Classification of amino acids based on statistical results of known structures and cooperativity 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 3 February 2002; revised manuscript received 6 March 2002; published 20 June 2002)

It has been found that the 20 kinds of amino acids have different frequencies of occurrence in α,β, and coil structures [P. Y. Chou and G. D. Fasman, Biochemistry **13**, 211 (1974)]. Based on more known structures of proteins, frequencies for each amino acid in α and β secondary structures are recalculated. Next step, under the approximation ignoring the chain connectivity of proteins, energy parameters to form α and β secondary structures for each amino acid are obtained. According to the hydrophobicity and energies in α and β secondary structures, 20 kinds of amino acids are classified. The results suggest that dividing amino acids to five or nine groups is desirable. At last, a protein model considering both two-body hydrophobic interaction and one-body energy to form secondary structures, hydrophobic-polar $\alpha\beta$ model, is introduced. It is shown that the consistency among various energy terms makes the cooperativity of protein folding closer to the experiments.

DOI: 10.1103/PhysRevE.65.061907 PACS number(s): 87.14.Ee, 87.15.Aa, 02.70.Rr

I. INTRODUCTION

Native structure of a protein is the global minimum of free energy [1,2]. To study protein folding problems theoretically, an appropriate description of both the structure of a protein and the intramolecular interactions is necessary. Proteins are composed of 20 different kinds of amino acids, each amino acid contains several atoms. The interactions between amino acids are very complicated. Molecular mechanics force field is useful for studying the structures and dynamics of proteins near their native states. However, it is not appropriate to study the complete conformation space of a protein because of the current computer capability. And it is difficult to tell us what kind of interaction is important for protein folding.

Now most of the theoretical studies of protein folding involve both simplified protein models and effective interactions. In simplified models, several atoms, such as a residue, are treated as a bead. The potentials of some models are artificial, such as Gō model [3] and random energy model [4], while some potentials come from statistical results of known structures of proteins. Now three-dimensional (3D) structures of more than ten thousand proteins have been obtained by x-ray or nuclear magnetic resonance experimental methods. Statistics of these known structures can tell us some information about the hidden interactions. Miyazawa and Jernigan obtained a two-body inter-residue interaction matrix (MJ matrix) from the numbers of residue-residue contacts observed in native structures of globular proteins [5]. The residues in the core of native structure have more residues in contact with them than the residues on the surface. Hydrophobic residues like to avoid water molecules and hide themselves in the core, while polar residues like to contact water molecules and appear on the surface. Therefore, in MJ matrix, interactions between hydrophobic residues are stronger than those between polar residues and those between hydrophobic residue and polar residue. From MJ matrix, Li *et al.* [6] found that the driving force of protein folding comes from hydrophobic interaction and a force of demixing, and the 20 kinds of amino acids can be classified to polar and hydrophobic amino acids, which is consistent with an HP (hydrophobic-polar) model.

More than 20 years ago, Chou and Fansman [7] found that different amino acids have different opportunities to appear in different regular secondary structures (α helix and β sheet). The result has been used to predict the secondary structures of polypeptides [8]. α helix and β sheet are both local conformations. Local conformation of main chain is described by the dihedral angles φ and ψ [1,9]. In Ramachandran plots, the (φ,ψ) distributions are clustered into three core regions corresponding to α, β, and α_L (left-hand α) structures. α_L region is much smaller than the other two. Therefore, different (φ,ψ) distributions for different amino acids reflect their different affinities to α helix and β sheet. The statistical result of Chou and Fansman is the coarse-grained result of Ramachandran plot and can be easily quantified. In this paper, like the work of Miyazawa and Jernigan [5], we can obtain a one-body energy term for each amino acid in α helices and β sheets. To obtain this secondary-structure-related energy term, more recent data are counted as Chou and Fansman.

One source of the complexity of protein folding problem is the fact that there are 20 kinds of amino acids. Traditionally, the 20 kinds of amino acids are classified to several classes (hydrophobic, charged, and polar) according to their chemical nature [10]. This classification is based on the chemical property of amino acids in water. But the environment around one amino acid in protein's native structure is different from water solution of amino acid monomers. Therefore, this classification is not appropriate to describe the roles of amino acids in protein's native structure. Recently, both experimentalists and theoreticians have interests to find out how many (less than 20) types of amino acids are enough to reconstruct protein's native structure [11–15]. Simplifying the 20-letter alphabet to a two-letter alphabet such as the HP model is the limit of simplification, because one-letter homopolymer does not have unique native structures like proteins. However, experimental work of Riddle and co-workers [11] showed that more than two kinds of amino acids are needed for proteins to fold to their native

structures. Theoretically, based on the MJ interaction matrix, Wang and Wang's work [14] showed that a five-letter alphabet may be a good choice to reconstruct protein's native structure and their result is in agreement with the suggestion of Riddle *et al*. Recently, work of Zheng and co-workers [15] based on MJ matrix and blocks substitution matrix (BLOSUM) [16] obtained different results. The result of experimental work is confined to the studied protein. Therefore, it is unknown if the result is the same for other proteins.

Most of the theoretical works are based on the simplified contact energy between residues. Recently, Vendruscolo and co-workers [17] tried to learn the energy parameters from crystal structures of proteins. They found that only pairwise contact interaction is not sufficient to stabilize the native states of proteins. The simplified amino acids alphabet and interactions not only need to stabilize the native structure of protein, but also need to satisfy the thermodynamic and kinetic property of real proteins. Folding transition of most global single domain proteins is calorimetric two state [18,19]. The quantitative criterion for calorimetric two-state transition is that the van't Hoff enthalpy, ΔH_{vH}, calculated at the peak of the specific heat, is approximately equal to the calorimetric enthalpy ΔH_{cal} of the entire transition, i.e., $\Delta H_{vH}/\Delta H_{cal} \approx 1$. Recently, it was found that the pairwise contact interaction is insufficient to satisfy the calorimetric criteria for two-state folding [20,21] even a 20-letter alphabet. So some other energy components are necessary.

In the present work, we try to classify the 20 kinds of amino acids according to their hydrophobicities and different affinities to α helix and β sheet. Conceptually based on the result of classification, we introduce a model of protein folding, in which not only two-body hydrophobic interaction like that in the HP model, but also a one-body energy component related to the formation of secondary structures is considered. This energy term can lead thermodynamic behaviors closer to experimental results.

II. ENERGY IN SECONDARY STRUCTURES

Since the work of Chou and Fansman [7], the data of protein 3D structures have increased many times. And some secondary databases have been established. Kabsch and Sander [22] designed a program to standardize secondary structure assignment by pattern recognition of hydrogen bonds. They established the database of secondary structure in proteins (DSSP) that assigns secondary structures for all protein entries in the protein data bank (PDB). Therefore, DSSP is an appropriate database for us to count the appearance frequencies of different amino acids in different secondary structures.

There are more than ten thousand entries in PDB, and most of them are homologous sequences. Therefore, we need not count all the proteins in PDB. Hobohm and Sander [23] established the pdbselect database, a subset of PDB that does not contain homologous sequences. Pdbselect database offers a representative selection that is about a factor of 5 or 6 smaller than PDB database. Here we use the pdbselect database Feb. 2001 release. It contains 1520 chains composed of proteins, DNAs, and RNAs. We download the corresponding 1419 protein entries in DSSP, and count amino acids in the chains included in the pdbselect database. The results are shown in Table I. The original work of Chou and Fansman only involved 15 proteins, and the number of all amino acids is 2473. Now the sampler space is 100 times bigger with total number of amino acids 248 848. Here we consider only α helices and β sheets, and all others are thought to be coil. The result is approximately consistent with that of Chou and Fasman [7] (Fig. 1). However, in the present results most frequencies of amino acids in β sheets are bigger, and the frequencies in α helices are smaller than the results of Chou and Fasman. Because the number of samples is 100 times more, the present results are more accurate.

TABLE I. The statistical results for 20 kinds of amino acids. The hydrophobic parameter h_i comes from the paper of Li *et al*. [6].

Name	No. of residues	Residues in α helix	Residues in β sheet	E_α (k_BT)	E_β (k_BT)	h_i (k_BT)
Ala (A)	19376	8728	3093	−1.56	−0.91	−1.43
Arg (R)	12274	4589	2374	−1.27	−0.99	−0.85
Asn (N)	11314	2590	1488	−0.39	−0.22	−0.87
Asp (D)	14419	3817	1587	−0.56	−0.06	−0.61
Cys (C)	4526	1008	1288	−0.62	−1.25	−2.34
Gln (Q)	9834	4032	1630	−1.38	−0.86	−0.69
Glu (E)	16135	6939	2430	−1.44	−0.77	−0.55
Gly (G)	18127	2445	2556	0.26	−0.16	−0.99
His (H)	5785	1579	1203	−0.77	−0.88	−1.33
Ile (I)	14065	4829	5187	−1.59	−2.05	−3.27
Leu (L)	21713	9232	5161	−1.65	−1.45	−3.70
Lys (K)	15612	5495	2720	−1.12	−0.80	−0.42
Met (M)	5330	2131	1122	−1.44	−1.18	−2.79
Phe (F)	9959	3108	3094	−1.23	−1.60	−3.65
Pro (P)	11383	1410	1032	0.43	0.36	−1.03
Ser (S)	15152	3620	2684	−0.52	−0.61	−0.80
Thr (T)	14202	3425	3667	−0.69	−1.14	−1.05
Trp (W)	3621	1226	992	−1.28	−1.45	−2.57
Tyr (Y)	8819	2593	2837	−1.15	−1.62	−2.07
Val (V)	17202	5011	6931	−1.37	−2.07	−2.70
All	248848	77807	53076	−1.00	−1.00	

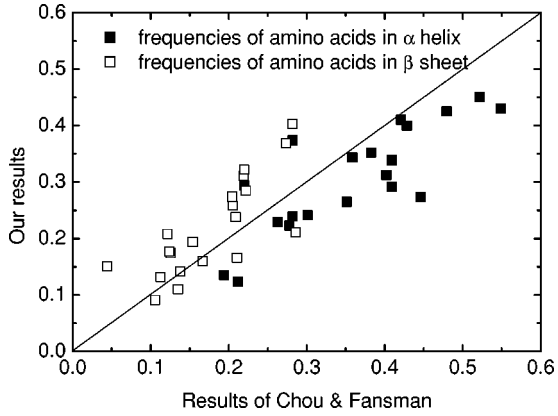


FIG. 1. The statistical results of Chan and Fasman vs our results based on more structural data.

Under coarse-grained treatment of local conformations, each amino acid has three states: α helix, β sheet, and coil. Not considering the 3D structure of a protein, we pay attention to the secondary structures in the protein's native structure. Then the native structure can be described by a one-dimensional vector whose component is the state of each amino acid. There are three kinds of states (α, β, and coil) the amino acid can occupy. We assume that there are three effective energy levels for each amino acid according to α helix, β sheet, and coil states, respectively. The degeneracies of the three states hold as constants for all amino acids, while the three energy levels are different for different amino acids.

Each amino acid distributes in these three kinds of states, the distribution can be obtained from the observed native structures. The observed distributions of amino acids in the three states can reflect the effective energy levels.

In a certain protein, the state of each amino acid is determined not only by the amino acid itself, but also by the whole protein sequence. To stabilize the native structure, there are correlations in the sequence of a protein [24]. But, the statistical results are based on more than one thousand proteins. Therefore, there is no significant contribution from the specific amino acid sequence of a certain protein in the statistical results. The statistical results can reflect the three hidden effective energy levels for each amino acid. The approximation of ignoring the correlation in a sequence, or the chain connectivity, can be treated as a first-order approximation to obtain the hidden effective energy levels from the known structures. The work of Miyazawa and Jernigan [5] is also based on this approximation.

First, not distinguishing which amino acid is concerned, we introduce a partition function to describe the amino acids distributing in the states of α helix, β sheet, and coil (c). The partition function takes the form

$$Z=\sum_{i=\alpha,\beta,c}\Omega_i \exp(-E_i), \tag{1}$$

where Ω_i is the degeneracy of the ith kind of secondary structure (Ω_i is normalized, so that $\Sigma_{i=\alpha,\beta,c}\Omega_i=1$), E_i is the energy advantage of the ith kind of secondary structure with unit k_BT_0. k_B is Boltzmann constant and T_0 is absolute physiological temperature. Similarly, we can construct the partition function of one kind of amino acid. The degeneracies of the three states are the same for all amino acids. The partition function of the kth amino acid takes the form

$$Z_k=\sum_{i=\alpha,\beta,c}\Omega_i \exp(-E_i^k), \tag{2}$$

where E_i^k is the energy of amino acid k in the ith secondary structure.

In regular secondary structures (α helix and β sheet), hydrogen bonds always exist between the main chains. The hydrogen bonds are important for the stability of regular secondary structures. 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 [25,26]. According to the strength of the hydrogen bond [27], we set the average energy advantage of one amino acid in regular secondary structures, such as α helix and β sheet, as -1; and energy of amino acid in coil as zero, i.e., $E_\alpha=E_\beta=-1$, $E_{coil}=E_{coil}^k=0$. We can use the statistical results above to determine the other parameters Ω_i and E_i^k. We have the equation

$$\frac{\Omega_i \exp(-E_i)}{Z}=\frac{n_i}{n_{all}}, \tag{3}$$

where n_{all} is the number of all amino acids, and n_i is the number of amino acids in the ith secondary structure (Table I). Therefore, we get $\Omega_\alpha=17.2\%$, $\Omega_\beta=11.8\%$, and $\Omega_{coil}=71.0\%$. In the early work of Chan and Dill [28], under 2D lattice model, it is found that the average proportion of secondary structure is high (about 50% to 70%) in the compact conformations. But under an off-lattice model, Socci *et al.* [29] found that compactness is not sufficient to create secondary structures. In the present result, the proportion of secondary structures in compact conformations (Ω_α and Ω_β) is smaller than the result of Chan and Dill based on lattice model, and bigger than the result of Socci *et al.* based on off-lattice model. It indicates that real protein is a hybrid of lattice and off-lattice models, because the chemical bonds have favorable directions.

Similar to Eq. (3), we have the equation

$$\frac{\Omega_i \exp(-E_i^k)}{Z_k}=\frac{n_i^k}{n_{all}^k}, \tag{4}$$

where n_{all}^k is the number of all the kth kind of amino acid, and n_i^k is the number of the kth kind of amino acid in the ith secondary structure. Thus we can get the value of E_i^k (Table I).

From the work of Li *et al.* [6], hydrophobicity of a residue is related to a parameter h_i. We show E_i^α and E_i^β as functions of h_i in Fig. 2. E_i^α and E_i^β for most of the hydrophobic residues ($h_i<-2.0$) are smaller than -1, which means that most hydrophobic residues tend to form α helix and β sheet. For polar residues ($h_i>-1.5$), the distribution of E_i^α and E_i^β is wide. Some residues tend to form α helix and β sheet ($E_i^\alpha, E_i^\beta<-1.0$), while some tend to break α helix and β sheet. Even E_i^α and E_i^β of proline and E_i^α of glycine are bigger than zero. This result is reasonable: most coils are on the surface and polar residues tend to appear on the surface, so on average, E_i^α and E_i^β of polar residues are bigger than those of hydrophobic residues that tend to hide themselves in the core.

III. CLASSIFICATION OF AMINO ACIDS

The work on classification of amino acids of Wang and Wang [14] is completely based on MJ matrix. Therefore, their classification is mainly based on hydrophobicity of

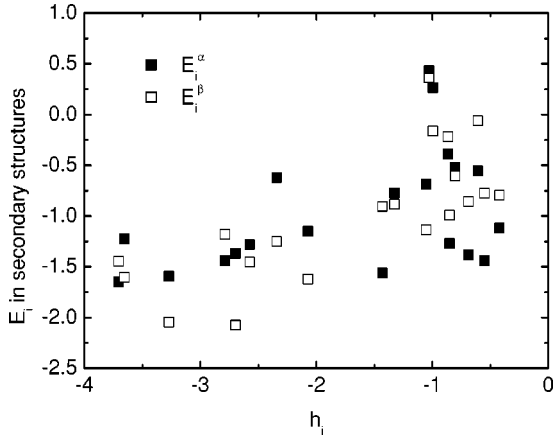


FIG. 2. E_i^{α} and E_i^{β} as functions of h_i for 20 kinds of amino acids.

amino acids [6]. The structural characteristics of a residue in protein's native structure include: (1) Is the residue on the surface or in the core? (2) Is the residue part of α helix, β sheet, or coil? The first question is related to the residue's hydrophobicity, and the second one is related to its energies in α helix and β sheet, which are calculated in the former section.

Here we use both the hydrophobicities of amino acids and their different energies in α and β structures to classify the 20 kinds of amino acids. Now each amino acid is related to a three-dimensional energy vector whose components include hydrophobicity h_i, energy in α helix E_{α}, and energy in β sheet E_{β}. The methods to obtain these parameters are similar (method of MJ and method in the former section), and their units are all $k_B T_0$. Therefore, we can treat these three components equally to classify the 20 kinds of amino acids. We use the optimization clustering algorithm [30] to classify amino acids.

In the optimization clustering method, the number of groups is fixed as g, and we minimize the target function

$$E = \sum_{m=1}^{g} \sum_{l=1}^{n_m} d_{ml,m}^2, \tag{5}$$

where $d_{ml,m}$ is the Euclidean distance between the lth amino acid in the mth group and the centroid of the group, n_m is the number of amino acids in mth group. We use the multicanonical Monte Carlo (MC) algorithm [31] to obtain the global minimum of E, $E_{\min}$. The details of the method will appear in another paper [32].

The results are shown in Table II. When $g=19$, 18, glutamine (Q) and glutaminic acid (E), asparagine (N) and aspartic acid (D) agglomerate first. Because structures of Q and E, N, and D are similar, their roles in the native structures of proteins are similar too. Though E and D(Q and N) are both charged (polar) residues, they do not aggregate to the same group until $g=2$ when the 20 amino acids are classified to hydrophobic and polar groups. For protein folding problem, the classification of amino acids by chemical nature is not appropriate to simplify the amino acid alphabet. The size and structure of the side chain are also important

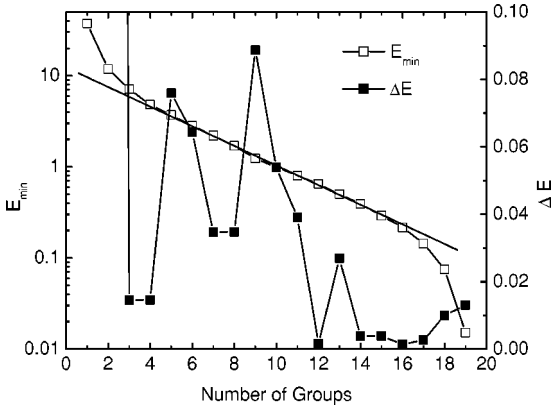


FIG. 3. Results of the classification of 20 amino acids using optimization clustering method. Open squares show the E_{min} [Eq. (5)] as a function of the number of groups (according to left axis), and filled squares show the ΔE as a function of the number of groups (according to right axis).

factors to determine the behavior of amino acid. Our classification is based on the behavior of the amino acids, and the structure information is not needed.

In optimization clustering method, $E_{\min}$ decreases with increasing the number of groups g. Plotting $E_{\min}$ against g can give some suggestive ''best'' number of groups. $E_{\min}$ as a function of g is shown in Fig. 3. We can see when $g=5$ to 15, there is a linear relation between $\ln(E_{\min})$ and g. When $g=1$ to 4, $\ln(E_{\min})$ is bigger than the linear prediction, while when $g=16$ to 19, $\ln(E_{\min})$ is smaller than the linear prediction. Therefore, the result indicates that a proper number of groups is 5 which is the same as Wang's result, but the partition is different.

In multicanonical MC simulation, not only the best partition with $E_{\min}$, but also the partition with E just above $E_{\min}$ (E_1) is obtained. Similar to energy level structures, we call the best partition with $E_{\min}$ ground state, and call the partition with E_1 the first excited state. The gap $\Delta E = E_1 - E_{\min}$ as a function of the number of group g is shown in Fig. 3 too. The bigger ΔE is, the more reliable the best partition is. The point $\Delta E(g=2)=1.21$ is much bigger than the others, thus it is out of the figure. Therefore, the best partition with $g=2$ is very robust. In this partition, 20 amino acids are divided to hydrophobic group and polar group, and the partition is the same as that of Li, *et al.* [6]. This result confirms that hydrophobic interaction is the most important driving force for protein folding. The simplest heteropolymer model of protein, HP model, grasps the most important driving force for protein folding. For other partitions with $g>2$, there are two peaks on the $\Delta E(g)$ curve, one peak with $g=5$, the other with $g=9$. The two peaks indicate that maybe dividing 20 amino acids into 5 or 9 groups is desirable. The suggestive group number 5 is in agreement with the result discussed in the preceding paragraph. We can also find that $\ln(E_{\min})$ with $g=9$ is just a little smaller than the linear prediction.

IV. LATTICE MODEL AND CALORIMETRIC COOPERATIVITY

According to the above results, we introduce a model of protein folding which is an extension of the HP model or the

TABLE II. Results of optimization clustering method based on h_i, E_i^{α}, and E_i^{β} of each amino acid. The first column indicated the number of groups.

1	A	H	T	R	Q	E	K	N	D	S	G	P	C	Y	M	W	V	I	L	F
2	A	H	T	R	Q	E	K	N	D	S	G	P								
3													C	Y	M	W	V	I	L	F
4	A	H	T	R	Q	E	K	N	D	S	G	P								
													C	Y	M	W	V	I	L	F
5	A	H	T	R	Q	E	K	N	D	S	G	P								
													C	Y	M	W	V	I	L	F
6	A	H	T	R	Q	E	K	N	D	S	G	P								
													C	Y	M	W	V	I	L	F
7	A	H	T	R	Q	E	K	N	D	S	G	P								
													C	Y	M	W		I	V	F
8	A	H	T	R	Q	E	K	N	D	S	G	P								
													C	Y	M	W		I	V	F
9	A	H	T	R	Q	E	K	N	D	S	G	P								
													C	Y		M	W	I	V	F
10	A	H	T	R	Q	E	K	N	D	S	G	P								
													C	Y	M	W		I	V	F
11	A	H	T	R	Q	E	K	N	D	S	G	P								
													C	Y	M	W		I	V	F
12	A	H	T	R	Q	E	K	N	D	S		G	P	C	Y	M	W	I	V	F
13	A	H	T	R	Q	E	K	N	D		S	G	P	C	Y	M	W	I	V	F
14	A	H	T	R	Q	E	K	N	D	S	G	P	C	Y	M	W		I	V	F
15	A	H	T	R	Q	E	K	N	D	S	G	P	C	Y	M	W		I	V	F
16	A	H	T	R	Q	E	K	N	D	S	G	P	C	Y	M	W		I	V	F
17	A	H	T	R	Q	E	K	N	D	S	G	P	C	Y	M	W		I	V	F
18	A	H	T	R	Q	E	K	N	D	S	G	P	C	Y	M	W		I	V	F
19	A	H	T	R	Q	E	K	N	D	S	G	P	C	Y	M	W		I	V	F
20	A	H	T	R	Q	E	K	N	D	S	G	P	C	Y	M	W		I	V	F

"helical-HP model" of Thomas and Dill [33]. We call it HP-$\alpha\beta$ model. In HP-$\alpha\beta$ model, there are six kinds of residues, H-α, H-β, H-coil, P-α, P-β, and P-coil residues. These six letters do not correspond to the classification result in the former section. They are conceptually artificial. The Hamiltonian of a given sequence $\{\sigma_i\}$ now takes the form

$$H=\sum_{i<j} E_{\sigma_i\sigma_j}C(i,j)+\sum_i E_{\alpha}^{\sigma_i}A_i+\sum_i E_{\beta}^{\sigma_i}B_i, \qquad (6)$$

where the first term is the same as that of the HP model, which comes from the hydrophobic interaction; and the second and third terms are the energy to form α and β secondary structures. $E_{\sigma_i\sigma_j}$, $E_{\alpha}^{\sigma_i}$, and $E_{\beta}^{\sigma_i}$ are determined by the six-letter sequence $\{\sigma_i\}$, while $C(i,j)$, A_i, and B_i come from the conformation. $E_{\sigma_i\sigma_j}$ is hydrophobic interaction between residue σ_i and σ_j, the values are set as: $E_{HH}=-3.3$, $E_{HP}=-2.0$, and $E_{PP}=-1.0$ [34,35]. Considering the relative strength of hydrogen bonds and hydrophobic interactions, we set $E_{\alpha}^{\sigma_i}=-0.5$ for H-α and P-α residues, and $E_{\alpha}^{\sigma_i}=0$ for other residues. Similarly, $E_{\beta}^{\sigma_i}=-0.5$ for H-β and P-β residues, and $E_{\beta}^{\sigma_i}=0$ for the others. If the ith and jth residues are nearest neighbors in the conformation and i,j are not adjacent along the chain, $C(i,j)=1$, and $C(i,j)=0$ otherwise. $A_i=1$ $(B_i=1)$ if the ith residue is part of an α helix (β sheet), and zero otherwise.

We use 2D square lattice model and enumeration method to obtain thermodynamics property of HP-$\alpha\beta$ model. In 2D square lattice, a protein is simplified as a sequence of beads in self-avoiding-walk conformation. α helices and β sheets can be identified by their special patterns in the *contact map* [26,28]. A conformation of a chain with 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 a topological contact between the ith and the jth beads, the corresponding matrix element $C(i,j)$ of the contact map is 1, and otherwise it is 0. Matrix element $C(i,j)$ is just the same as $C(i,j)$ in Eq. (6). Figure 4 shows an example of a compact conformation together with its contact map. The patterns of α helix and β sheet are indicated by arrows. In the present work, the smallest α helix is composed of six beads, and the smallest β sheet (parallel and antiparallel) is composed of four beads. Under this definition, there is the case in which one bead is both a part of an α helix and a part of a β sheet, which is not true in real protein. Therefore, in this case, we set the bead apart of the α helix, not apart of the β sheet.

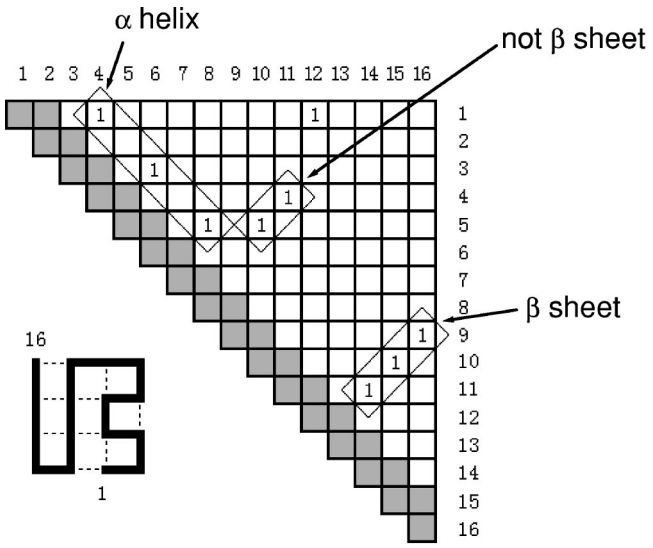
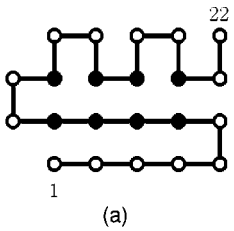


FIG. 4. An example of conformation together with its contact map (open lattice represents 0). The specific patterns of α helix and β sheet are indicated by arrows. Contacts (4,11) and (5,10) do not form a β sheet because residues 4 and 5 are part of α helix.

Proteins are not random sequences of amino acids [35], and they are a small subset of all possible sequences. The native conformation of protein must be the energy minimum in a funnel-like energy landscape. The easiest way to obtain a proteinlike sequence is to design a sequence with a target conformation as its native conformation [36]. Here we select a target conformation including both α helix and β sheet, and there is a small coil connecting them [Fig. 5(a)]. To design a sequence with the target conformation as its native state, we select the simplest design strategy: the coloring method [37], which determines the type of each unit only according to the position of the unit in the target conformation. We assign H residues to the units in the core, and P residues to those on the surface. Also, the units in α helix (β sheet) are assigned as H-α or P-α (H-β or P-β) residues, and the other units are assigned as H-coil or P-coil residues. The designed HP-$\alpha\beta$ sequence is shown in Fig. 5(b). H-coil letter does not appear in the designed sequence. The de-



(a)

P-β,P-β,P-β,P-β,P-coil,P-coil,H-β,H-β,H-β,H-β,P-coil,
P-coil,H-α,P-α,P-α,H-α,H-α,P-α,P-α,H-α,P-α,P-α.

(b)

FIG. 5. (a) The target conformation with both α helix and β sheet. The color of units indicts the designed HP sequence (black-H, white-P). (b) The designed sequence for HP-$\alpha\beta$ model.

signed HP sequence is shown in Fig. 5(a) together with the target conformation. The target conformation is the ground state for both the designed HP-$\alpha\beta$ sequence and the designed HP sequence.

We enumerate all the 301 064 158 conformations and calculate the thermodynamic properties of the designed sequences by the standard formulas of canonical ensemble. We compare the results of HP and HP-$\alpha\beta$ model. Figure 6 shows the specific heat capacity of the two designed sequences and the probability of native conformation. The transition is much sharper for HP-$\alpha\beta$ model.

Experimentally, most small single domain proteins can be described by a two-state model. Often it can be found that the van't Hoff enthalpy ΔH_{vH} around the transition midpoint is approximately equal to the calorimetric enthalpy ΔH_{cal} of the entire folding transition [20,21]. $\Delta H_{vH}/\Delta H_{cal}$ takes the form

$$\Delta H_{vH}/\Delta H_{cal}=2T_{max}\sqrt{k_B C(T_{max})}/\Delta H_{cal}, \qquad (7)$$

where specific heat capacity $C(T)$ is maximum at $T=T_{max}$, k_B is Boltzmann constant that is set as 1 in our calculation. It has been found that contact energies cannot reproduce the calorimetric two-state picture [20,21]. Even some works show that there are three phases: random coil, molten globule, and native conformation [35,38]. Thus, there will be two transitions with decreasing temperature. From Fig. 6, we can see that the transition of HP model happens at a wide temperature range of more than one order of magnitude. The C_v curve for the HP model is nearly the sum of two Gaussian functions that reflect the two transitions, one from random coil to molten globule and the other from molten globule to native conformation. While the C_v curve for HP-$\alpha\beta$ model is one Gaussian function, and the transition is much sharper. $\Delta H_{vH}/\Delta H_{cal}$ is 0.184 for the HP sequence, and 0.276 for the HP-$\alpha\beta$ sequence. Though the HP-$\alpha\beta$ sequence here does not satisfy the calorimetric two-state criterion, it has changed a three-state picture of HP model to a two-state picture. The size of the example studied here is only 22 that is very small. If we study a 3D model with the size big enough, the transition will be much sharper, and the thermodynamic behavior will be closer to experimental results.

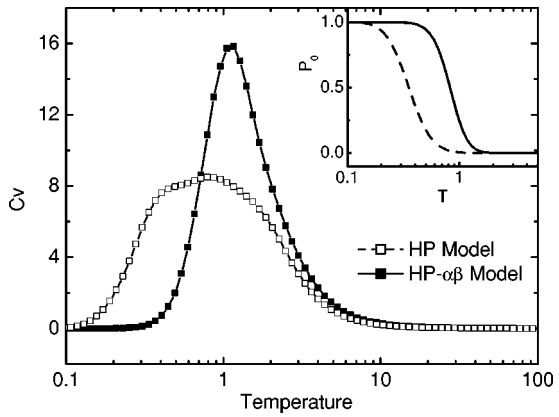


FIG. 6. Heat capacity C_v as a function of temperature for both the designed HP sequence and HP-$\alpha\beta$ sequence in Fig. 5. The inner figure shows the proportion of native structure (P_0) as a function of temperature.

V. DISCUSSION

From the known structural information of proteins, the secondary structure related energy terms E_i^α and E_i^β for each amino acid are obtained. Here the correlation in a sequence is ignored, and the obtained energy terms E_i^α and E_i^β are only related to the state of one amino acid. But the state of one amino acid in a protein is influenced by the whole sequence, especially the neighboring amino acids. Under the approximation of ignoring the chain connectivity, the chain connectivity is treated as the environment of amino acids that induces the distribution of amino acids in the three states (α, β, and coil). There is a typical length for α helix and β sheet, i.e., the formation of α helix (β sheet) needs several continuous α helix (β sheet) favorable amino acids. Therefore, the chain connectivity is also important for the formation of α helix and β sheet. The effect of the chain connectivity on the formation of secondary structures needs further study. A simple extension of the present work is to study the structure of two or three adjacent amino acids along protein sequence.

Based on the physical source, the interactions in proteins can be classified to bond energy (mainly bond angle) and noncovalent interactions (van der Waals, electrostatic, hydrogen bonds, and hydrophobic interactions). These interactions all serve to stabilize protein's native structure. There are not only long-range interactions between two residues apart from each other along the sequence, but also short-range interactions related to the conformations of the residue itself and the adjacent residues along the sequence. Long-range interactions mainly come from noncovalent interactions (disulfide bond between Cys residues also belongs to long-range interaction), while short-range interactions mainly come from bond angle energies and hydrogen bonds between sequence-adjacent residues. The widely used MJ matrix only grasps long-range interactions. In our work, energies in secondary structures E_i^α and E_i^β are short-range interactions that are related to the local conformation of the chain. Upon residue's hydrophobicity, E_i^α and E_i^β, the classification of amino acids is consistent with some known results, and gives some sug-

gestions. The classification is based on the behaviors of amino acids in protein's native structure. Our results are different from the traditional classification of amino acids based on their chemical nature and side chain structures. The reason is that the environment around one amino acid in protein's native structure is different from water solution of amino acid monomers. Our classification is more instructive to protein folding problem.

The consistency of long-range interaction and short-range interaction makes the thermodynamic behavior of model proteins closer to experiments [39,40]. It indicates that cooperativity of protein folding comes partly from the consistency among various energy terms, such as long-range and short-range interactions.

ACKNOWLEDGMENTS

The authors thank C. Tang for the sharing of data and useful suggestions. Numerical calculations are performed at the State Key Lab 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] N. Gō and H. Abe, Biopolymers **20**, 991 (1981).
[4] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[5] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985); J. Mol. Biol. **256**, 623 (1996).
[6] H. Li, C. Tang, and N. S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[7] P. Y. Chou and G. D. Fasman, Biochemistry **13**, 211 (1974); **13**, 222 (1974).
[8] B. Rost and C. Sander, Annu. Rev. Biophys. Biomol. Struct. **25**, 113 (1996).
[9] C. M. Venkatachalam and G. N. Ramachandran, Annu. Rev. Biochem. **38**, 45 (1969).
[10] *Introduction to Protein Structure*, edited by C. Branden and J. Tooze (Garland, New York, 1999).
[11] D. S. Riddle, J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi, and D. Baker, Nat. Struct. Biol. **4**, 805 (1997).
[12] P. G. Wolynes, Nat. Struct. Biol. **4**, 871 (1997).
[13] H. S. Chan, Nat. Struct. Biol. **6**, 994 (1999).
[14] J. Wang and W. Wang, Nat. Struct. Biol. **6**, 1033 (1999).
[15] W. M. Zheng, e-print physics/0106074; X. Liu, J. Qi, D. Liu, and W. M. Zheng (private communication).
[16] S. Henikoff and J. G. Henikoff, Proc. Natl. Acad. Sci. U.S.A. **89**, 10915 (1992).
[17] M. Vendruscolo, R. Najmanovich, and E. Domany, Phys. Rev. Lett. **82**, 656 (1999); C. Clementi, M. Vendruscolo, A. Maritan, and E. Domany Proteins **37**, 544 (1999).
[18] P. L. Privalov and N. N. Khechinashvili, J. Mol. Biol. **86**, 665 (1974).
[19] P. L. Privalov, Annu. Rev. Biophys. Biophys. Chem. **18**, 47 (1989).
[20] H. Kaya and H. S. Chan, Phys. Rev. Lett. **85**, 4823 (2000).
[21] H. S. Chan, Proteins **40**, 543 (2000).
[22] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983); http://www.sander.ebi.ac.uk/dssp/
[23] U. Hobohm, M. Scharf, R. Schneider, and C. Sander, Protein Sci. **1**, 409 (1992); U. Hobohm and C. Sander, *ibid.* **3**, 522 (1994); http://www.cmbi.kun.nl/swift/pdbsel/
[24] E. N. Govorun, V. A. Ivanov, A. R. Khokhlov, P. G. Khalatur, A. L. Borovinsky, and A. Y. Grosberg, Phys. Rev. E **64**, 040903 (2001).
[25] L. F. Yan and Z. R. Sun, *Molecular Structure of Protein* (Tsinghua University Press, Beijing, 1999).
[26] H. Chen, X. Zhou, and Z. C. Ou-Yang, Phys. Rev. E **64**, 041905 (2001).
[27] G. Nemethy, M. S. Pottle, and H. A. Scheraga, J. Phys. Chem. **87**, 1883 (1983).
[28] H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (1989).
[29] N. D. Socci, W. S. Bialek, and J. N. Onuchic, Phys. Rev. E **49**, 3440 (1994).
[30] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis* (Oxford University Press, New York, 2001).
[31] B. A. Berg and T. Neuhaus, Phys. Rev. Lett. **68**, 9 (1992).
[32] H. Chen, X. Zhou, and Z. C. Ou-Yang (unpublished).
[33] P. D. Thomas and K. A. Dill, Protein Sci. **2**, 2050 (1993).
[34] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[35] H. Chen, X. Zhou, and Z. C. Ou-Yang, Phys. Rev. E **63**, 031913 (2001).
[36] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993); E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[37] A. R. Khokhlov and P. G. Khalatur, Phys. Rev. Lett. **82**, 3456 (1999).
[38] A. Dinner, A. Šali, M. Karplus, and E. Shakhnovich, J. Chem. Phys. **101**, 1444 (1994).
[39] N. Gō and H. Takeromi, Proc. Natl. Acad. Sci. U.S.A. **75**, 559 (1978).
[40] N. Gō, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).