# Difference between ''proteinlike'' and ''nonproteinlike'' heteropolymers

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 20 April 2000; revised manuscript received 7 August 2000; published 27 February 2001)

Based on a simple two-dimensional (2D) hydrophobic-polar (H-P) lattice model, properties of amino acid chains are studied by enumeration and Monte-Carlo simulation methods. Among them some chains with large average energy gap ($\overline{E_g}$) are thought to be ''proteinlike'' while the others are ''nonproteinlike.'' The large $\overline{E_g}$ between the low excited conformations and the native conformation guarantees not only the thermodynamic stability of protein but also its fast-folding property. The phase transition from molten globule to the native conformation for the ''proteinlike'' polymer is found to be of first order, while that for the ''nonproteinlike'' polymer is not. Some properties of chains as a function of $\overline{E_g}$ shows that the transition from ''nonproteinlike'' to ''proteinlike'' heteropolymers is continuous. The simulation of folding at different temperature indicates that the main reason why some polymers fold slowly to its native conformation is their low folding temperature which makes the effective energy barrier ($E_b/T_f$) much higher than ''proteinlike'' chains.

DOI: 10.1103/PhysRevE.63.031913 PACS number(s): 87.14.Ee, 87.15.Cc, 05.20.−y, 02.70.Rr

## I. INTRODUCTION

Proteins are important biological macromolecules which are linear heteropolymers composed of 20 kinds of different amino acids. They have important biological functions for life and these functions depend highly on their three-dimensional (3D) structures [1]. Anfinsen [2] concluded that folded structure information of protein is coded in the amino acid sequence and the folded structure is the global minimum of the free energy. This is called the thermodynamic hypothesis. For single domain globular proteins, the length of the chain is of the order of 100. Thus the number of possible sequences is astronomical ($20^{100}$). Are the natural protein sequences being selected randomly from the possible sequences composed of the 20 kinds of amino acids or there are only a special part of them can act as protein? Now the trend is in favor of the latter [3,4]. Shakhnovich and co-workers [5,6] had developed a sequence design method to design sequences with some proteinlike properties. Natural proteins are products of evolution, therefore they must have unique properties distinguished from random sequences of amino acids. As has been pointed out by Kardar [7], proteins must satisfy the following conditions: (i) proteins must have a nondegenerate and thermodynamically stable native conformation; (ii) proteins must have the ability to fold quickly to the native structure from a stretched conformation.

Due to the above necessary properties of proteins, only a small part of heteropolymers can act as proteins. In this paper they are called ''proteinlike'' heteropolymers and the others are called ''nonproteinlike'' heteropolymers. As much as we know, much work focused on the ''proteinlike'' heteropolymers has been done. But what about the properties of ''nonproteinlike'' heteropolymers and what about the difference between properties of ''proteinlike'' and ''nonproteinlike'' heteropolymers?

The protein's thermodynamic properties, such as phase transitions, have been studied [8–10]. The phases of heteropolymers include random coil, molten globule, and native conformation [11]. The random coil consists of a very large number of rapidly interconverting conformations. The molten globule is composed of a large number of fluctuating relatively compact conformations. And the native conformation fluctuates only around its own neighborhood. It has been pointed out that the phase transition of the proteins from molten globule to native conformation is of first order [8]. How about the phase transition of ''nonproteinlike'' heteropolymers and what is the difference between the ''proteinlike'' and ''nonproteinlike'' heteropolymers?

The thermodynamic stability of native conformation and the fast-folding process are both necessary properties for proteins. It has been demonstrated that the thermodynamic stability of native conformation can also solve the problem of kinetic accessibility of native conformation, thus the fast-folding property can also be satisfied. What determines that the proteins have these properties? And what about the ''nonproteinlike'' heteropolymers?

## II. MODEL AND METHOD

A real protein is very complex; there are 20 different kinds of residues with each residue composed of many atoms. Up to now there has not been a theoretical method which can find the ground state of a real amino acid sequence reliably. To study what kind of amino acid sequences can act as proteins is even more difficult. From the work of Dill and co-workers [12,13], the lattice model has been widely used to study the thermodynamic properties and folding process of protein. In some lattice models [11,14], the interactions between monomers are set as random values satisfying a Gaussian probability distribution function. Analysis of the interactions between the 20 kinds of amino acids from the Miyazawa-Jernigan (M-J) matrix [15,16] shows that the amino acids can be divided into two different kinds: the hydrophobic (H) and the polar (P) amino acids, according to their affinity to water [17,18]. In this paper, we use the H-P model in a two-dimensional (2D) square lattice which includes the following points. (i) The protein is simplified as a heteropolymer composed of H and P monomers and the

 

monomers are treated as beads linked by covalent bonds to form a chain. (ii) The monomers can only occupy the 2D square lattice sites, and two or more monomers cannot occupy the same site. (iii) Only interactions between nearest-neighbor monomers are considered which depend on the monomer type (H or P).

Because only interactions between the monomers which are adjacent in the position and not adjacent in the sequence are taken into consideration, the Hamiltonian of a given sequence $\{\sigma_i\}$ is taken in the form

$$H=\sum_{i<j} E_{\sigma_i\sigma_j}\Delta(\mathbf{r}_i,\mathbf{r}_j), \tag{1}$$

where $\sigma_i$ and $\sigma_j$ represent H or P, $E_{\sigma_i\sigma_j}$ represents $E_{\text{H-H}}$, $E_{\text{H-P}}$, or $E_{\text{P-P}}$, the energy of H-H, H-P, or P-P interaction, respectively, and $\mathbf{r}_i$ is the position of the ith monomer and describes a self-avoiding-walk (SAW) conformation of a chain. Therefore if $\mathbf{r}_i$ and $\mathbf{r}_j$ are nearest-neighboring sites and $i$, $j$ are not adjacent along the chain, $\Delta(\mathbf{r}_i,\mathbf{r}_j)=1$ and $\Delta(\mathbf{r}_i,\mathbf{r}_j)=0$ otherwise.

The main driving force of the protein folding is hydrophobicity. Most H residues are buried in the core of the native conformation of protein and most P residues are exposed on the surface. So there is the relation $E_{\text{H-H}}<E_{\text{H-P}}<E_{\text{P-P}}$. Because different types of monomers tend to separate from each other, therefore the condition $2E_{\text{H-P}}>E_{\text{P-P}}+E_{\text{H-H}}$ should be satisfied [19]. According to the analysis of the M-J interaction matrix [16] by Wang and co-workers [18], among the 20 kinds of residues, the H residues include eight kinds: Cys, Met, Phe, Ile, Leu, Val, Trp, Tyr, and the other 12 kinds of residues are P residues. We have taken the average of all the interactions of H-H, H-P, and P-P in the M-J matrix, respectively, and found the average interactions between them as $-5.7,-3.4$, and $-1.7$ with standard deviations 0.8, 0.6, and 0.5, respectively. When the interaction of P-P is taken as a unit, then $E_{\text{H-H}}=-3.3$, $E_{\text{H-P}}=-2$, and $E_{\text{P-P}}=-1$. These values satisfy the relations above and agree with the work done by R. Mélin *et al.* [3], if the parameter $EC$ in their formula is taken as 1.

If the energies of all the conformations, including compact and incompact SAW conformations, are known, all the thermodynamic properties of a heteropolymer can be determined. The number of conformations associated with an energy level is called the degeneracy of the energy level. If the chain is short enough, the energy levels and their degeneracy can be obtained by enumeration method. Here we select a heteropolymer with length $N=16$, thus the most compact conformation would be a $4\times4$ structure. A similar model and enumeration method has been used by others [11,20,21]. Analysis of the composition of real proteins gives an average proportion of the H residues as 31.6% [22]. Here we focus on the chains with 5 H monomers, thus the proportion of the H monomers is now equal to $5/16=31.3\%$, similar to the real value, and the 5 H monomers can be used to construct the core of the most compact $4\times4$ conformations too. To make this heteropolymer ''proteinlike,'' its native conformation must be unique and the energy gap between the native conformation and the excited conformations must be large enough. Differing from the work done by Mélin *et al.* [3], the native conformation of sequence can be searched among all the conformations, not only in the most compact conformations. Many authors have selected $E_{\text{H-H}}=-1$, $E_{\text{H-P}}=E_{\text{P-P}}=0$ in their work [23]. If we make the same selection, among all the chains considered in this paper, there are only 19 chains with a nondegenerate native conformation, much less than the case (657, in the result part) of the present selection of energy type. And the energy spectrum is too simple, there are only five or six different energy levels and the minimal energy level is $-4$ or $-5$, the energy gaps between the native conformation and the low excited conformations are all 1. Therefore the difference between different chains is not easy to study. To a certain extent, complexity of energy type and length of chain may be complementary to separate different chains.

From the energy levels and their degeneracies, we can obtain the partition function $Z(T)$:

$$Z(T)=\sum_i n_i e^{-E_i/k_BT}, \tag{2}$$

where $n_i$ is the degeneracy of the ith energy level with energy $E_i$, $T$ is the temperature and $k_B$ is the Boltzmann constant. Here we set $k_B=1$. From the partition function we can find the average energy $\langle E(T)\rangle$ and the heat capacity $C_v(T)$ [24]:

$$\langle E(T)\rangle=\frac{\Sigma_i E_i n_i e^{-E_i/T}}{Z(T)}, \tag{3}$$

$$C_v(T)=\frac{\langle E^2(T)\rangle-\langle E(T)\rangle^2}{T^2}. \tag{4}$$

To study the folding process of a nascent polypeptide to its native structure, the Monte Carlo (MC) method has been widely used [3,26]. In this paper we use the standard Metropolis algorithm to simulate the folding process. In this algorithm, if the energy of the new conformation is higher than that of the old conformation, the probability to accept the new conformation is $e^{-\triangle E/T}$, where $\triangle E$ is the energy difference between the new conformation and the old conformation, and $T$ is the simulating temperature. The move set includes end flip, corner flip, and crankshaft of three bonds and global rigid rotation [23]. The site to perform a deformation is selected randomly. In the course of deformation, a lattice site cannot be occupied by more than one monomer. In the native-searching process, if the new conformation cannot be accepted on account of either the energy increasing or the overlapping of different monomers, a step must be registered in the folding time. We perform the simulation from a stretched conformation until the unique native conformation (known by the enumeration method) is reached. The folding time is defined as the MC steps used in the searching process. An MC step is defined as the $N$ move attempts in the simulation.

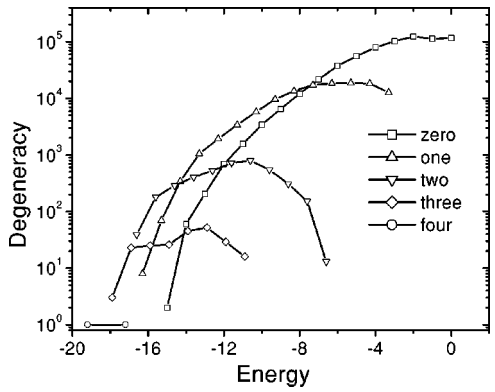


FIG. 1. The degeneracy of all the energy levels of a chain. The points can constitute several lines. The points on the different lines correspond to the conformations which have zero, one, two, or more H-H adjacent contacts.

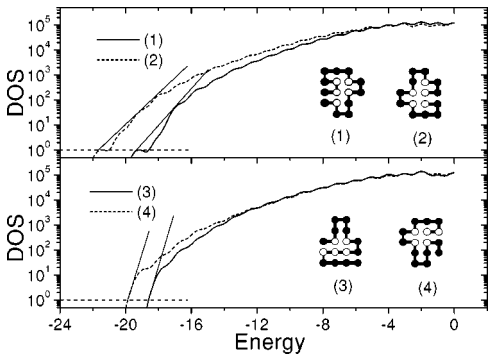


FIG. 2. The DOS curves for four different chains in the inner figures. The inner figures are the native conformations of the four chains; chains (1) and (2) are ''protein-like'' and chains (3) and (4) are ''nonproteinlike.'' The open circles are H monomers and the closed circles are P monomers. For the ''proteinlike'' chains, a line can be drawn passing through the native state (DOS=1) and tangent to the curve at the segment of the non-native states with energy much higher than the native conformation. For the ''nonproteinlike'' chains, the only tangent point is located at the native conformation.

## III. RESULT

### A. Phase transition and thermodynamic stability

For a chain, the energy levels and their degeneracies are obtained by the enumeration method. We enumerate all conformations of the chain with length $N=16$. The total number of conformations is 802 075. Considering the rotating and the mirror symmetry, the degeneracy of most of the conformations (802 074) is 8, except the degeneracy of the stretched conformation which is 4. Since the weight of the stretched conformation is very small, it can be neglected.

Because the interaction energy of the H-P contact is just twice the interaction energy of the P-P contact, the degeneracies of many energy levels is much larger. The degeneracy of all the energy levels for a chain as a function of the energy is shown in Fig. 1. From Fig. 1, we find that the points fall on several lines. Analysis of the values of the energy levels shows that the points on different lines represent the conformations with zero, one, two, or more H-H contacts.

From the energy spectrum and the degeneracy of all the energy levels, we can obtain all the thermodynamic properties. There are some criteria to distinguish the ''proteinlike'' polymers and the ''nonproteinlike'' polymers. To be ''proteinlike,'' first, the polymer must have one and only one conformation with minimal energy as the native conformation. Without reverse-labeling symmetry, the total number of chains with $N=16$ and $H=5$ is 2184, but there are only 657 chains with a nondegenerate native conformation, and only 284 chains' native conformations are compact $(4\times4)$.

We set the average difference between the lowest 10 excited conformations and the native conformation as a parameter of the different chains, which is called the average energy gap ($\overline{E_g}$). In enumerating the conformations, the values of $\overline{E_g}$ for all the 657 chains are calculated. The chains with large $\overline{E_g}$ are more likely the ''proteinlike'' heteropolymers while the chains with small $\overline{E_g}$ are more likely the ''nonproteinlike'' heteropolymers. To analyze the difference between these two kinds of chains, two chains with large $\overline{E_g}$ (1.97,1.8) and two chains with small $\overline{E_g}$ (both 0.3) are selected as representatives of these two kinds of chains (Fig. 2).

From the energy spectrums of all the 657 chains, we obtained average ground state energy as $-19$ and average number of energy levels of a chain as about 50. Thus the average separation of the energy levels is $19/50=0.38$. Continuous density of states (DOS) can be obtained approximately from the degeneracy of all the energy levels. The degeneracy of an energy level (points in Fig. 1) is replaced by a Gaussian function, as its peak value with the number of the conformations, the peak is at the position of the energy level and the standard deviation is 0.4 which is selected to make most part of the DOS curve smooth enough. The DOS curves (Fig. 2) for the chains are similar on the high-energy segment, but differ strongly on the low-energy segment. The logarithm of the DOS curve is the entropy curve $S(E)$. In the low excited energy part, there is an apparent concave segment on the entropy curves of the chains with large $\overline{E_g}$, but no concave segment for the chains with small $\overline{E_g}$. Thus on the $S(E)$ curve of the ''proteinlike'' chain we can draw a straight line passing though the native state and being tangent to the $S(E)$ curve at the part of the non-native states whose energy is much higher than the native conformation. At the temperature which is the reciprocal of the slope of the line, the free energy $F=E-TS$ of the native conformation, and that of the non-native conformations segment which contacts the line are the same, and the free energy of the low excited states is higher than them. With decreasing temperature, the weight of the native conformation increases. We call the temperature at which the weight of the native conformation begins to overtop that of the other conformations $T_{\mathrm{trans}}$. At this temperature, the free energy of the native conformation is equal to the free energy of the conformations with one of the excited energy levels. For the ''proteinlike'' chains, energy of the conformations whose weight is equal to that of the native conformation is much larger than the minimal energy, and there are low excited conformations (the concave part in Fig. 2) with weight less than that of the native conformation.

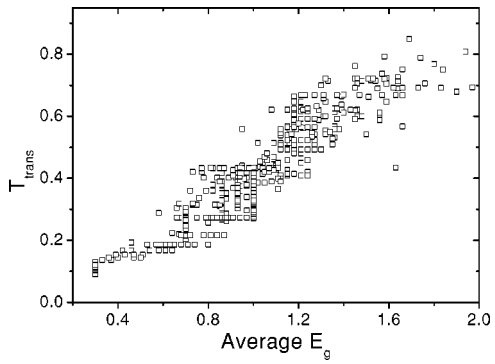


FIG. 3. $T_{\text{trans}}$ of all 657 chains with nondegenerate native conformation as a function of $\overline{E_g}$.

Therefore, the histogram of conformations as a function of energy at $T_{\text{trans}}$ will be bimodal and the phase transition from the molten globule to the native conformation is of first order. For the ''nonproteinlike'' chains, energy of the conformations whose weight is equal to that of the native conformation at $T_{\text{trans}}$ is just a little higher than the minimal energy. With decreasing temperature, the weight of the native conformation increases, but conformations with energy just a little higher than the native conformation will compete with the native conformation. The transition to the native conformation happens at much lower temperature and comes from the low excited conformations whose energy is close to the native conformation. Therefore, At $T_{\text{trans}}$ the histogram of conformations as a function of energy will not be bimodal, and the transition is not first order. In fact, as temperature decreases, the ''nonproteinlike'' polymers often cannot fold to their native conformations and they will come into glass state. For all the 657 chains, $T_{\text{trans}}$ as a function of $\overline{E_g}$ is shown in Fig. 3. The chains with large $\overline{E_g}$ tend to have a higher $T_{\text{trans}}$, i.e., ''proteinlike'' chains will transit to native conformations at a much higher temperature than the ''nonproteinlike'' chains.

$\langle E(T)\rangle$ and $C_v(T)$ are calculated by Eqs. (3) and (4), respectively, and are shown in Fig. 4 for the four chains. There are big differences, in the $C_v(T)$ curves, the peaks on the curves for the ''proteinlike'' chains reflect the phase transition from the molten globule to the native conformation

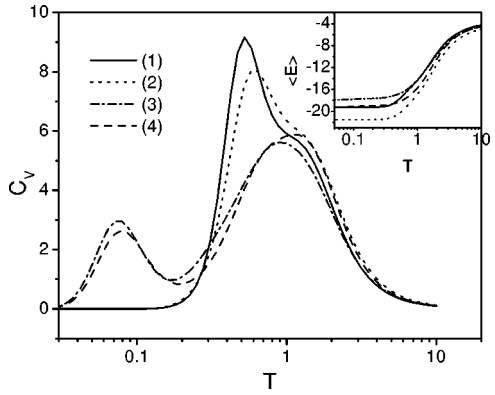


FIG. 4. $C_v$ of the four chains in Fig. 2 as a function of temperature. The inner figure is $\langle E\rangle$ as a function of temperature.

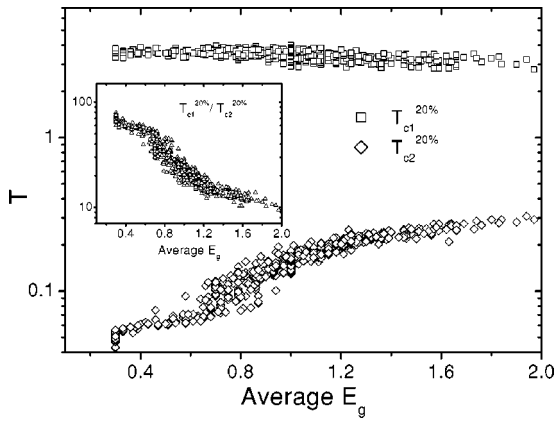


FIG. 5. $T_{c1}^{20\%}$ and $T_{c2}^{20\%}$ (defined in the text) as a function of $\overline{E_g}$. $T_{c1}^{20\%}/T_{c2}^{20\%}$ as a function of $\overline{E_g}$ is shown in the inner figure.

which happens at a higher temperature. They correspond to the small peaks on the $C_v(T)$ curves for the ''nonproteinlike'' chains at a much lower temperature. These small peaks reflect the transition from low excited conformations to native conformation for the ''nonproteinlike'' chains. There is a shoulder on each $C_v(T)$ curve for the ''proteinlike'' chains and a high peak on each $C_v(T)$ curve for the ''nonproteinlike'' chains. Their positions and heights are similar. All the shoulders and the peaks reflect the phase transition from random coil to molten globule. Because all the chains considered here have the same proportion of the H monomers, the incompact conformations with high energy for different chains have similar property. Therefore the phase transition from random coil to molten globule have similar property for all the chains. At the high- and low-temperature limits, $C_v(T)$ will be zero (Fig. 4). From the $C_v(T)$ curves of all the 657 chains, we calculate the temperatures at which $C_v$ increase to 20% of $C_v^{\max}$, the maximal value of $C_v$ for each chain, when decreasing temperature from high temperature and increasing temperature from low temperature. These two temperatures are called $T_{c1}^{20\%}$ and $T_{c2}^{20\%}$, which are shown as a function of $\overline{E_g}$ in Fig. 5. $T_{c1}^{20\%}$ is approximately constant for all the chains, but $T_{c2}^{20\%}$ increases nearly a order of magnitude with increasing $\overline{E_g}$. The transition width from coil to native conformation can be represented by $T_{c1}^{20\%}/T_{c2}^{20\%}$ (inner figure of Fig. 5). Transition width is much narrower for the chains with large $\overline{E_g}$.

The probability of the native conformation ($P_0$) as a function of the temperature for the four chains are drawn in the inner figure of Fig. 6(a). We see that the transition for the ''proteinlike'' chains is much sharper than that of the ''nonproteinlike'' chains. To measure the thermodynamic stability of the native conformation, let us define another transition width as

$$\delta = \frac{T_f^{10\%} - T_f^{90\%}}{T_f^{10\%} + T_f^{90\%}}. \tag{5}$$

Here, $T_f^x$ is the temperature at which the probability for the chain to be at its native conformation is $x$. The transition

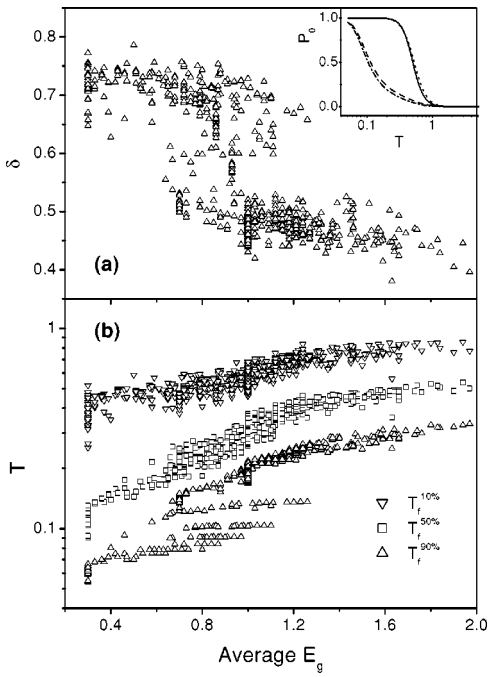


FIG. 6. (a) The transition width $\delta$ [Eq. (5)] of all 657 chains as a function of $\overline{E_g}$. For the chains with large $\overline{E_g}$, the transition width $\delta$ is smaller and their native conformations are more stable. The inner figure shows probability of the native conformation as a function of the temperature for the four chains in Fig. 4. The line types are the same as those in Fig. 2. (b) $T_f^{10\%}$, $T_f^{50\%}$, and $T_f^{90\%}$ (defined in the text) as a function of $\overline{E_g}$ for the 657 chains.

width of all the 657 chains with a unique native conformation are shown in Fig. 6(a). The transition of chains with large $\overline{E_g}$ trends to be sharper than chains with small $\overline{E_g}$, thus native conformations of chains with large $\overline{E_g}$ are thermodynamically more stable. Figure 6(b) shows $T_f^{10\%}$, $T_f^{50\%}$, and $T_f^{90\%}$ as a function of $\overline{E_g}$. $T_f^{10\%}$, $T_f^{50\%}$, and $T_f^{90\%}$ all increase with increasing $\overline{E_g}$, but $T_f^{90\%}$ increases faster than $T_f^{10\%}$.

From Figs. 3, 5, and 6, we can see that $\overline{E_g}$ is a good parameter to describe different chains, and the properties concerned here change almost monotonously and continuously with $\overline{E_g}$. But there is not a determinate criterion to distinguish the ''proteinlike'' and ''nonproteinlike'' heteropolymers. Therefore, it indicates that the transition of chains from ''nonproteinlike'' to ''proteinlike'' is continuous. What we can say is that chains with large $\overline{E_g}$ are ''proteinlike'' and they have a more stable native conformation than the other chains.

### B. Folding simulation

The folding time (measured by the MC steps before the protein reaches the native conformation for the first time, i.e., first passage time) depends highly on the simulating temperature [26]. We have selected one of the fast-folding chains to simulate the folding process from a stretched conformation to the native conformation. The folding time is averaged over 100 folding simulations, and average folding time at a wide range of temperature is obtained [Fig. 7(a)].

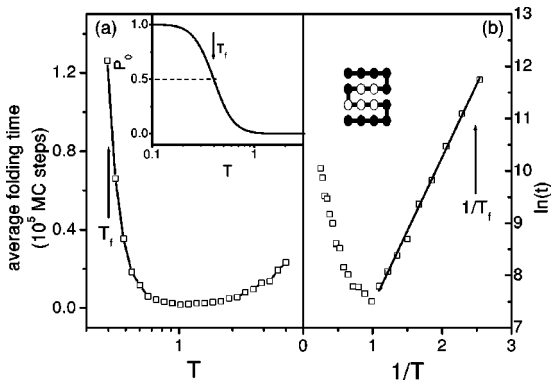


FIG. 7. (a) The average folding time $t$ of a fast-folding chain as a function of temperature. The inner figure is the curve of the probability of the native conformation as a function of temperature. (b) The logarithm of the average folding time of the same chain as a function of $1/T$. At the low-temperature segment, there are an apparent linear relationship between $\ln(t)$ and $1/T$. The slope of the line is the average energy barrier $E_b$. The inner figure is the native conformation of the chain.

Unlike the Go model [25], the temperature at which the protein folds most quickly to its native conformation is much higher than folding temperature $T_f$ (defined as the temperature at which the probability of the native conformation is 50%) [inner figure in Fig. 7(a)]. Figure 7(b) shows the logarithm of the folding time as a function of $1/T$. At low temperature, the logarithm of the folding time is a linear function of $1/T$, thus folding time $t \sim \exp(E_b/T)$. This suggests that the folding process at low temperature is an activated process of overcoming the energetic barriers. The average energetic barrier $E_b$ is the slope of the line.

In the body, the protein must be able to fold to its native conformation at the body temperature at which the protein's native conformation is thermodynamically stable. The body temperature must be of the same order of magnitude as $T_f$. In order to make the simulation fast enough, we select $T_f^{10\%}$ as the temperature at which the chains fold to their native conformation from the stretched conformation. We have averaged the folding time over 10 folding simulations for each chain. In order to save the time of computation, we have set a maximal MC steps for each simulation. If the chain could not fold to its native conformation until $N^5 = 104\,857\,6$ MC steps, we stopped the simulation and estimated a lower limit for the folding time. Figure 8 shows the folding time of all the 657 chains as a function of $\overline{E_g}$. Among them, 203 chains have not found their native conformations in all 10 simulations and only a lower limit of the folding time is obtained. From Fig. 8 we can see that chains with large $\overline{E_g}$ tend to fold faster. All chains with $\overline{E_g}$ greater than 1.5 will fold quickly to their native conformations, and the value of $\overline{E_g}$ of all the slow folding chains is less than 1.5. Most of the chains with very small $\overline{E_g}$ (less than 0.6) cannot fold quickly to their native conformations.

What factor affects the folding time of chains? Mélin *et al.* [3] concluded that some chains fold slow because of the proliferation effect, which is caused by a large number of low-energy conformations competing with native conforma-

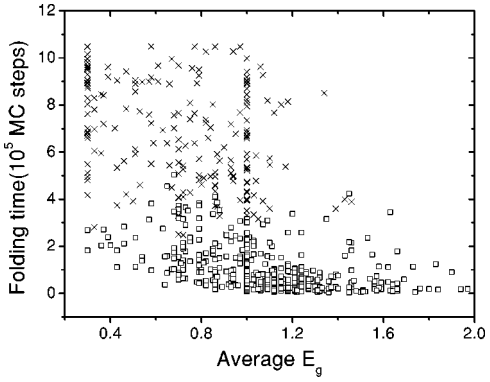


FIG. 8. The average folding time of the 657 chains as a function of $\overline{E_g}$. The open squares represent the chains which fold to their native conformations in all the ten simulations, and the cross symbols represent the chains for which not all the ten simulations find the native conformation and only a lower limit of the folding time is obtained.

tion in folding process. To test this effect, as the conformations being enumerated, we counted the number of conformations with energy less than the energy of the native conformation plus 2.0 for all the 657 chains. Figure 9 shows the folding time as a function of this number. Folding time tends to be longer for the chains with more low energy conformations. Chains with less than 17 low excited conformations can all fold fast, and slowly folding chains tend to have a large number of the conformations with energy close to the native conformation. This result confirms that the proliferation effect do exist.

Folding simulation is done at a temperature ($T_f^{10\%}$) lower than the fastest-folding temperature. Folding precess at this temperature includes two stage, a rapid collapse, and slow random searching for the native conformation among all the low-energy conformations [27,28], and the bottleneck is the second stage. Therefore, not only the number of low-energy conformations affect the folding time, but also the energy barriers which separate the local minimums with the native conformation. We make additional folding simulation of the four chains in Fig. 2 at different temperature. The logarithm of the folding time as a function of $1/T$ is shown in Fig. 10.

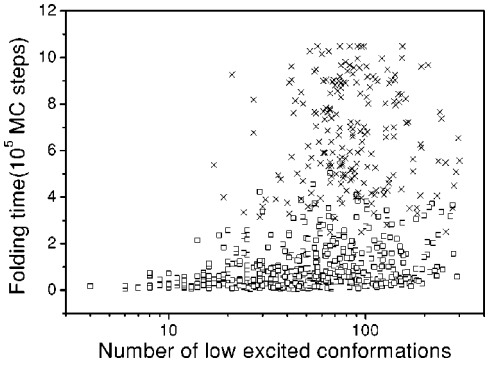


FIG. 9. The average folding time of the 657 chains as a function of the number of the conformations with energy less than the energy of the native conformation plus 2.0. The symbols are the same as those in Fig. 8.

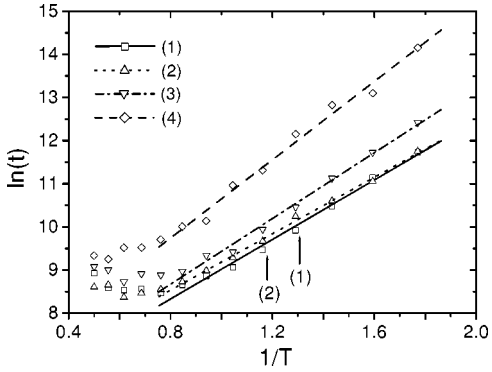


FIG. 10. The logarithm of the average folding time of the four chains in Fig. 2 as a function of $1/T$. From the slope of the fitting lines, average energy barrier $E_b$ can be obtained. The reciprocal of the folding temperature ($1/T_f^{10\%}$) for the two ''proteinlike'' chains is shown by the two arrows, and that for the ''nonproteinlike'' chains is out of the range of the figure.

For each chain and each temperature, folding time is the average of 100 simulations. At low temperature, the logarithm of the folding time is a linear function of $1/T$, i.e., $\ln(t) \sim E_b/T$. We make linear fitting of the eight $\ln(t)$ points at low temperatures for each chain to determine average energetic barrier $E_b$ from slope of the line (Table I). From Fig. 10, we can see that the fastest folding temperatures for all the four chains are almost the same. Because the folding simulation temperature $T_f^{10\%}$ of chains (3) and (4) is lower than that of chains (1) and (2), the effective energy barrier ($E_b/T_f^{10\%}$, Table I) is higher for chains (3) and (4). If the number of the low-energy conformations competing with native conformation is $n$, the folding time can be expressed as $t \sim n \exp(E_b/T_f^{10\%})$. The dependence of folding time on the number of low-energy conformations is linear, but the dependence of folding time on the effective energy barrier is exponential. Therefore, folding temperature is more important than the number of low-energy conformation to affect folding time. If we can make folding simulation at $T_f$ rather than $T_f^{10\%}$, the effect of temperature will be more dominant.

$E_b$ of chains (1) and (2) is a little smaller than that of chains (3) and (4) too (Table I), which can be understood because chains (1) and (2) have large $\overline{E_g}$. But $\overline{E_g}$ cannot determine $E_b$ because $\overline{E_g}$ concerns only the ten low excited conformations.

TABLE I. Some parameters for the four chains in Fig. 2. $n$ is the number of the conformations with energy less than the energy of the native conformation ($E_{\min}$) plus 2.0, and $t$ is folding time.

| Chain no. | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| $E_{\min}$ | $-19.2$ | $-21.5$ | $-17.9$ | $-19.2$ |
| $\overline{E_g}$ | 1.97 | 1.8 | 0.3 | 0.3 |
| $E_b$ | $3.45 \pm 0.13$ | $3.27 \pm 0.08$ | $3.81 \pm 0.14$ | $4.54 \pm 0.22$ |
| $T_f^{10\%}$ | 0.771 | 0.846 | 0.327 | 0.404 |
| $E_b/T_f^{10\%}$ | 4.47 | 3.87 | 11.65 | 11.24 |
| $n$ | 4 | 6 | 126 | 78 |
| $t$ | 18897 | 18956 | 803214[a] | 894440[a] |

[a]Lower limit of folding time.

Back to Fig. 8, similar to the result of Mélin *et al.* [3], we have found that many chains with $\overline{E_g}$ between 0.6 and 1.5 can also fold quickly to their native conformations, though folding simulation temperature $T_f^{10\%}$ for them will be lower than the chains with larger $\overline{E_g}$. This shows that though large $\overline{E_g}$ can make the native conformation more accessible, the topology of the native conformation also seems important for its accessibility.

## IV. CONCLUSION

Although the 2D H-P lattice model with only 16 beads in a chain used in this paper is simple, many properties of heteropolymer and protein can be found in it. Because proteins must have some necessary properties which have been discussed in the introduction, only a small fraction of the heteropolymers can act as proteins. In this paper we have calculated some properties of all the chains with $N=16$ and $H=5$. These properties change monotonically and continuously with $\overline{E_g}$, indicating a continuous transition from random polymers to proteins.

For a heteropolymer to be a protein, it must have a nondegenerate ground state as native conformation, and the average energy gap ($\overline{E_g}$) must be large enough, such that conformations with energy close to the native conformation are few and the native conformation is more likely to be in a deeper energy funnel. The large $\overline{E_g}$ not only guarantees the thermodynamic stability of native conformation, but also the fast-folding property.

Phases of the heteropolymer include random coil, molten globule, and native conformation. The main difference between ''proteinlike'' and ''nonproteinlike'' chains comes from their native conformation and the low excited conformations. For ''proteinlike'' and ''nonproteinlike'' chains, the phase transitions from random coil to molten globule are similar, but phase transitions from molten globule to native conformation are different. For ''proteinlike'' chains, it is first-order phase transition at a higher temperature, while for the ''nonproteinlike'' chains, it is not of first order and happens at a lower temperature.

Folding simulation results confirm the proliferation effect caused by other low-energy conformations [3]. But folding simulation at different temperature shows that the main reason is that the folding simulation temperature $T_f^{10\%}$ of the chains with small $\overline{E_g}$ is lower than that of the chains with large $\overline{E_g}$, which make the effective energy barrier $E_b/T_f^{10\%}$ higher. Thus it can be concluded that it is by means of the folding temperature $T_f$ that fast folding property correlates with thermodynamic stability.

Why some chains fold fast to their native conformation while some chains fold slowly is not a simple problem. The folding time is determined by the energy landscape which is difficult to describe. Another factor which can affect the folding time is the topology of native conformation, which is not a concern in the current work. The accessibility difference of conformations caused by their topology should be able to be described by a parameter which can be abstracted from the vicinity of the native conformation in the energy landscape or from the native conformation itself. This problem needs further study.

## ACKNOWLEDGMENTS

We are grateful to H. J. Zhou and Y. Zhang for discussions and critical comments. This work was supported by CERNET CHPCC.

[1] *Protein Folding*, edited by T. E. Creighton (Freeman, New York, 1992).
[2] C. Anfinsen, Science **181**, 223 (1973).
[3] R. Mélin, H. Li, N.S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).
[4] C. Tang, e-print cond-mat/9912450.
[5] E.I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[6] E.I. Shakhnovich and A.M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).
[7] M. Kardar, Science **273**, 610 (1996).
[8] Y. Zhou, C.K. Hall, and M. Karplus, Phys. Rev. Lett. **77**, 2822 (1996).
[9] Y. Zhou and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **94**, 14 429 (1997).
[10] M.H. Hao and H.A. Scheraga, Physica A **244**, 124 (1997).
[11] A. Dinner, A. Šali, M. Karplus, and E. Shakhnovich, J. Chem. Phys. **101**, 1444 (1994).
[12] K.A. Dill, Biochemistry **24**, 1501 (1985).
[13] K.F. Lau and K.A. Dill, Macromolecules **22**, 3986 (1989).
[14] E.I. Shakhnovich and A.M. Gutin, Nature (London) **346**, 773 (1990).
[15] S. Miyazawa and R.L. Jernigan, Macromolecules **18**, 534 (1985).
[16] S. Miyazawa and R.L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[17] H. Li, C. Tang, and N.S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[18] J. Wang and W. Wang, Nature Struct. Biolo. **6**, 1033 (1999).
[19] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[20] V. Shahrezaei, N. Hamedani, and M.R. Ejtehadi, Phys. Rev. E **60**, 4629 (1999).
[21] H.S. Chan and K.A. Dill, Macromolecules **22**, 4559 (1989).
[22] L. F. Yan and Z. R. Sun, *Molecular Structure of Protein* (Tsinghua University Press, People's Republic of China, Tsinghua 1999).
[23] H.S. Chan and K.A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[24] H. L. Friedman, *A Course in Statistical Mechanics* (Prentice-Hall, Englewood Cliffs, NJ, 1985).
[25] M. Cieplak, T.X. Hoang, and M.S. Li, Phys. Rev. Lett. **83**, 1684 (1999).
[26] A.M. Gutin, V.I. Abkevich, and E.I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).
[27] E. Shakhnovich, G. Farztdinov, A.M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[28] A. Šali, E. Shakhnovich, and M. Karplus, Nature (London) **369**, 248 (1994).