# Secondary-structure-favored hydrophobic-polar lattice model of protein folding

Hu Chen,[1] Xin Zhou,[2] and Zhong-Can Ou-Yang[1,2]

[1]*Center for Advanced Study, Tsinghua University, Beijing 100084, People's Republic of China*

[2]*Institute of Theoretical Physics, Academia Sinica, P.O. Box 2735, Beijing 100080, People's Republic of China*

(Received 12 December 2000; revised manuscript received 8 March 2001; published 20 September 2001)

Protein folding is studied using a two-dimensional lattice model with the Hamiltonian including both hydrophobic interactions and main chain hydrogen bond interactions of amino acids. Since compact conformations have different designabilities and only highly designable conformations can act as native structural candidates [H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996)], it is shown that hydrophobic interaction alone is insufficient to explain the appearance of a high proportion of regular secondary structures, especially $\beta$ sheets whose content decreases with increasing designability, but interactions of main chain hydrogen bonds can account for this. Thus the emergence of only a small number of structure types (folds) among all possible structures can be understood to some extent.

DOI: 10.1103/PhysRevE.64.041905 PACS number(s): 87.14.Ee, 87.15.By

## I. INTRODUCTION

The functions of a protein greatly depend on its three-dimensional (3D) structure [1]. It is believed that the 3D structure of a protein is determined by the sequence of amino acids in the protein. The native conformation is in a state of minimum free energy [2]. Most native structures of globular proteins are compact, and more strikingly have a high degree of order (the presence of secondary structures, such as $\alpha$ helices and $\beta$ sheets). In the protein database SCOP (structural classification of proteins) [3], proteins with known structures are classified according to their structural and sequential information [4]. The *fold*, defined by the arrangement of the different secondary structures of the protein and the topology of their connections, is an important level of classification. The number of protein sequences is practically infinite, but, surprisingly, it is predicted that the number of protein folds is only about 1000 [5]. It is still not quite clear why proteins have a large amount of secondary structures and only a small number of possible folds.

In the early work of Chan and Dill [6], it was found that under a two-dimensional (2D) lattice model the average proportion of secondary structures is rather high (about 50% to 70%) in a compact conformation. However, later on, an off-lattice model [7] showed that compactness does not create enough secondary structures in real proteins, although the content of secondary structures increases with increasing compactness. It is known that in the secondary structure hydrogen bonds always exist between the main chains. They must be important for the stability of the secondary structure. Among the inter-residue interactions in proteins, such as van der Waals interactions, electrostatic interactions, hydrophobic interactions, and hydrogen bonds, only the formation of hydrogen bonds depends on the special orientation of the interacting groups. In a secondary structure, regular arrangement of the residues is advantageous for the formation of main chain hydrogen bonds [8]. In the recent work of Chan *et al.* [9,10], it was found that hydrophobic interactions alone cannot account for the calorimetric two-state picture of proteins, and helical cooperativity and hydrogen bonding can cause thermodynamic behaviors closer to experiment. Recently, Hansmann and Okamoto [11] studied the formation of an $\alpha$ helix using an all-atom model (except water). In the folding process, like the van der Waals energy term, the hydrogen bond energy term obviously decreases with decreasing temperature. This implies that the hydrogen bond is also a driving force for the formation of secondary structure.

In order to characterize the different compact conformations, Li *et al.* introduced an important concept, the *designability* [12]. The designability of a conformation is defined as the number of amino acid sequences with that conformation as the native structure. Among the compact conformations, not all can be used as the native structure of some amino acid sequences, i.e., there are some compact conformations with zero designability. There are also a small number of the compact conformations with exceedingly high designability. Only these highly designable structures can be used as native structure candidates for proteins, and they correspond to the few fold structures in the SCOP database. Under the HP (hydrophobic-polar) model, to some extent, the appearance of some highly designable structures is understandable [13]. However, some characteristics of highly designable conformations, such as secondary structures, have not yet been fully studied. We are interested in the following questions. (i) Do highly designable structures contain more secondary structures than less designable ones? (ii) Are the hydrogen bonds important for the appearance of secondary structures?

## II. MODEL AND METHOD

In this paper, the relationship between the designability of a compact conformation and its content of regular secondary structures is studied using the 2D square lattice model. In order to study the effect of hydrogen bonds in regular secondary structures, a sequence-independent energy term originating from the hydrogen bonds in regular secondary structures is added to the Hamiltonian of the standard HP model. The present model, taking hydrogen bonds in both $\alpha$ helices and $\beta$ sheets into consideration, may be considered as an extension of the ‘‘helical-HP model’’ of Thomas and Dill [14].

In the 2D square lattice, a polypeptide is simplified as a

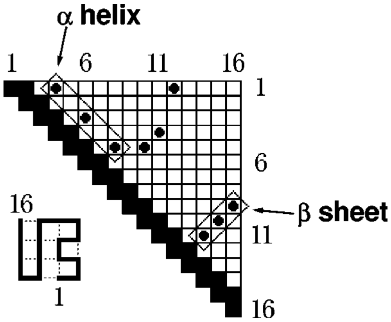

FIG. 1. Compact conformation with length $N=16$ and its contact map (filled circle represents 1 and open lattice represents 0). The specific patterns of $\alpha$ helix and $\beta$ sheet are indicated by arrows.

sequence of beads in self-avoiding-walk conformation. Secondary structures, such as $\alpha$ helices and $\beta$ sheets, can be identified by their special patterns in the *contact map* [6]. A chain of length $N$ corresponds to an $N\times N$ matrix in the contact map. If the residues $i$ and $j$ are nearest neighbors in space and nonadjacent along the chain, we say that there is a *topological contact* between them. If there is topological contact between the $i$th and the $j$th beads, the corresponding matrix element $C(i,j)$ of the contact map is 1, and otherwise it is 0. Figure 1 shows an example of a compact conformation together with its contact map. In the present work, only $\alpha$ helices and $\beta$ sheets (parallel and antiparallel) are considered as regular secondary structures with minimal units containing six beads. The case of two neighboring beads in one sequence being close neighbors of two antiparallel $\beta$ sheets is treated as a part of the $\beta$ sheet. For a compact conformation, the proportions of $\alpha$ helices, $\beta$ sheets (parallel or antiparallel), and all secondary structures are defined as the number of beads participating in the corresponding structures divided by the number of beads $N$.

The Hamiltonian of a given sequence $\{\sigma_i\}$ now takes the form

$$H=\sum_{i<j} E_{\sigma_i\sigma_j}C(i,j)+E_{hb}N_{hb}, \qquad (1)$$

where the first term comes from the hydrophobic interactions, and the second term from the hydrogen bond interactions within the secondary structures, namely, the *secondary-structure-favored* term. $E_{\sigma_i\sigma_j}$ represents the hydrophobic interaction between residue $\sigma_i$ and $\sigma_j$, such as $E_{HH}$, $E_{HP}$, or $E_{PP}$ (the energies of *H-H*, *H-P*, or *P-P* interactions). $C(i,j)$ is the element of the contact map as defined above. $E_{hb}$ is the average energy of a hydrogen bond and $N_{hb}$ is the number of hydrogen bonds in a conformation. In regular secondary structures (both $\alpha$ helices and $\beta$ sheets), it is considered that there exists a hydrogen bond between any two beads in topological contact.

The hydrophobic interaction parameters $E_{HH}$, $E_{HP}$, and $E_{PP}$ should satisfy the following physical constraints: (i) the native structures of most globular proteins are compact; (ii) most $H$ residues are buried in the core and most $P$ residues are exposed on the surface; (iii) different types of residues tend to separate from each other. Therefore, $E_{HH}$, $E_{HP}$, and $E_{PP}$ must satisfy the relations $E_{HH}<E_{HP}<E_{PP}<0$ and $E_{PP}+E_{HH}<2E_{HP}$ [12]. On account of constraint (i), in what follows, we focus only on the fully compact conformations to find the native conformation of a chain. Because compact conformations of a given chain have the same number of contacts, the value of $E_{HH}$, $E_{HP}$, or $E_{PP}$ can be shifted without any change in the result. We set $E_{HH}=-2.3$, $E_{HP}=-1$, and $E_{PP}=0$ as in the work of Li *et al.* [12] and study the effect of various values of $E_{hb}$.

In this paper, we focus our attention on the 2D $6\times6$ compact conformations. They are big enough to have a core-surface ratio of 16:20 similar to that of real proteins. The restriction of the designability calculation to maximally compact conformations may induce errors for the standard HP model [15]. Because of the limitation of current computer capacity, it is impossible to include noncompact conformations in the calculation. However, if we adopt the ''perturbed homopolymer model'' [16] in our calculation, no error will be introduced. The perturbed homopolymer model assumes that each monomer is strongly attracted by all other monomers, i.e., the contact energies are $E_{\sigma_i\sigma_j}-C$ and $C\rightarrow\infty$. Thus all native conformations are maximally compact.

## III. RESULTS

All of the 28 728 $6\times6$ compact conformations unrelated under rotation, reflection, and reverse-labeling operations are obtained by enumeration. To calculate the designabilities of all compact structures, only a random sampling of the $2^{36}$ sequences was performed. To assure the reliability of the sampling method, two sets of the same number of independent sequences were used to calculate the designabilities of all compact structures, and the correlation of the results was tested. Finally, the sequence sample includes $28\,728\times800=22\,982\,400$ sequences (enough to suppress statistical fluctuations). In the following results, the designabilities ($N_s$) of $6\times6$ structures refer only to the sequence samples, not to all the $2^{36}$ sequences. Only sequences with energy differences between the native conformation and the first excited state greater than 0.1 are considered in designing the native conformation.

When $E_{hb}=0$, the present model reduces to the model of Li *et al.* [12]. The average proportion of all secondary structures as a function of the designability is shown in Fig. 2. It is nearly a constant, except that it jumps by about 10% for the conformations with the highest designabilities. To see what happens with increasing designability of the conformation, the average proportions of $\alpha$ helices and $\beta$ sheets (parallel and antiparallel) are also shown in Fig. 2. Surprisingly, when designability is greater than 300, with increasing designability the proportion of $\alpha$ helices increases, but the proportion of antiparallel $\beta$ sheets decreases. With designability less than 300, proportions of both $\alpha$ helices and antiparallel $\beta$ sheets increase with increasing designability, but the rates of increase are very small. For parallel $\beta$ sheets, the proportion is smaller than those of $\alpha$ helices and antiparallel $\beta$ sheets. The proportion of parallel $\beta$ sheets decreases with

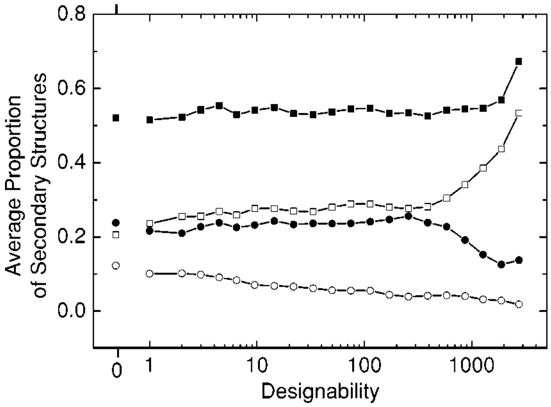

FIG. 2. Average proportion of $\alpha$ helices (open squares), parallel $\beta$ sheets (open circles), antiparallel $\beta$ sheets (filled circles), and all secondary structures (filled squares) as functions of designability for standard HP model.

increasing designability, and it is nearly a linear function of the logarithm of designability.

If structures with the highest designability are the native structures of proteins and compactness is the basic reason for the appearance of secondary structures, then the amount of $\alpha$ helices in proteins should be nearly four times the amount of $\beta$ sheets, a conclusion inconsistent with real fact. The appearance of large amount of secondary structures in the present model is considered to come from the lattice constraint [7]. Thus the origin of the large amount of secondary structures in proteins must be reconsidered.

The average designabilities of structures with the same proportion of regular secondary structures is plotted as a function of the proportion of regular secondary structures in Fig. 3(a) together with their standard deviations. It shows that the average designability has some weak dependence on the proportion of secondary structures, as suggested by Li *et al.* [12]. The average designability of structures without any secondary structure is 314, and that of structures with 100% of secondary structures is only 559. All information on average designabilities is essentially suppressed by the large standard deviations. Structures with either a large or a small amount of regular secondary structures can have high designability. But many structures with a large amount of regular secondary structures have only small designability. Some of them may even be good models of typical folds in the real protein world (such as the first conformation in the all-$\alpha$ class and the first conformation in the all-$\beta$ class shown in Fig. 5 below, they are good 2D models of $\alpha$ bundle and $\beta$ sandwich folds, respectively).

When the second term of the Hamiltonian is taken into account, one finds Fig. 3(b) and Fig. 3(c). It is surprising that a small value of $E_{hb}$, such as $-0.01$, will greatly change the designability of many structures. The average designability of the structures with a high proportion of secondary structures obviously overtops the others. As $E_{hb}$ $(<0)$ decreases, the highly designable structures converge to the structures with a high proportion of secondary structures, and the designabilities of structures with small amounts of secondary structures decrease. When $E_{hb}$ decreases to $-1.0$, all structures with less than 75% secondary structures have zero designability, i.e., no sequence will select them as the native structure. The results are understandable: if the conformation contains a high proportion of regular secondary structures, its energy will be lower even if the chain is a homopolymer; thus the conformation will be a deeper trap in conformation space and there will be more sequences with it as native conformation.

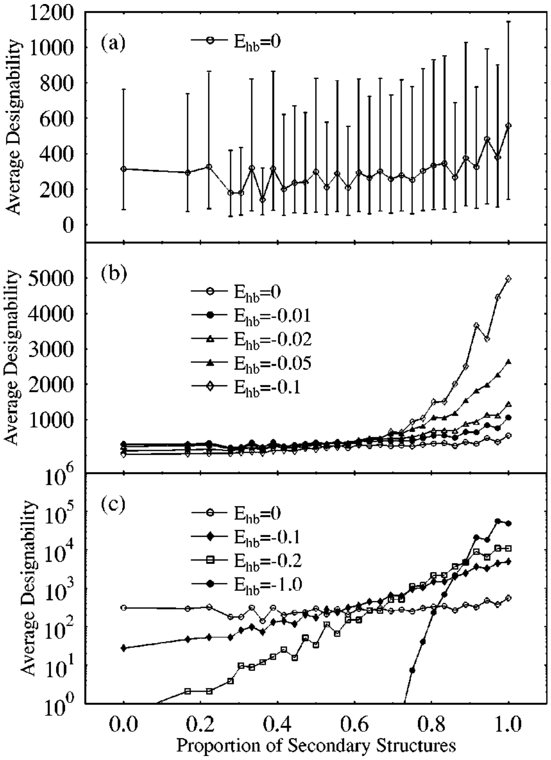

FIG. 3. Average designabilities of 6×6 compact conformations with the same proportion of secondary structures as a function of the proportion of the secondary structures. (a) $E_{hb}=0$ with standard deviation. (b) Small $|E_{hb}|$. (c) Large $|E_{hb}|$ with semilogarithm coordinates.

The number of sequences with nondegenerate native structures increases with decreasing $E_{hb}$. Thus the average designability $\langle N_s \rangle$, the number of sequences with nondegenerate native structure divided by the number of compact structures, increases too. The values are shown in Table I.

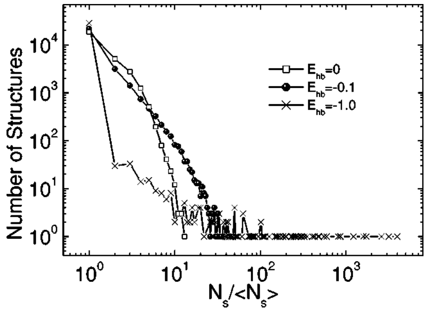

FIG. 4. Histogram of designabilities with $E_{hb}=0$, $-0.1$, and $-1.0$ obtained by random sampling of 22 982 400 sequences. The bin size is $\langle N_s \rangle$.

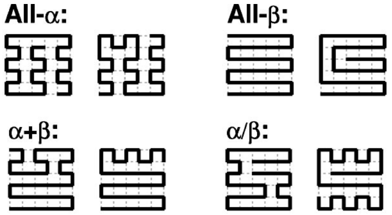

FIG. 5. Examples of highly designable compact conformations under the Hamiltonian with $E_{hb}=-1.0$ and the second definition of hydrogen bonds in secondary structures. They are classified into four classes.

This is understandable: with decreasing $E_{hb}$, the native structure candidates converge to structures with a high proportion of secondary structures, i.e., the native structure of a sequence is now selected from a smaller set, and the opportunity of having degenerate ground states decreases.

With the addition of the secondary-structure-favored and sequence-independent energy term in the Hamiltonian, the finding of Li *et al.* [12] becomes clearer. Figure 4 is a histogram of designabilities ($N_s$) of all $6\times6$ structures. The bin sizes for different $E_{hb}$ are set as $\langle N_s\rangle$. With decreasing $E_{hb}$, the designabilities of highly designable structures become larger, i.e., the tail of the histogram $N_s \gg \langle N_s\rangle$ becomes longer for lower $E_{hb}$.

To test how sensitively our result depends on the definition of the hydrogen bonds in secondary structures, we have repeated the calculation using another definition of the hydrogen bonds. Recalling the hydrogen bonds in a real $\alpha$ helix, where a hydrogen bond exists between the $i$th and the $(i+4)$th residues, we consider a second definition of the hydrogen bonds in secondary structures. We suppose that there are $n-3$ hydrogen bonds in an $\alpha$ helix of length $n$ with the definition of hydrogen bonds in $\beta$ sheets unchanged. The results are similar to those using the first definition.

The strength of the hydrogen bond is of the same order of magnitude as the hydrophobic interaction [17]. Therefore we focus our attention on the result at $E_{hb}=-1$. We find that only a small number of compact conformations have high designabilities and all of them have a high proportion of secondary structures. Recalling the classification of protein structures as folds, we try to classify the highly designable conformations. Under the first definition of the hydrogen bond, each topological contact in the secondary structure is considered as a hydrogen bond, and the hydrogen-bond-induced energy per bead in the $\beta$ sheets is lower than that in the $\alpha$ helices. Thus when $E_{hb}=-1$ most of the highly designable structures are composed of $\beta$ sheets. The second definition of the hydrogen bond agrees more closely with the structure of a real protein, and the hydrogen-bond-induced energies in $\alpha$ helices are nearly the same as in $\beta$ sheets. Thus the highly designable structures can be composed of both $\alpha$ helices and $\beta$ sheets. Some examples are shown in Fig. 5. According to the proportion of $\alpha$ helices and $\beta$ sheets and their connections, the highly designable conformations may be classified into four classes, all $\alpha$, all $\beta$, $\alpha+\beta$, and $\alpha/\beta$. Each class contains several different folds.

TABLE I. The percentage of sequences that have a unique native structure and the average designabilities $\langle N_s\rangle$ when $E_{hb}$ is set at different values.

| $E_{hb}$ | 0.0 | $-0.01$ | $-0.02$ | $-0.05$ | $-0.1$ | $-0.2$ | $-0.5$ | $-1.0$ |
|---|---|---|---|---|---|---|---|---|
| Seq. (%) | 36.0 | 48.4 | 52.0 | 57.4 | 60.7 | 70.3 | 77.7 | 81.0 |
| $\langle N_s\rangle$ | 288 | 387 | 416 | 459 | 486 | 562 | 622 | 648 |

## IV. DISCUSSION

In the traditional Hamiltonian of the HP model we have introduced a sequence-independent secondary-structure-favored energy term originating from the main chain hydrogen bonds in the secondary structures. It is found that this energy term strongly affects the designabilities of different conformations. The innate energy advantage of the structures with higher proportions of secondary structures makes them physical attractors [18] and highly designable. Therefore, the sequence-independent term in the Hamiltonian is decisive in the explanation of the existence of a large amount of regular secondary structure in proteins. The possible combinations of secondary structure elements are limited in compact conformations; therefore the emergence of only a small number of structure types (folds) among all possible structures is reasonable. This indicates that the structure-specific interaction term in the Hamiltonian plays an important role in the emergence of protein folds.

As a next step, one may consider different effective strengths for the hydrogen bonds in $\alpha$ helices and $\beta$ sheets. If we consider a 20-letter amino acid sequence, the effective strength of the hydrogen bonds in $\alpha$ helices and $\beta$ sheets may have different values for different amino acids according to their different tendencies to form $\alpha$ helices and $\beta$ sheets. Since the formation of a hydrogen bond then depends on the structure of its neighbors, the cooperativity in the folding process will be strengthened [9,10]. Also, the effects of hydrogen bonds on the formation of $\alpha$ helices and $\beta$ sheets need further study.

## ACKNOWLEDGMENTS

The authors thank Y. Zhang for discussions and critical comments. Numerical calculations were performed at the State Key Laboratory of Scientific and Engineering Computing and CHPCC.

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

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

[3] http://scop.mrc-lmb.cam.ac.uk/scop/

[4] C.A. Orengo, D.T. Jones, and J.M. Thornton, Nature (London) **372**, 631 (1994); A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995); C. Chothia, T. Hubbard, S. Brenner, H. Barns, and A. Murzin, Annu. Rev. Biophys. Biomol. Struct. **26**, 579 (1997).

[5] C. Chothia, Nature (London) **357**, 543 (1992); Z.-X. Wang, Protein Eng. **11**, 621 (1998).

[6] H.S. Chan and K.A. Dill, Macromolecules **22**, 4559 (1989).

[7] L.M. Gregoret and F.E. Cohen, J. Mol. Biol. **219**, 109 (1991); D.P. Yee, H.S. Chan, T.F. Havel, and K.A. Dill, *ibid.* **241**, 557 (1994); N.G. Hunt, L.M. Gregoret, and F.E. Cohen, *ibid.* **241**, 214 (1994); A. Kolinski and J. Skolnick, J. Chem. Phys. **97**, 9412 (1992); M.H. Hao, S. Rackovsky, A. Liwo, M.R. Pincus, and H.A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **89**, 6614 (1992); N.D. Socci, W.S. Bialek, and J.N. Onuchic, Phys. Rev. E **49**, 3440 (1994).

[8] L.F. Yan and Z.R. Sun, *Molecular Structure of Protein* (Tsinghua University Press, Beijing, 1999).

[9] H.S. Chan, Proteins **40**, 543 (2000).

[10] H. Kaya and H.S. Chan, Phys. Rev. Lett. **85**, 4823 (2000).

[11] U.H.E. Hansmann and Y. Okamoto, J. Chem. Phys. **110**, 1267 (1999).

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

[13] H. Li, C. Tang, and N.S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **95**, 4987 (1998); M.R. Ejtehadi, N. Hamedani, and V. Shahrezaei, Phys. Rev. Lett. **82**, 4723 (1999); R. Tatsumi and G. Chikenji, Phys. Rev. E **60**, 4696 (1999); C.T. Shih, Z.Y. Su, J.F. Gwan, B.L. Hao, C.H. Hsieh, and H.C. Lee, Phys. Rev. Lett. **84**, 386 (2000).

[14] P.D. Thomas and K.A. Dill, Protein Sci. **2**, 2050 (1993).

[15] K. Yue, K.M. Fiebig, P.D. Thomas, H.S. Chan, E.I. Shakhnovich, and K.A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 325 (1995); R. Backofen, S. Will, and E. Bornberg-Gauer, Bioinformatics **15**, 234 (1999); H. Chen, X. Zhou, and Z.C. Ou-Yang, Phys. Rev. E **63**, 31913 (2001).

[16] H.S. Chan and K.A. Dill, Proteins **24**, 335 (1996).

[17] G. Nemethy, M.S. Pottle, and H.A. Scheraga, J. Phys. Chem. **87**, 1883 (1983).

[18] L. Holm and C. Sander, Science **273**, 595 (1996).