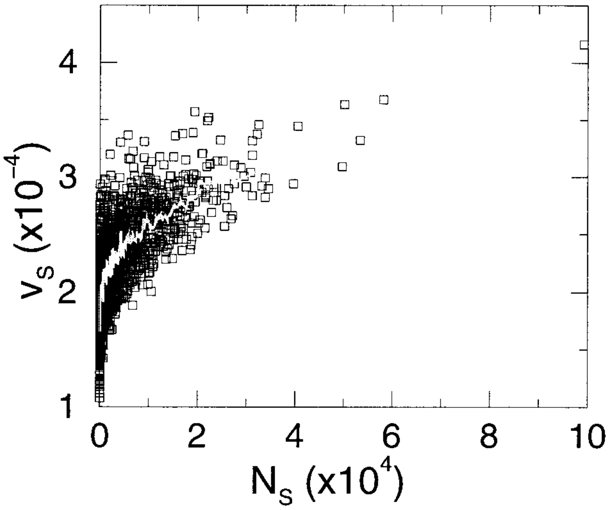

Fig. 6. The average variance v_s of a cluster against the designability N_s of the cluster for the 23-mer. The 5000 most compact configurations participate in the calculation, λ = 0.4 Å, and r_β = 1.9 Å. Gray line: running average with bin size 30.

N_s: structures with well-formed cores are not necessarily highly designable.

A zinc finger-like fold emerges from our calculation as one of the most designable structures. The fold [Fig. 1(*b*)] does not simply replicate 1PSV [Fig. 1(d)], on which we optimized our angle set. The structure of 1PSV is too open to be designable within our model because the small, uniformly sized side-chains cannot fill the large opening between the α-helix and the β-β turn in 1PSV. It is of interest that the model produces a highly designable solution by collapsing the α-helix onto the β-β turn.

Another of our most designable structures is similar to another small natural fold, the helix-turn-helix [see Fig. 1(c)]. Some of our most designable structures [e.g., that shown in Fig. 1(a)] do not resemble any known natural folds. These structures are candidates for the design of truly novel folds.

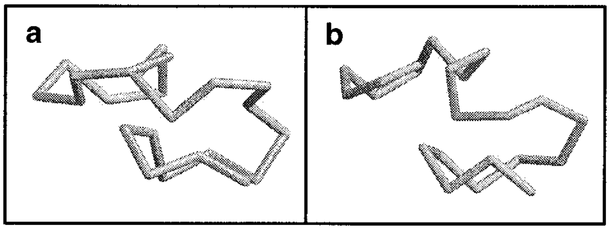

Fig. 7. **a:** Backbone configuration of the 11th most designable 23-mer structure, using untargeted angle set (see text): (ϕ, ψ) = (−55°, 135°), (−126°, 145°), and (−85°, −25°), with a mean crms of 3.6 Å on a representative subset of natural structures segmented into subchains of 21 amino acids. For this calculation, the amino acids are represented by spheres of radius r_α = 1.52 Å centered on the C_α carbons only. **b:** Backbone configuration of the zinc finger 1NC8, truncated to 23 amino acids.[25]

Targeting a fold by fitting the angle set to a single chosen structure is not essential. For example, we can obtain a suitable angle set by choosing two pairs of dihedral angles (ϕ, ψ) within the β-sheet region and one pair from the α-helix region, locally optimizing on 160 representative natural structures from the PDB database.[21] Among the most designable structures emerging for this angle set is the zinc finger-like structure in Figure 7(*a*), shown next to its apparent natural counterpart, 1NC8 [Fig. 7(b)].[25]

Recently, many studies have been conducted on the relation between the folding kinetics and the topology of native states.[26–36] In particular, it has been shown that folding rates and the topology of the transition states are closely related to the topology of the native states. In other words, the native state topology, which in this context is often measured in terms of contact order,[26,35] largely determines how a protein folds. It would be interesting to compare the two roles the native state topology plays: in folding kinetics and in the designability and thermodynamic stability. However, such a comparative study would preferably be done in systems of longer chains than used in the current study. Although it is tempting to think that there is a deep connection between the two roles of topology, one should note that there is a huge variation in folding rates among natural proteins,[33] which are presumably highly designable and thermodynamically stable. It appears that designability is largely governed by the surface-core patterning,[15] whereas folding kinetics depends more on the ease of forming native contacts (the contact order).

In summary, we have computed the designabilities of structures within an off-lattice model of realistic protein-backbone configurations. Highly designable structures emerge with remarkable insensitivity to model parameters. The sequences that design these structures have strongly enhanced mutational stability and a large energy gap between the native fold and the lowest non-native conformation. In this light, it is interesting that recent mutation studies on some small proteins show that they maintain their native folds even when about half of their residues are replaced by alanine.[37,38] Some of our highly designable structures correspond closely to natural folds, such as the zinc finger and helix-turn-helix motifs. Others do not resemble existing structures and are candidates for ab initio design of novel protein folds.

REFERENCES

1. Beasley JR, Hecht MH. Protein design: the choice of de novo sequences. J Biol Chem 1997;272:2031–2034.
2. Baltzer L. Functionalization of designed folded polypeptides. Curr Opin Struct Biol 1998;8:466–470.
3. Cao AN, Lai LH, Tang YQ. The current state and prospect of de novo protein design. Prog Biochem Biophys 1998;25:197–201.
4. Giver L, Arnold FH. Combinatorial protein design by in vitro recombination. Curr Opin Chem Biol 1998;2:335–338.
5. Regan L, Wells J. Engineering and design: recent adventures in molecular design—editorial overview. Curr Opin Struct Biol 1998;8:441–442.
6. Schafmeister CE, Stroud RM. Helical protein design. Curr Opin Biotechnol 1998;9:350–353.
7. Shakhnovich EI. Protein design: a perspective from simple tractable models. Fold Design 1998;3:R45–R58.
8. DeGrado WF, Summa CM, Pavone V, Nastri F, Lombardi A. De novo designa and structural characterization of proteins and metalloproteins. Annu Rev Biochem 1999;68:779–819.
9. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. Science 1998;282:1462–1467.
10. Struthers MD, Cheng RP, Imperiali B. Design of a monomeric 23-residue polypeptide with defined tertiary structure. Science 1996;271:342–345.
11. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82–87.
12. Dahiyat BI, Sarisky CA, Mayo SL. De novo protein design: towards fully automated sequence selection. J Mol Biol 1997;273: 789–796.
13. Chothia C. One thousand families for the molecular biologist. Nature 1992;357:543–544.
14. Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science 1996;273: 666–669.
15. Li H, Tang C, Wingreen NS. Are protein folds atypical? Proc Natl Acad Sci USA 1998;95:4987–4990.
16. Govindarajan S, Goldstein RA. Searching for foldable protein structures using optimized energy functions. Biopolymers 1995;36: 43–51.
17. Govindarajan S, Goldstein RA. Why are some protein structures so common? Proc Natl Acad Sci USA 1996;93:3341–3345.
18. Finkelstein AV, Ptitsyn OB. Why do globular proteins fit the limited set of folding patterns? Prog Biophys Mol Biol 1987;50:171–190.
19. Yue K, Dill KA. Forces of tertiary structural organization in globular proteins. Proc Natl Acad Sci USA 1995;92:146–150.
20. Mélin R, Li H, Wingreen NS, Tang C. Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. J Chem Phys 1999;110:1252–1262.
21. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. J Mol Biol 1995;249:493–507.
22. Flower DR. SERF: A program for accessible surface area calculations. J Mol Graph Model 1997;15:238–244.
23. Creighton TE. Proteins. New York: Freeman; 1993. p160–162, 236–237.
24. Lau KF, Dill KA. Lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules 1989;22:3986–3997.
25. Kodera Y, Sato K, Tsukahara T, Komatsu H, Maeda T, Kohno T. High-resolution solution NMR structure of the minimal active domain of the human immunodeficiency virus type-2 nucleocapsid protein. Biochemistry 1998;37:17704–17713.
26. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
27. Chan HS. Protein folding: matching speed and locality. Nature 1998;392:761–763.
28. Portman JJ, Takada S, Wolynes PG. Variational theory for site resolved protein folding free energy surfaces. Phys Rev Lett 1998;81:5237–5240.
29. Goldenberg DP. Finding the right fold. Nat Struct Biol 1999;6:987–990.
30. Alm E, Baker D. Matching theory and experiment in protein folding. Curr Opin Struct Biol 1999;9:189–196.
31. Fersht AR. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. Proc Natl Acad Sci USA 2000;97:1525–1529.
32. Maritan A, Micheletti C, Banavar JR. Role of secondary Motifs in fast folding polymers: a dynamical variational principle. Phys Rev Lett 2000;84:3009–3012.
33. Baker D. A surprising simplicity to protein folding. Nature 2000;405:39–42.
34. Clementi C, Nymeyerson Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298: 937–953.
35. Plaxco KW, Simons KT, Ruczinski I, Baker D. Sequence, stability, topology and length; the determinants of two-state protein folding kinetics. Biochemistry 2000;39:11177–11183.
36. Guerois R, Serrano L. Protein design based on folding models. Curr Opin Struct Biol 2001;11:101–106.
37. Kuroda Y, Kim PS. Folding of bovine pancreatic trypsin inhibitor (BPTI) variants in which almost half the residues are alanine. J Mol Biol 2000;298:493–501.
38. Brown BM, Sauer RT. Tolerance of Arc repressor to multiple-alanine substitutions. Proc Natl Acad Sci USA 1999;96:1983–1988.