Available online at www.sciencedirect.com
SCIENCE DIRECT
Polymer 47 (2006) 1755–1762
polymer
www.elsevier.com/locate/polymer

What is the origin of those common structures of protein-model chains?

Lei Huang, Xiaojing Ma, Haojun Liang *

Department of Polymer Science and Engineering, University of Science and Technology of China Hefei, Anhui 230026, People's Republic of China

Received 3 March 2005; received in revised form 27 September 2005; accepted 13 November 2005

Abstract

Starting from a kinetically foldable criterion for designing fast-folding structures, we have investigated the foldabilities of all possible sequences coded in two letters through an exhaustive enumeration of model chains of a 16-mer protein that we performed using a simple off-lattice model. From a set of 32,896 sequences, we found only 145 sequences that were foldable. Through a comparison of the geometrical similarities of those foldable sequences, we reduced the corresponding 145 native structures to a structural set of 69 good candidates for target structures in the de novo design of fast-folding sequences. We make the following conclusions: (1) a preferred proportion of compositions exist for sequence design. (2) Foldable sequences having different numbers of hydrophobic residues possess very similar sequences. (3) The stability of some special structures toward mutations may be the origin of common protein structures; our results demonstrate that the presence of hydrophobic residues in certain positions of a sequence can result in firm and mutation-resistant skeletons. It appears that a simple, but robust, chain topology and structural symmetry lead to high designability.

Keywords: Monte Carlo simulation; Protein folding; Conformation

1. Introduction

Studies of the de novo design of proteins, which is an inverse protein folding problem, have attracted considerable attention both experimentally and theoretically for many years, and much progress is being made [1–9]. Generally, a suitable predetermined structure should be selected as a target structure at the very beginning of de novo sequence design [4]. The target structure should be appropriate for design; that is, it should somewhat resemble naturally occurring proteins. It may not be possible to find sequences that can fold into any arbitrary compact structure because it seems that the number of naturally occurring protein fold families may be finite [10–12]. That is to say, Nature may employ a somewhat limited set of functional proteins such that many different sequences share common folds. Jones and Thornton [13] have demonstrated that certain 'super folds' dominate the current structural databases. Goldstein and coworkers [14] have explained this finding by using the energy landscape theory. It remains an open problem as to why some certain folds are so common. Several authors have proposed possible physical mechanisms behind Nature's selection of protein folds. Finkelstein and coworkers [15–17] have argued that certain motifs are easier to stabilize and, thus, are more common either because they have lower energies or because they have unusual energy spectra over random sequences. Yue and Dill [18] have demonstrated that protein-like folds are associated with sequences that have a minimal number of degenerate lowest-energy states. Goldstein et al. [19,20] stated that the robustness of proteins to site mutations results from population dynamics during the evolutionary process. Recently, Chan and coworkers [21,22] demonstrated that evolutionary populations depend greatly on the topologies of proteins.

The exhaustive enumeration method is a well known, simple—but formidable—tool for studying protein folding problems; Tang and coworkers [23–25] and several other groups [26–28] have obtained fruitful results using this approach. Tang and coworkers studied the designabilities, which are measured by the number of sequences that can be designed for the structure, of all the compact conformations for a 27-monomer chain in a $3\times3\times3$ lattice [23]. The authors suggest that a structure is designable with certain sequences when they have their non-degenerate lowest energy at the target structure. From their analysis, they observed that some structures are more designable than others. Such highly designable structures possess 'protein-like' secondary structures and even tertiary symmetries and they are thermodynamically more stable than other structures. Broglia

* Corresponding author.
E-mail address: hjliang@ustc.edu.cn (H. Liang).

0032-3861/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2005.11.036

and Tiana [29] argued that the analysis reported by Tang and coworkers [23] had two basic problems. Firstly, sequences composed of only two kinds of residues are not very suitable for mimicking real proteins; secondly, only fully compact conformations were enumerated. Based on a lattice model using a 'hydrophobic' energy function, de Araujo [30] also pointed out that maximally compact structures might not be the most designable structures in which monomers occupy completely buried or completely exposed positions. For the first problem, recent work [24] by Tang and coworkers used the Miyazawa–Jernigan (MJ) matrix to provide a good qualitative agreement between the two-letter HP model and a 20-letter model. The authors studied an off-lattice model to investigate the designabilities of all possible conformations for a 23-monomer chain in a discrete off-lattice space [25] to avoid the ambiguity deriving from the second problem: they reached similar conclusions.

Following Bryngleson and Wolynes [31,32], a number of other simulations [33–36] have explored the dynamics of protein folding. As demonstrated by Wolynes et al. [37], natural proteins are both thermodynamically and kinetically foldable. The definition of design provided by Tang and coworkers [23–25], however, states that a structure is designable with certain sequences when its non-degenerate lowest energy form is the target structure. We argue that the target structure designed according to this criterion of designability should be thermodynamically designable, but not necessary kinetically designable, because kinetically foldable sequences constitute only a subset of thermodynamically foldable sequences [37]. Based on a lattice model, Du and coworkers [38] demonstrated that a collapsed chain exhibits ergodicity breaking, in which the disjoint regions of phase space do not arise uniformly, but instead as small chambers whose number increases exponentially with respect to the polymer density. A chain would be frustrated in finding the native state when a chain collapses near the glass temperature. Herein, we present an alternative criterion of foldability/designability that considers the requirement for kinetic designability. Our model should be more general because there are no rigid restrictions on conformations (e.g. discrete or dihedral angles), which, therefore, creates a continuous conformational space. A sequence can be considered foldable when the chain can fold into a target structure with a significant statistical probability (e.g. 0.8), starting from an arbitrary initialized conformation during an appointed finite time. The corresponding target structure is deemed designable. We argue that fast-folding sequences can, at least, be obtained when using this design procedure and employing the new criterion. We performed an exhaustive enumeration to obtain all 32,896 possible sequences of a 16-monomer chain coded by two letters and have checked their foldabilities with our fast-fold criterion. Details of the foldability calculations for certain sequences are provided below. A thorough analysis of the relationships among those foldable sequences and designable structures leads us to a conclusion that is essentially coincident with that of Tang and coworkers [25].

2. Model and algorithms

The conformation of a chain made up of n beads is defined by n coordinates (r_1, r_2,…,r_n) of beads in a three-dimensional space. Each bead may represent one C_α atom of an amino acid. The total internal energy is a sum of the Lennard–Jones (LJ) potential between non-bonded beads and the harmonic spring potential between bonded beads, respectively. The system has a Hamiltonian of

$$H = 4\sum_{i=1}^{n-2}\sum_{j=i+2}^{n}\eta_{s_i s_j}\left(\left(\frac{\sigma}{r_{ij}}\right)^{12}-\left(\frac{\sigma}{r_{ij}}\right)^{6}\right)+\frac{k}{2} \times\sum_{i=1}^{n-1}(r_{i,i+1}-r_0)^2 \quad (1)$$

The first term is the contribution of the LJ potential between non-bonded beads; the latter is contributed by the energies of the fluctuation of $n-1$ bonds in the n-mer. According to the protein data bank (PDB), the mean distance between two neighboring C_α atoms in a protein chain is $r_0=3.8$ Å. In our model, each randomly selected residue on the chain is allowed to move around its position [39] with bond fluctuation restricted to between $r_{i,i+1}=3.7$ and 3.9 Å. In our model, sequences are coded by two kinds of residues, labeled by H and P, which are similar to the hydrophobic and polar residues in the HP model [23]. The coefficient $\eta_{s_i s_j}$ in the LJ potential term reflects the interaction between residues s_i and s_j and can be expressed in a 2×2 interaction matrix, having $\eta_{HH}=40$, $\eta_{HP}=\eta_{PH}=20$, and $\eta_{PP}=5$, following the choices suggested by Clementi et al. [40] for the interactions among four types of amino acids. We chose the parameters $\eta_{HH}=40$ and $\eta_{PP}=5$ to ensure that stronger interactions exist between hydrophobic beads and relatively weaker interactions between hydrophilic beads. The term σ is determined from the requirement that the average number of C_α–C_α contacts for each amino acid roughly equals the respective number obtained using the all-atom definition of contacts. We set $\sigma=6.5$ Å to ensure that a strong interaction occurs between two monomers when their inter-monomer distance is below 9 Å. We assigned a value of 20.0 to the constant k.

For tractability in computation, we exhaustively enumerated in sequence space a chain of 16 residues coded by two letters, H and P. From a consideration of symmetry, we eliminated those sequences that were redundant. All of the sequences we obtained were checked for their foldabilities. In this present study, we did not introduce any prior determined target structure. We applied a simulated annealing algorithm to determine the lowest-energy conformation, starting from 50 different initialized conformations and various random numbers. We made comparisons between all of these 50 conformations. The root-mean-square coordinate deviation,

$$D=\left(\frac{\sum_{i=1}^{n}(r_i-r_{i'})^2}{n}\right)^{0.5}$$

was calculated to characterize the degree of similarity between the geometries of two conformations, which we considered to be the same if D was below 0.25 Å, which is a much more rigid requirement than the experimental resolution. Assuming the number of conformations that share a certain structure, labeled *, is $n_{f,}$, we define the sequence for structure * to be foldable if n_f is not less than 40 (i.e. its probability is higher than 80%); otherwise, we do not consider it to be foldable. We believe that although the number of individual samples, the threshold number for the criterion of a foldable chain, the time span of the simulated annealing run, and even the parameters in the simulated annealing, are all set arbitrarily, they do not alter our final results significantly; essentially, our conclusion should be valid if some parameters are changed within a reasonable range because we used the same parameters to study all of the sequences. In this study, we began our simulations of simulated annealing at a temperature $T=100$; the system reached equilibrium after 5000 steps/bead during a continuous cooling process in which the temperature was lowered each time by a factor of 0.90. The simulations were run until the temperature had decreased to below 10^{-7}, which requires 10^6 MC steps during the entire simulated annealing span. We classify the foldable sequences as fast-folding if the checked chains converge at same conformation during this time, but they are considered not foldable, or at least not fast-folding, if they fail to converge to certain conformations.

3. Results and discussions

In total, there are 32,896 possible sequences into which a 16-mer chain can be coded using two letters. We investigated the foldabilities of these sequences; they are listed in Table 1. As expected, the balance of hydrophobic and hydrophilic residues on a protein chain plays a significant role in determining the foldabilities of the protein chains. There should be an optimal component, which guarantee a protein's junctions, in naturally occurring proteins. Table 1 displays that the optimal proportion of hydrophobic residues—i.e. where the largest foldable sequence number is found—is $n_H/n_p=6/10$. Having more or fewer hydrophobic residues in the chain leads to a decrease in foldability. Several groups [41,42] have observed that incorporating too many or too few hydrophobic residues results in a failure to design protein-like sequences.

As shown by Shakhnovich et al. [43], the residues that are involved in nucleus formation during protein folding are widely conserved during Nature's selection process. Different proteins should have many of the same segments in common along their sequences. Trinquier et al. [44] reached a similar conclusion based on their use of a lattice chain model. On the other hand, as various new sequences emerge (as a result of mutation), only the very few whose folding provides some certain physiological function are reserved. We can consider that mutations create protein sequence candidates based on a limited number of existing protein sequences and that natural selection acts as a quality auditor to guarantee that they have a special function. We found a trace of mutation design during the analysis of foldable sequences having various compositions of hydrophobic (H) residues. Sequences of n monomers are represented by an ordered set of n letters, $\{a_i\}=\{a_1,\ldots,a_n\}$, where a_i stands for the monomer type, H or P, at position i along the sequence. Sequence fitness between two sequences can be measured by the number of different residues along a sequence. In a simple fashion, when two sequences having m and n hydrophobic residues, respectively, only have $|m-n|$ different residues along their sequences, we classify these two sequences as fitting one another; otherwise, they do not fit. As an illustration, consider the following four sequences:

$$S_{2a}=\{P\ P\ P\ \mathbf{H}\ P\ P\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\ P\}$$

$$S_{3a}=\{P\ P\ \mathbf{H}\ \mathbf{H}\ P\ P\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\ P\}$$

$$S_{3b}=\{P\ P\ P\ \mathbf{H}\ P\ \mathbf{H}\ P\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\}$$

$$S_{3c}=\{P\ P\ P\ \mathbf{H}\ P\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ \mathbf{H}\ P\ P\ P\}$$

We say that S_{2a} and S_{3a} fit, but S_{2a} and S_{3b} do not. In this example, S_{3b} and S_{3c} are symmetrical and are considered the same. Herein, we would use 'fit' and 'fitness' to mean those sequences that are a minimal hamming distance away. Those sequences in a fit sequence pair can be changed into their counterparts by $|m-n|$ point mutations. For two compositions having H/P proportions of n_{Ha}/n_{Pa} and n_{Hb}/n_{Pb}, respectively, their foldable sequence sets are $\{\{a_i\}\}n_{Ha}$ and $\{\{a_i\}\}n_{Hb}$, respectively. For $n_{Ha}>n_{Hb}$, we can calculate the fitness fraction between those two sequence sets, where the sequence $\{a_i\}^*$ in $\{\{a_i\}\}n_{Ha}$ is considered to fit the sequence set $\{\{a_i\}\}n_{Hb}$ if there is at least one sequence in $\{\{a_i\}\}n_{Hb}$ that fits $\{a_i\}^*$; the fitness fraction between those two sets is defined as the number of sequences in $\{\{a_i\}\}n_{Ha}$ that fit $\{\{a_i\}\}n_{Hb}$ divided by the element number in $\{\{a_i\}\}n_{Ha}$. An example is helpful in understanding this definition. The foldable sequence sets having n_H equal to 2 and 3 are indicated as follows:

$$\{S_2\}=\begin{Bmatrix} S_{2a}=\{P\ P\ \mathbf{H}\ P\ P\ P\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\}, \\ S_{2b}=\{P\ P\ P\ \mathbf{H}\ P\ P\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\ P\} \end{Bmatrix}$$

$$\{S_3\}=\begin{Bmatrix} S_{3a}=\{P\ \mathbf{H}\ \mathbf{H}\ P\ P\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\ P\}, \\ S_{3b}=\{P\ P\ \mathbf{H}\ \mathbf{H}\ P\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\ P\}, \\ S_{3c}=\{P\ P\ P\ \mathbf{H}\ \mathbf{H}\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\ P\}, \\ S_{3d}=\{P\ P\ P\ \mathbf{H}\ \mathbf{H}\ P\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\}, \\ S_{3e}=\{P\ P\ P\ P\ \mathbf{H}\ \mathbf{H}\ P\ P\ P\ P\ P\ \mathbf{H}\ P\ P\ P\} \end{Bmatrix}$$

To calculate the fitness between $\{S_3\}$ and $\{S_2\}$, we must check how many sequences in $\{S_3\}$ fit $\{S_2\}$. In our example, S_{3b} and S_{3c} both fit S_{2b}. Namely, they can be obtained by a single mutation starting from S_{2b}. Sequences S_{3d} and S_{3e} also fit S_{2b} when symmetry is considered. Sequence S_{3a} does not fit either

Table 1
The number of sequences obtained through exhaustive enumeration for various H/P proportions and the number of foldable sequences by our designability criterion

n_H/n_P	0/16	1/15	2/14	3/13	4/12	5/11	6/10	7/9	8/8
n_s	1	8	64	280	924	2184	4032	5720	6470
n_d	0	0	2	5	15	31	40	28	18
n_H/n_P	16/0	15/1	14/2	13/3	12/4	11/5	10/6	9/7	
n_s	1	8	64	280	924	2184	4032	5720	
n_d	0	0	0	0	0	1	3	6	

Rows n_H/n_P, n_s, and n_d represent the H/P proportion, the number of possible sequences, and the number of foldable sequences, respectively.

Table 2
The fitness between foldable sequences sets having various numbers of H residues, presented in a matrix style

n_H	n_H									
	2	3	4	5	6	7	8	9	10	11
2		0.80	0.59	0.87	1.00	1.00	1.00	1.00	1.00	1.00
3	0.50		0.47	0.71	0.88	0.93	1.00	1.00	1.00	1.00
4	1.00	0.80		0.65	0.83	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	0.67		0.76	0.93	1.00	1.00	1.00	1.00
6	1.00	1.00	0.73	0.68		0.71	1.00	1.00	1.00	1.00
7	1.00	1.00	0.73	0.71	0.57		0.55	0.83	1.00	1.00
8	1.00	1.00	0.87	0.81	0.63	0.43		0.83	1.00	1.00
9	1.00	1.00	0.80	0.77	0.60	0.39	0.28		0.50	1.00
10	1.00	1.00	0.87	0.81	0.68	0.57	0.78	0.17		1.00
11	1.00	1.00	0.73	0.65	0.45	0.61	0.33	0.17	0.33	

The fitness in the upper diagonal represents the possibility of designing a fast fold starting from foldable sequences having fewer H residues and proceeding to more H units through mutations; that in the lower diagonal corresponds to the reverse process, i.e. starting from foldable sequences having more H residues and proceeding to fewer H units. The term n_H represents the number of H residues in the 16-mer.

S_{2a} or S_{2b}. The fitness between $\{S_3\}$ and $\{S_2\}$ is 0.8 because four out of the five sequences in $\{S_3\}$ fit $\{S_2\}$. Using this procedure, we can conclude that the fitness between two sequences sets, $\{S_{n_{Ha}}\}$ and $\{S_{n_{Hb}}\}$, represents the probability of finding all foldable sequences having more H residues starting from those foldable sequences having fewer H residues through an iterative single point mutation of replacing one P unit with an H residue. We stress that the requirement of fit is a very rigorous one because we only count in the sequences that fit others perfectly and we do not explicitly consider mutations through crossover mutations in our calculation. On the other hand, for $n_{Ha} < n_{Hb}$, the fitness indicates the possibility of the inverse design in which H residues are replaced by P units starting from a foldable sequence with a relatively higher proportion of H units.

We have calculated the fitness between all possible n_{Ha}/n_{Hb} pairs, displayed in Table 2, for both of the cases $n_{Ha} < n_{Hb}$ and $n_{Ha} > n_{Hb}$. As pointed out above, the up diagonal matrix in Table 2 indicates the probability of a foldable sequence design, starting from a foldable sequences set having a relatively low proportion of H units; the down diagonal matrix represents the opposite scenario, i.e. starting from foldable sequences having a high H proportion. It is clear that most sequences sets have high fitness, which means that they can be designed through a single point mutation in a step-by-step manner or through multiple point mutations. The mechanism thus imposed alleviates the need for an exhaustive search in sequence space, as pointed out by Yomo et al. [45]. Some low fitness is suggested to derive from crossover mutation between two residues, which we ignore in our discussion. The effect is very significant for small values of $|n_{Ha} - n_{Hb}|$ near the diagonal. By comparing the two parts in Table 2, a sequence design that starts from fewer H residues has, in the most part, a higher probability than one starting with more H residues (the main exceptions occur in several instances in the upper-left-hand corner). We can comprehend this phenomenon by considering the hierarchy sketch in Fig. 1, which does not display the sequences that do not fit sequences in the neighboring hierarchy. From top to bottom, some sequences lead to several fit sequences upon increasing the number of H residues, but others do not, which leads us to the conclusion that a design proceeding from fewer to more H residues has a higher probability than a design procedure occurring in the reverse manner. For example, as indicated in Fig. 1, seven sequences in level d can be obtained by mutation from sequences in level c; if only the sequences in our scheme are considered, only two out of five sequences in level c can be generated from seven sequences in level d. The situation may be complex when considering sequences that do not fit sequences in

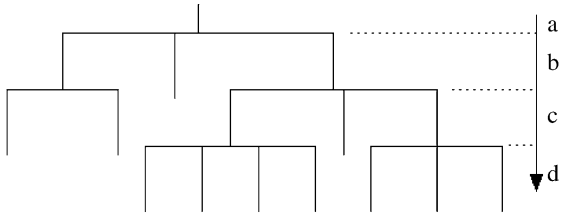

Fig. 1. A sketch of the hierarchy in evolution relative to the design sequence sets. Proceeding from the top to the bottom (from (a) to (d)), the four levels have correspondingly increased proportions of H residues.

Table 3
All of the foldable sequences, except for those corresponding to isolated structures

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	P	P	P	*H*	*H*	*H*	P	P	P	P	P	*H*	P	P	P	P
	P	P	**H**	*H*	*H*	*H*	P	P	P	P	P	*H*	P	P	P	P
	P	P	P	*H*	*H*	*H*	P	P	P	P	**H**	*H*	**H**	P	P	P
	P	P	**H**	*H*	*H*	*H*	P	P	P	P	P	*H*	**H**	P	P	P
	P	P	P	*H*	*H*	P	P	**H**	P	P	**H**	*H*	**H**	P	P	P
	P	P	**H**	*H*	*H*	*H*	P	P	P	P	**H**	*H*	P	P	**H**	P
	P	**H**	P	*H*	*H*	*H*	P	P	P	P	**H**	*H*	**H**	P	P	P
	P	P	**H**	*H*	*H*	*H*	P	P	P	P	**H**	*H*	**H**	P	P	P
	P	P	**H**	*H*	*H*	*H*	P	P	P	P	**H**	*H*	P	**H**	P	P
	P	P	P	*H*	*H*	*H*	P	**H**	P	P	**H**	*H*	**H**	P	P	P
	P	P	**H**	P	*H*	*H*	P	**H**	P	P	**H**	*H*	**H**	P	P	P
	P	P	**H**	*H*	*H*	*H*	**H**	P	P	P	**H**	*H*	**H**	P	P	P
	P	P	**H**	*H*	*H*	*H*	P	**H**	P	P	**H**	*H*	**H**	P	P	P
	P	P	**H**	*H*	*H*	*H*	P	P	P	P	**H**	*H*	**H**	**H**	P	P
	P	P	**H**	*H*	*H*	*H*	P	P	P	P	**H**	*H*	**H**	P	**H**	P
	P	**H**	P	*H*	*H*	*H*	P	**H**	P	P	**H**	*H*	**H**	P	P	P
	H	P	**H**	*H*	*H*	*H*	P	P	P	P	**H**	*H*	**H**	**H**	P	P
	H	P	**H**	*H*	*H*	*H*	**H**	P	P	P	**H**	*H*	**H**	P	P	P
2	P	P	P	*H*	*H*	P	P	P	P	P	P	P	*H*	P	P	P
	P	P	**H**	*H*	*H*	**H**	P	P	P	P	P	P	*H*	P	P	P
	P	P	**H**	*H*	*H*	P	P	P	P	P	P	P	*H*	**H**	P	P
	P	**H**	P	*H*	*H*	P	P	P	P	P	P	P	*H*	**H**	P	P
	P	P	**H**	*H*	*H*	P	**H**	P	P	P	P	P	*H*	**H**	P	P
	P	P	**H**	*H*	*H*	P	**H**	P	P	P	P	P	*H*	P	**H**	P
	H	P	**H**	*H*	*H*	**H**	P	P	P	P	P	P	*H*	**H**	P	P
	P	P	**H**	*H*	*H*	P	**H**	P	P	**H**	P	P	*H*	**H**	P	P
	H	P	**H**	*H*	*H*	**H**	P	P	P	P	P	P	*H*	P	**H**	P
3	P	P	P	*H*	*H*	P	P	P	P	P	P	*H*	P	P	P	P
	P	**H**	**H**	*H*	*H*	**H**	P	P	P	P	P	*H*	P	P	P	P
	P	P	**H**	*H*	*H*	**H**	P	P	**H**	P	P	*H*	P	P	P	P
	P	P	**H**	*H*	*H*	P	P	P	**H**	P	P	*H*	**H**	P	P	P
	P	**H**	**H**	*H*	*H*	**H**	P	P	P	P	P	*H*	P	**H**	P	P
	P	P	**H**	*H*	*H*	**H**	P	P	**H**	P	P	*H*	**H**	P	P	P
	P	**H**	**H**	*H*	*H*	**H**	P	P	**H**	P	P	*H*	**H**	P	P	P
	P	**H**	**H**	*H*	*H*	**H**	P	P	**H**	P	**H**	*H*	**H**	**H**	P	P
	P	**H**	**H**	*H*	*H*	**H**	P	P	**H**	P	**H**	*H*	**H**	P	**H**	P
4	P	P	*H*	*H*	P	P	P	P	P	P	P	*H*	P	P	P	P
	P	**H**	*H*	*H*	**H**	P	P	P	P	P	P	*H*	P	P	P	P
	P	**H**	*H*	*H*	P	P	P	P	P	P	P	*H*	**H**	P	P	P
	P	**H**	*H*	*H*	P	**H**	P	P	P	P	P	*H*	**H**	P	P	P
	H	**H**	*H*	*H*	**H**	P	P	P	P	P	P	*H*	P	P	**H**	P
	P	**H**	*H*	*H*	P	**H**	P	P	**H**	P	P	*H*	**H**	P	P	P
	H	**H**	*H*	*H*	**H**	P	P	P	P	P	P	*H*	**H**	P	P	P
5	P	P	P	*H*	*H*	*H*	P	*H*	P	P	P	*H*	*H*	P	P	P
	P	P	P	*H*	*H*	P	P	*H*	P	**H**	P	*H*	*H*	P	P	P
	P	P	**H**	*H*	*H*	*H*	P	*H*	P	P	P	*H*	*H*	P	P	P
	P	P	**H**	*H*	*H*	*H*	P	*H*	P	P	P	*H*	*H*	**H**	P	P
	P	**H**	P	*H*	*H*	*H*	P	*H*	P	P	P	*H*	*H*	**H**	P	P
	P	P	**H**	*H*	*H*	*H*	P	*H*	P	**H**	P	*H*	*H*	**H**	P	P
	H	P	**H**	*H*	*H*	*H*	P	*H*	P	P	**H**	*H*	*H*	**H**	**H**	P
6	P	P	*H*	*H*	*H*	P	*H*	P	P	P	*H*	*H*	P	P	P	P
	P	P	*H*	*H*	P	P	*H*	P	P	P	*H*	*H*	**H**	P	P	P
	P	**H**	*H*	*H*	*H*	P	*H*	P	P	P	*H*	*H*	P	P	P	P
	P	**H**	*H*	*H*	*H*	P	*H*	P	P	P	*H*	*H*	P	**H**	P	P
	P	**H**	*H*	*H*	*H*	P	*H*	P	P	**H**	*H*	*H*	**H**	P	P	P
	P	**H**	*H*	*H*	*H*	P	*H*	P	P	**H**	*H*	*H*	**H**	P	**H**	P
7	P	P	P	P	*H*	*H*	P	P	P	P	P	P	*H*	P	P	P
	P	P	P	**H**	*H*	*H*	P	P	P	P	P	P	*H*	**H**	P	P
	P	P	**H**	P	*H*	*H*	P	P	P	P	P	P	*H*	**H**	P	P
	P	P	P	**H**	*H*	*H*	**H**	P	P	**H**	P	P	*H*	**H**	P	P
8	P	P	*H*	*H*	*H*	P	P	P	P	P	*H*	*H*	P	P	P	P
	P	P	*H*	*H*	*H*	P	**H**	P	P	**H**	*H*	*H*	P	P	P	P

(*continued on next page*)

Table 3 (*continued*)

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	P	**H**	*H*	*H*	*H*	P	P	P	P	**H**	*H*	*H*	**H**	P	P	P
	P	**H**	*H*	*H*	*H*	P	P	P	P	**H**	*H*	*H*	P	P	**H**	P
9	P	P	P	*H*	*H*	*H*	P	P	P	P	P	*H*	**H**	P	P	P
	P	**H**	P	*H*	*H*	*H*	P	P	P	P	P	*H*	**H**	P	P	P
	P	P	**H**	*H*	*H*	*H*	**H**	P	P	P	P	*H*	P	P	P	P
	P	**H**	P	*H*	*H*	*H*	**H**	P	P	P	**H**	*H*	**H**	**H**	P	P
10	P	**H**	P	*H*	P	P	P	P	P	P	P	*H*	*H*	P	P	P
	P	P	**H**	*H*	P	P	P	P	P	P	P	*H*	*H*	P	P	P
	P	P	**H**	*H*	P	P	**H**	P	P	P	P	*H*	*H*	P	P	P
11	P	P	*H*	*H*	P	P	P	P	P	P	P	P	*H*	*H*	P	P
	H	P	*H*	*H*	P	P	P	P	P	P	P	P	*H*	*H*	P	P
	P	**H**	*H*	*H*	P	P	P	P	P	P	P	P	*H*	*H*	**H**	P
12	P	P	*H*	*H*	*H*	P	P	P	P	P	P	*H*	H	P	P	P
	P	P	*H*	*H*	*H*	P	P	P	P	P	P	*H*	P	**H**	P	P
	P	P	*H*	*H*	*H*	P	**H**	P	P	P	P	*H*	H	P	P	P
13	P	P	P	*H*	*H*	*H*	P	P	P	P	**H**	*H*	P	*H*	P	P
	P	P	P	*H*	*H*	*H*	P	P	P	**H**	P	*H*	P	*H*	P	P
	P	**H**	P	*H*	*H*	*H*	P	P	P	P	**H**	*H*	P	*H*	P	P
14	P	*H*	*H*	P	P	P	P	P	P	P	P	*H*	P	P	P	P
	P	*H*	*H*	**H**	P	P	**H**	P	P	P	P	*H*	P	P	P	P
15	P	P	*H*	P	*H*	*H*	P	P	P	P	P	*H*	P	P	P	P
	P	P	*H*	P	*H*	*H*	P	P	P	P	P	*H*	**H**	P	P	P
16	P	*H*	*H*	P	*H*	P	P	P	P	P	P	*H*	P	P	P	P
	P	*H*	*H*	**H**	*H*	P	P	**H**	P	P	P	*H*	P	P	P	P
17	P	P	P	**H**	*H*	*H*	*H*	P	P	P	P	P	*H*	*H*	P	P
	P	P	**H**	P	*H*	*H*	*H*	P	P	P	P	**H**	*H*	*H*	**H**	P
18	P	*H*	*H*	*H*	*H*	P	P	P	P	P	*H*	P	P	*H*	P	P
	P	*H*	*H*	*H*	*H*	P	P	P	**H**	P	*H*	P	P	*H*	P	P
19	*H*	P	*H*	*H*	*H*	P	P	P	P	P	*H*	P	P	*H*	P	P
	H	P	*H*	*H*	*H*	P	P	P	P	**H**	*H*	P	P	*H*	P	P
20	P	P	P	*H*	*H*	P	P	*H*	P	P	P	P	*H*	*H*	*H*	P
	P	**H**	P	*H*	*H*	**H**	P	*H*	P	P	P	P	*H*	*H*	*H*	P
21	P	*H*	P	*H*	*H*	P	**H**	P	*H*	P	P	*H*	*H*	P	P	P
	P	*H*	P	*H*	*H*	**H**	P	P	*H*	P	P	*H*	*H*	P	P	P
22	P	P	*H*	*H*	*H*	P	P	P	P	P	*H*	P	P	P	P	P
	P	P	*H*	*H*	*H*	P	P	**H**	P	**H**	*H*	**H**	**H**	P	P	P

From our results, 145 structures can be grouped into 22 groups and 47 isolated structures. The H residues are presented in bold; those at 'hot' sites involved in forming the skeleton are rendered in italics.

the neighboring hierarchy, but the scenario is mainly the same in our case. The deviation at the upper-left-hand corner of Table 2 presumably derives from the existence of many isolated sequences for $n_H = 4$ and 5. The origin of the difference between the up and down diagonal matrix in Table 2 is the emergence of some preferred sequences that lead to several sequences in the following hierarchy having more H residues. A more fundamental reason for this effect is explained below.

Only 145 sequences remain to fit our criterion of fast folding and there is one corresponding conformation for each foldable sequence. That is to say, a de novo design would fail to produce a fast-folding structure if its pre-determined target structure is not within the structure set comprising those 145 structures. There is such an example presented in Ref. [46]; a de novo design for a 16-mer chain coded in two letters failed because a random compact conformation was selected as the target structure. The analysis of those structures having high foldabilities should be very instructive. To characterize the structural and sequence behaviors of these distinct structures, first we separated the chains into several groups according to their geometric similarities. The chains in each group converge to a special structure; that is to say, there exists a central structure that is similar to one in all the other structures in the group. We classify two structures as being identical if the D value between the two structures is below 0.25 Å. Using this criterion, we classified the 145 sequences into 22 groups and 47 isolated structures. It seems that the conclusion made by Tang and coworkers [25] regarding the designabilities of different structures—that different structures have various designabilities and many protein-like sequences prefer sharing some certain common structures—should hold in our model. Our largest group contains 18 structures and there are at least six structures in groups 1–6; i.e. many sequences share the same structure in each group. Table 3 displays all of the chain sequences in groups; all H residues have been rendered in bold. The most intriguing observation is that almost all the sequences in each group possess some common positions, which are shown in italics that are occupied by H units. As an illustration, H residues occupy three sites—positions 4, 5, and 13—in all of the sequences in group 3. These three H residues play a crucial role in determining the skeleton structure of the 16-mer chain; this skeleton is so stable that it fixes the chain into the special

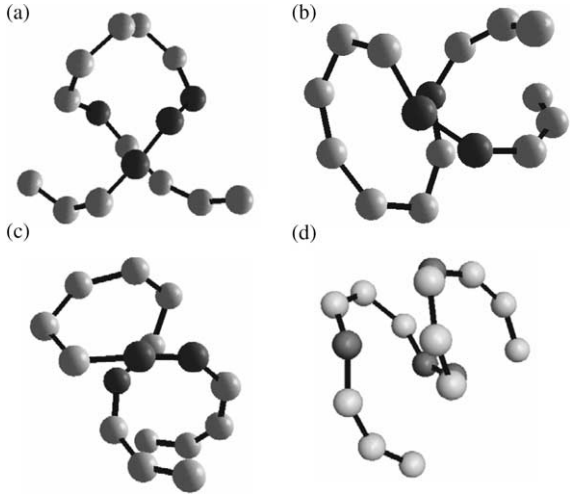

Fig. 2. Some typical native structures. The structures in a to d possess 18, 9, 4, and 1 sequences, respectively. The light- and dark-grey spheres represent polar and hydrophobic residues, respectively.

structure presented in Fig. 2(b) when some certain H residues are introduced. It seems that we are approaching the answer to the question regarding the origin of those common protein structures. For those common folds, the skeleton is very stable to mutations and various protein-like sequences share the same structure that emerges from the sequence's design procedure, charged by evolution and promoted by mutation, starting from a sequence having that special structure as its native state. This concept also provides some new insight into the high probability of sequence design starting from a low proportion of H units and replacing P with H residues.

Tang et al. [23] pointed out that structures having symmetry possess high designabilities; other groups [47,48] have also argued that symmetry can cause proteins to fold into their native states quickly. In this study, we also observe symmetry in these structures. Fig. 2 displays several typical native structures. It should be noted that the measure of symmetry is different from the regular symmetry elements in point group. We found that the two sub-chain 8-mers formed by splitting a chain at its central position possess geometric similarities, which implies a small value of D in their structural comparison. The D values for the two sub-chains of the four structures in Fig. 2(a)–(d) are 0.28, 1.27, 2.21, and 0.03 Å, respectively. Relative to the structure in Fig. 2(a), the two structures in Fig. 2(b) and (c) share similar topologies, but the structure in Fig. 2(d) does not, which is supported by geometric comparisons. In a geometric comparison using the structure in Fig. 2(a) as a reference, the D values for the structures in Fig. 2(b)–(d) are 1.56, 2.01, and 6.21 Å, respectively. As we indicate in the caption to Fig. 2, upon proceeding from (a) to (d), we observe a decrease in the number of foldable sequences for the corresponding structures. This discussion implies that there are complex relationships between the designabilities of structure and symmetry. We suppose that the structures in Fig. 2(a)–(c) exhibit twisted hairpin-like topologies, which leads to their high designability because of its robustness toward mutation. With a similar topology, the designability may benefit from a higher structural symmetry, as is implied from the comparison of the three structures in Fig. 2. Despite its perfect symmetry, the structure in Fig. 2(d) is an isolated example because its delicate conformation will change significantly when mutations occur. The topology—widely recognized as playing an important role in protein folding [49–55]—may largely determine the designability of a structure, which also can be improved by high symmetry. Herein, our native structures seem to differ significantly from those of common protein structures [11] and we believe that these differences derive from the simple force-field that we employed for our model.

4. Conclusions

Starting from a kinetic criterion of foldability/designability, we investigated the foldabilities of all possible sequences coded in two letters through an exhaustive enumeration of 16-mers. We consider only 145 out of 32,896 sequences to be good candidates for fast folding. The native structures of these 145 sequences converge to 69 conformations and constitute a structure set of good candidates for target structures for de novo fast-fold design. Based on our discussion, we reach the following conclusions: (1) a preferred proportion of compositions exist for sequence design. (2) Foldable sequences possessing different numbers of hydrophobic residues have very similar sequences; this situation indicates that we can design, through single point mutations of a known fast-folding sequence, new fast-folding structures having different numbers of H residues. (3) It seems that the common structures are those that are stable against mutations; their skeletons are so robust that structural change is small when more H residues are introduced. It seems that a simple, but robust, chain topology and structural symmetry will lead to high designability.

Acknowledgements

We are grateful for the financial support provided by the Outstanding Youth Fund (No. 20525416), the Programs of the National Natural Science Foundation of China (Nos. 20490220, 20374050, and 90403022), and the National Basic Research Program of China (No. 2005CB623800).

References

[1] Summa CM, Rosenblatt MM, Hong JK, Lear JD, DeGrado WF. J Mol Biol 2002;321:923.
[2] Bolon DN, Voigt CA, Mayo SL. Curr Opin Chem Biol 2002;6:125.
[3] Kraemer-Pecore CM, Wollacott AM, Desjarlais JR. Curr Opin Chem Biol 2001;5:690.
[4] Saven JG. Chem Rev 2001;101:3113.
[5] Pokala N, Handel TM. J Struct Biol 2001;134:269.
[6] Gibney BR, Dutton PL. Adv Inorg Chem 2001;51:409.
[7] Hill RB, Raleigh DP, Lombardi A, Degrado NF. Acc Chem Res 2000;33:745.
[8] DeGrado WF, Summa CM, Pavonem V, Nastri F, Lombardi A. Annu Rev Biochem 1999;68:779.
[9] Shakhnovich EI. Fold Des 1998;3:R45.
[10] Chothia C. Nature 1992;357:543.

[11] Orengo CA, Jones DT, Thornton JM. Nature 1994;372:631.
[12] Wolf YI, Grishin NV, Koonin EV. J Mol Biol 2000;299:897.
[13] Jones D, Thornton J. J Comput Aided Mol Des 1993;7:439.
[14] Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Proc Natl Acad Sci USA 1992;89:4918.
[15] Finkelstein AV, Ptitsyn OB. Prog Biophys Mol Biol 1987;50:171.
[16] Finkelstein AV, Gutin AM, Badretdinov AY. FEBS Lett 1993;325:23.
[17] Finkelstein AV, Badretdinov AY, Gutin AM. Proteins 1995;23:142.
[18] Yue K, Dill KA. Proc Natl Acad Sci USA 1995;92:146.
[19] Taverna DM, Goldstein RA. J Mol Biol 2002;315:479.
[20] Taverna DM, Goldstein RA. Proteins 2002;46:105.
[21] Bornberg-Bauer E, Chan HS. Proc Natl Acad Sci USA 1999;96:10689.
[22] Cui Y, Wong WH, Bornberg-Bauer E, Chan HS. Proc Natl Acad Sci USA 2002;99:809.
[23] Li H, Helling R, Tang C, Wingreen NS. Science 1996;273:666.
[24] Li H, Tang C, Wingreen NS. Proteins 2002;49:403.
[25] Miller J, Zeng C, Wingreen NS, Tang C. Proteins 2002;47:506.
[26] Shakhnovich E, Gutin A. J Chem Phys 1990;93:5967.
[27] Shahrezaei V, Hamedani N, Ejtehadi MR. Phys Rev E 1999;60:4629.
[28] Irback A, Troein C. J Biol Phys 2002;28:1.
[29] Broglia RA, Tiana G. Protein folding, evolution and design. Ohmsha: IOS Press; 2001.
[30] de Araujo AFP. Proc Natl Acad Sci USA 1999;96:12482.
[31] Bryngelson JD, Wolynes PG. Proc Natl Acad Sci USA 1987;84:7524.
[32] Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Proc Natl Acad Sci USA 1992;89:4918.
[33] Socci ND, Onuchic JN. J Chem Phys 1994;101:1519.
[34] Abkevich VI, Gutin AM, Shakhnovich EI. J Mol Biol 1995;252:460.
[35] Socci ND, Onuchic JN. J Chem Phys 1995;103:4732.
[36] Melin R, Li H, Wingreen NS, Tang C. J Chem Phys 1999;110:1252.
[37] Onuchic JN, Luthey-Schulten Z, Wolynes PG. Annu Rev Phys Chem 1997;48:545.
[38] Du R, Grosberg AY, Tanaka T, Rubinstein M. Phys Rev Lett 2000;84:2417.
[39] Gerroff I, Milchev A, Binder K, Paul W. J Chem Phys 1993;98:6526.
[40] Clementi C, Maritan A, Banavar JR. Phys Rev Lett 1998;81:3287.
[41] Yue K, Dill KA. Proc Natl Acad Sci USA 1992;89:4163.
[42] Wang JY, Wang J, Wang W. Chin Phys Lett 2001;18:449.
[43] Shakhnovich E, Abkevich V, Ptitsyn O. Nature 1996;379:96.
[44] Trinquier G, Sanejouand Y. Phys Rev E 1999;59:942.
[45] Yomo T, Saito S, Sasai M. Nat Struct Biol 1999;6:743.
[46] Liang HJ. J Chem Phys 2000;113:4827.
[47] Wolynes PG. Proc Natl Acad Sci USA 1996;93:14249.
[48] Nelson ED, Teneyck LF, Onuchic JN. Phys Rev Lett 1997;79:3534.
[49] Alm E, Baker D. Curr Opin Struct Biol 1999;9:189.
[50] Riddle DS, Santiago JV, BrayHall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. Nature Struct Biol 1997;4:805.
[51] Kim DE, Gu HD, Baker D. Proc Natl Acad Sci USA 1998;95:4982.
[52] Baker D. Nature 2000;405:39.
[53] Shea JE, Onuchic JN, Brooks CL. Proc Natl Acad Sci USA 1999;96:12512.
[54] Micheletti C, Banavar JR, Maritan A, Seno F. Phys Rev Lett 1999;82:3372.
[55] Abkevich VI, Gutin AM, Shakhnovich EI. Biochemistry 1994;33:10026.