Origin of the designability of protein structures

Rie Tatsumi and George Chikenji
Department of Physics, Graduate School of Science, Osaka University, Machikaneyama-cho 1-1, Toyonaka, Osaka 560-0043, Japan

(Received 7 December 1998; revised manuscript received 3 May 1999)

We examined what determines the designability of two-letter codes (H and P) lattice proteins from three points of view. First, whether the native structure is searched within all possible structures or within maximally compact structures. Second, whether the structure of the used lattice is bipartite or not. Third, the effect of the length of the chain, namely, the number of monomers on the chain. We found that the bipartiteness of the lattice structure is not a main factor that determines the designability. Our results suggest that highly designable structures will be found when the length of the chain is sufficiently long to make the hydrophobic core consisting of a large enough number of monomers. [S1063-651X(99)18810-X]

PACS number(s): 87.15.By, 87.10.+e, 02.70.Lq

INTRODUCTION

Natural proteins fold into unique compact structures in spite of the huge number of possible conformations [1]. For most single domain proteins, each of these native structures corresponds to the global minimum of the free energy [2].

It has been proposed phenomenologically that the number of possible structures of natural proteins is only about 1000 [3], which suggests that many sequences can fold into one preferred structure. There have been theoretical studies for the existence of such preferred structures [4–8].

In many theoretical studies for the protein folding, a simplified model called the hydrophobic polar (HP) model [4,5,7–12] is adopted. The HP model is one of two-letter codes lattice models where a protein is represented by a self-avoiding chain of beads placed on a lattice with two types of beads, hydrophobic (H) and polar (P). In the HP model, the energy of a structure is given by the nearest-neighbor topological contact interactions as

$$H=-\sum_{i<j} E_{\sigma_i\sigma_j}\Delta(r_i-r_j), \tag{1}$$

where i and j are monomer indexes, $\{\sigma_i\}$ are monomer types (σ=H or P); $\Delta(r_i-r_j)=1$ if r_i and r_j are topological nearest neighbors not along the sequence, and $\Delta(r_i-r_j)=0$ otherwise.

Based on the HP model, a concept of *designability* has recently been introduced [4]; the number of sequences that have a given structure as their nondegenerate ground state (native state) is called the *designability* of this structure. When many sequences have a common native structure, one can say that the structure is *highly designable*. Adding to the importance in the protein design problem, the designability also has evolutional significance because highly designable structures are found to be relatively stable against mutations.

In the original study of Li *et al.* [4], HP models on the square and cubic lattices are employed, with the energy parameters in Eq. (1) being $E_{HH}=-2.3$, $E_{HP}=-1.0$, and $E_{PP}=0.0$. For each sequence, they calculated the energy over all maximally compact structures and picked up the native structure. The results indicated that highly designable structures actually exist on both lattices.

Irbäck and Sandelin studied the HP models on the square and triangular lattices [5]. They adopted different energy parameters from Li *et al.* [4], namely, $E_{HH}=-1$ and $E_{HP}=E_{PP}=0$. In the calculation of the designability they considered all the possible structures not restricting themselves to the maximally compact ones. For the square lattice, they confirmed the existence of the highly designable structures as in Ref. [4]. For the triangular lattice, however, no such structures were found. In addition to the nearest-neighbor topological contact interactions, they considered local interactions represented by the bend angle and calculated the designability. Indeed the local interactions reduced degeneracy (i.e., the number of sequences which have nondegenerate ground state increased) and made the designability higher. But they found that the designability on the square lattice was still much higher than that on the triangular lattice. They concluded that the difference in the designability for these two lattices are related to the even-odd problem, that is, whether the lattice structure is bipartite or not.

Quite recently, Li *et al.* proposed a model based on the HP model on the square lattice [6]. In the model, the hydrophobic interaction is treated in such a way that the energy decreases if the hydrophobic residue is buried in the core. They justify this treatment with two reasons: (1) the hydrophobic force that is dominant in folding [13,14] originates from aversion of hydrophobic residues from water. (2) The Miyazawa-Jernigan matrix [15] contains a dominant hydrophobic interaction of the linear form $E_{\alpha\beta}=h_\alpha+h_\beta$ [16]. They took

$$H=-\sum_{i=1}^{N} s_i h_i, \tag{2}$$

where $\{h_i\}$ represent a sequence ($h_i=1$ if the ith amino acid is H-type, and $h_i=0$ if it is P-type), and $\{s_i\}$ represent a structure ($s_i=0$ if the ith amino acid is on the surface and $s_i=1$ if it is in the core). They calculated the designability over all maximally compact structures, whose result is consistent with their former study [4] (see Table I).

In our view, there are many points to be explored further for the designability problem. First, since the structures of natural proteins are compact but not necessarily ‘‘maximally compact’’ in general, how can we justify the discussion

TABLE I. The difference among three studies is shown. Each variable in the Hamiltonian is defined in the text.

	Li *et al.* [4]	Irbäck and Sandelin [5]	Li *et al.* [6]
Lattice	square and cubic	square and triangular	square
Interaction	nearest-neighbor		depend on the position of an H
Hamiltonian	$H=-\sum_{i<j} E_{\sigma_i\sigma_j}\Delta(r_i-r_j)$		$H=-\sum_{i=1}^{N} s_i h_i$
Energy parameter (E_{HH},E_{HP},E_{PP})	$(-2.3,-1.0,0.0)$	$(-1,0,0)$	
Conformational space	maximally compact	all	maximally compact
Highly designable structures	found on both lattices	found on square lattice but not found on triangular one	found

where only the maximally compact structures are taken into account? Second, is it adequate to consider only nearest-neighbor interactions? Properties of a system with only nearest-neighbor interactions are directly affected by the lattice structure, in particular, whether the lattice is bipartite or not. Is it good, only from these facts, to conclude immediately that the absence of the highly designable structures on the triangular lattice should be ascribed to the even-odd problem associated with the triangular lattice [5]? One should discuss the problem on the triangular lattice by using a model like the one in Ref. [6] where the interactions do not depend on the contact between monomers, hence, *do not directly reflect the nonbipartiteness*.

Our aim of this paper is to examine the above points and clarify what determines the designability of protein structures. For that purpose, we introduce a model called the ‘‘solvation model’’ and calculate the designability over *all possible* structures on the square and the triangular lattices. Comparing the results with those of the HP model, we investigate which properties of designability are less sensitive to a choice of models and energy parameters. In the solvation model, a sequence consists of two-type amino acids (H and P) and based on Ref. [6], the energy increases if the hydrophobic residue is exposed to the solvent. In brief, the solvation model is a two-letter codes lattice model where the hydrophobic force to form a core is dominant and the interactions do not directly reflect the bipartiteness.

MODELS

In the solvation model based on Ref. [6] a protein is represented by a self-avoiding chain of beads with two types H and P, placed on a lattice. A sequence is specified by a choice of monomer types at each position on the chain.

We used two-dimensional lattice models because a computable length by numerical enumeration of the full conformational space is limited (square lattice, 18; triangular and cubic lattices, 13). Even with this chain-length limitation, we can make a ‘‘hydrophobic core’’ in two dimensions, in contrast with the three-dimensional case.

A structure is specified by a set of coordinates for all the monomers and is mapped into the number of contacts with the solvent. In our model, the total energy is given in terms of the monomer-solvent interactions, and depends only on the number of contacts with the solvent:

$$H=\sum_{i=1}^{N} E_{s_i} h_i , \qquad (3)$$

where $\{h_i\}$ represent a sequence, $h_i=1$ if the ith monomer is the H-type, and $h_i=0$ if it is P-type. The variable s_i denotes the number of contacts with the solvent, for example, $s_i=\{0,1,2,3\}$ on the square lattice and $s_i=\{0,1,2,3,4,5\}$ on the triangular lattice. In other words, $s_i=0$ means that the ith monomer is buried away from the solvent. We take $E_0=0$, $E_1=\sqrt{2}$, $E_2=\sqrt{7}$, $E_3=\sqrt{13}$, $E_4=\sqrt{19}$, and $E_5=\sqrt{23}$. That is, the possible minimum energy is zero. And these parameters are selected so that the larger the number of contacts with the solvent is, the more the degree of energy increase is; the hydrophobic residue is energetically unfavorable to be at the corner [17,18]. Although the choice of these values is somewhat arbitrary, we have considered the following points: (1) these values should not increase too rapidly with the increase in the number of contacts with the solvent and (2) the way of choosing these values must not bring about nonessential accidental degeneracies (due to simple rational ratios between the parameters) [19].

Using the model on the square and triangular lattices, we calculate the designability for all the 2^N sequences, where N is the number of monomers, by the exact computer-enumeration method over the full conformational space. To get correct data we exclude overcounting coming from redundant structures that are mutually related by rotation, reflection, and reverse labeling.

On the basis of data obtained by the solvation model and the HP model, we examine what determines the designability from three points of view: (1) the effect of the search-space restriction, namely, the search within maximally compact structures (in this paper, we just used maximally compact structures as the simplest example of the search-space restriction, and we may consider other ones, e.g., structures with the biggest core), (2) the effect of the lattice structure, namely, whether the lattice is bipartite or not (or, equivalently, the even-odd problem), and (3) the effect of the number of monomers (or, the length of the chain).

RESULTS AND DISCUSSION

Let us now give results of calculations.

(i) *The effect of the search within maximally compact structures*. In Fig. 1, we show the designability calculated on

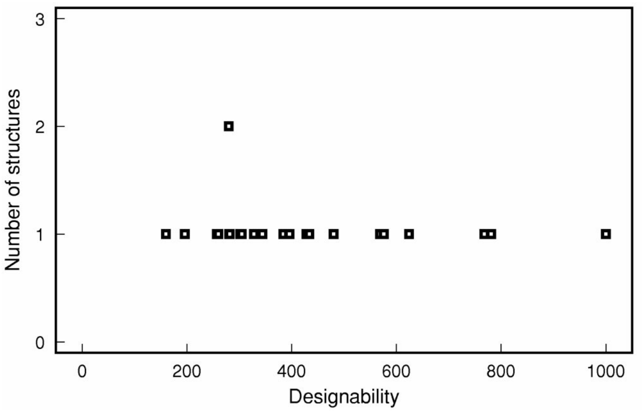

FIG. 1. The designability calculated over maximally compact structures on the square lattice for $N=16$. The vertical axis indicates the number of structures of the same designability.

the square lattice for $N=16$, using maximally compact structures. For comparison, in Fig. 2 we show the designability of the same system without the search-space restriction (i.e., search over all possible structures). In both cases, there are some highly designable structures. However, these structures are not common to both cases. In Fig. 2, the number of sequences that have native structures is 8277, but the number of sequences that have maximally compact structures as native is only 1087 out of 8277. That is, most sequences that have native structures have nonmaximally compact structures as native. The importance of nonmaximally compact structures has also been pointed out for the HP model [5,20–23]. These facts imply that it is not good to calculate the designability over only maximally compact structures. Such calculation picking up a ''native'' structure out of maximally compact structures is not correct if the true native structure is nonmaximally compact. Further, when the lowest-energy nonmaximally compact structure and the lowest-energy maximally compact structure are degenerate, there is no native structure (native structure must be nondegenerate), but the restricted-search-space calculation gives a false result that there is a native (and maximally compact) structure. We should say that the designability calculated over only maximally compact structures may be erroneous.

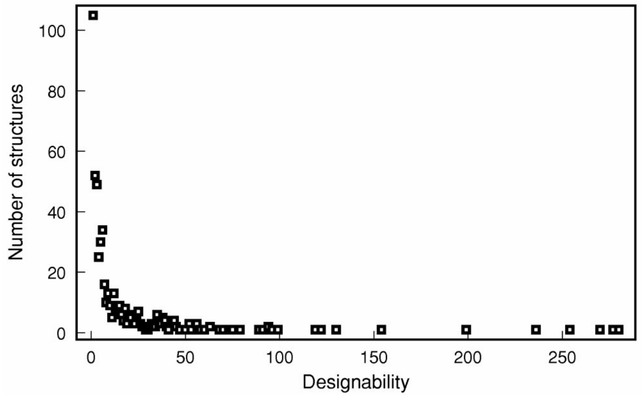

FIG. 2. The designability calculated over all possible structures on the square lattice for $N=16$.

TABLE II. S_n and D_h on the triangular lattice for $N=13$. The parentheses correspond to energy parameters (E_{HH}, E_{HP}, and E_{PP}). The data in the HP model with the energy parameters being $E_{HH}=-1$ and $E_{HP}=E_{PP}=0$ was obtained by Irbäck and Sandelin [5]. S_n and D_h are defined in the text.

	S_n	D_h
HP model $(-1,0,0)$	0	0
HP model $(-2.3,-1.0,0.0)$	129	3
Solvation model	7	1

(ii) *The effect of the lattice structure: bipartite or nonbipartite*. In two previous studies using the HP model [4,5], interactions of the system directly reflected whether the lattice is bipartite or not. Moreover, the designability on the triangular lattice was calculated with the energy parameters in Eq. (1) being $E_{HH}=-1$ and $E_{HP}=E_{PP}=0$, which would cause accidental degeneracies. In their results, highly designable structures were not found for the triangular lattice. Also, it seemed that native structures are likely to contain the hydrophobic core where a group of hydrophobic monomers contact with each other; such contact can be made only if the distance between the monomers along the sequence is odd. Therefore, the bipartiteness has been thought to be a main source of the designability [4,5,24]. If so, highly designable structures do not actually exist, i.e., the concept of *designability* itself could be meaningless. On the other hand, if such preferred structures should exist on the basis of the proposal by Chothia [3], the use of the lattice model would be inadequate. Then, we used the solvation model, which does not directly reflect the bipartiteness, and calculated the designability on the square and triangular lattices. Besides, we also calculated the designability on the triangular lattice using the HP model, with the energy parameters being $E_{HH}=-2.3$, $E_{HP}=-1.0$, and $E_{PP}=0.0$.

In Table II, we show the total number of sequences that have nondegenerate ground state (S_n) and the highest designabilities (D_h) on the triangular lattice for $N=13$, obtained by using different interactions. This result shows that, even if we take different values of energy parameters, or even if we use the solvation model, the triangular lattice is still unfavorable for the designability although S_n varies largely. On the other hand, for the square lattice, highly designable structures are found in the solvation model as well

TABLE III. The designability calculated over all possible structures on the square lattice for $N=10$. The right column indicates the number of structures of the same designability.

Designability	Number of structures
1	1
2	2
3	4
4	3
5	4
6	1
10	2
12	1

TABLE IV. The designability calculated over all possible structures on the square lattice for $N=11$.

Designability	Number of structures
1	5
2	11
3	4
4	1
5	3
8	1
10	1
13	1
18	1
29	1
36	1
43	1

as in the HP model (Fig. 2). These results imply that the absence of the highly designable structures for the triangular lattice should not be ascribed to the even-odd problem (or, the nonbipartiteness), but to other reasons. The properties that highly designable structures are found on the square lattice and no such structures are found on the triangular lattice might be general in two-letter codes lattice models where the hydrophobic force is dominant.

(iii) *The effect of the number of monomers*. Then, why are the highly designable structures absent for the triangular lattice? The smallness of the number of monomers (in other words, the length of a chain is too short) may be a possible reason. An important object in the protein structure is the hydrophobic core, which consists of buried monomers having no contact with the solvent. Recall that the limit of a computable length by exact enumeration of the full conformational space on the triangular lattice is 13. The biggest core, which we can make by using this limited length, is the one that consists of only three monomers; the length is too short for the hydrophobic force to form a core. This monomer-number effect is also found on the square lattice. Consider the following conditions: at least ten sequences have a given structure as their native state, and at the same time, there are at least five such structures. Only if these conditions are satisfied, let us say that ''there are highly designable structures.'' Then, at $N=10$ or less, there are no highly designable structures even for the square lattice (Tables III and IV). This result implies that when we discuss whether there are highly designable structures or not, we need a long enough chain to make a core of enough size. This further implies that in a three-dimensional case, we will need a chain of longer length than that in a two-dimensional case to make a core.

Let us see Tables III, IV, and V. In Table V we show the designability calculated on the triangular lattice for $N=13$. On the square lattice for $N=10$, the biggest core consists of two monomers. Both on the triangular lattice for $N=13$ and on the square lattice for $N=11$, the biggest core consists of three monomers. We see that the triangular lattice is unfavorable for designability compared with the square lattice, even when the biggest possible core size is the same or a little larger. A possible reason would be the number of all possible structures, particularly the number of structures with the biggest core. As the length of a chain becomes long, the number of all possible structures increases almost exponentially as μ^N ($2<\mu<3$ for the square lattice and $4<\mu<5$ for the triangular lattice) [25]. On the triangular lattice for $N=13$, the number of all possible structures is 6 279 601 and the number of structures with the biggest core is 4110 out of them. On the other hand, on the square lattice for $N=10$ and 11, the number of all possible structures is 2034, 5513 and the number of structures with the biggest core is 23 and 5, respectively. Thus the number of all possible structures and the number of structures with the biggest core on the triangular lattice are much larger than those on the square lattice [26]. In consequence, the degeneracy tends to grow, which is unfavorable for designability. In this view, designable structures on the triangular lattice would be more difficult to appear than on the square lattice.

TABLE V. The designability calculated over all possible structures on the triangular lattice for $N=13$.

Designability	Number of structures
1	7

SUMMARY

We have calculated the designability using the solvation model and the HP model on the square and the triangular lattices to deduce what determines the designability of protein structures. The solvation model introduced in this paper satisfies two conditions: (1) the hydrophobic force is dominant and (2) the model does not directly reflect the bipartiteness. We have examined what determines the designability from three points of view: effect of restricted search within maximally compact structures, the bipartite/nonbipartite effect, and the length of the chain.

In conclusion, we have found that it is inadequate to calculate the designability within maximally compact structures. Our results imply that the reason why no highly designable structures on the triangular lattice have been found is not the nonbipartiteness. We suppose that the main factor, which affects the designability, is the chain length because for sufficiently large hydrophobic core to form, long enough chains are required. A triangular lattice is more unfavorable for the designability than a square lattice irrespective of models or energy parameters, probably because the number of all possible structures is large. However, if we can deal with a longer chain than in the present study, it is possible that we may find highly designable structures even on the triangular lattice. The calculations of the designability for longer chains on the triangular lattice are highly desirable. These conclusions would apply to a wide variety of two-letter codes lattice models, where the hydrophobic force is dominant, regardless of energy parameters and further details of the model.

Though a concept of designability is currently defined for a two-letter codes lattice model, our final goal is to examine whether natural proteins have highly designable structures. Therefore, it is an interesting problem to extend the study of the designability for a 20-letter codes model [27] (e.g., the Miyazawa-Jernigan model [15] and Kolinski-Godzik-

Skolnick model [28]) and an off-lattice model. Substituting 20-letter codes for two-letter codes certainly reduces degeneracy, and most of all sequences come to have a structure as a nondegenerate ground state (i.e., native structure).

ACKNOWLEDGMENTS

We would like to thank Y. Akutsu and M. Kikuchi for useful discussions and careful readings of the manuscript.

[1] *Protein Folding*, edited by T. E. Creighton (Freeman, New York, 1992).

[2] C. Anfinsen, Science **181**, 223 (1973).

[3] C. Chothia, Nature (London) **357**, 543 (1992).

[4] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).

[5] A. Irbäck and E. Sandelin, J. Chem. Phys. **108**, 2245 (1998).

[6] H. Li, C. Tang, and N. S. Wingreen, Proc. Natl. Acad. Sci. USA **95**, 4987 (1998).

[7] M. R. Ejtehadi, N. Hamedani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad, J. Phys. A **31**, 6141 (1998).

[8] R. Melin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).

[9] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[10] F. Seno, M. Vendruscolo, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **77**, 1901 (1996).

[11] J. M. Deutsch and T. Kurosky, Phys. Rev. Lett. **76**, 323 (1996).

[12] C. J. Camacho and D. Thirumalai, Phys. Rev. Lett. **71**, 2505 (1993).

[13] K. A. Dill, Biochemistry **29**, 7133 (1990).

[14] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).

[15] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).

[16] H. Li, C. Tang, and N. S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).

[17] K. Yue and K. A. Dill, Proc. Natl. Acad. Sci. USA **92**, 146 (1995).

[18] K. Yue and K. A. Dill, Phys. Rev. E **48**, 2267 (1993).

[19] M. R. Ejtehadi, N. Hamedani, and V. Shahrezaei, Phys. Rev. Lett. **82**, 4723 (1999).

[20] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, Proc. Natl. Acad. Sci. USA **92**, 325 (1995).

[21] H. Frauenkron, U. Bastolla, E. Gerstner, P. Grassberger, and W. Nadler, Phys. Rev. Lett. **80**, 3149 (1998).

[22] M. Vendruscolo, A. Maritan, and J. R. Banavar, Phys. Rev. Lett. **78**, 3967 (1997).

[23] D. K. Klimov and D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).

[24] E. Shakhnovich, V. Abkevich, and O. Ptitsyn, Nature (London) **379**, 96 (1996).

[25] N. Madras and G. Slade, *The Self-Avoiding Walk* (Birkhäuser, Boston, 1993).

[26] M. Vendruscolo, B. Subramanian, I. Kanter, E. Domany, and J. Lebowitz, Phys. Rev. E **59**, 977 (1999). In this paper, it is shown that the larger the number of all possible structures, the larger the number of structures with the same number of contacts between monomers.

[27] M. Vendruscolo, R. Najmanovich, and E. Domany, Phys. Rev. Lett. **82**, 656 (1999). Although pairwise contact potentials with 20-letter codes are shown to be unsuitable for the prediction of the native conformation of natural proteins in this paper, we consider that a 20-letter codes model is still meaningful for theoretical understanding of general properties of protein folding.

[28] A. Kolinski, A. Godzik, and J. Skolnick, J. Chem. Phys. **98**, 7420 (1993).