Thermodynamic stability and kinetic foldability of a lattice protein model

Jie Li, Jun Wang, Jian Zhang, and Wei Wang[a)]
National Lab of Solid State Microstructure and Physics Department, Nanjing University, Nanjing 210093, China

(Received 3 September 2003; accepted 6 January 2004)

By using serial mutations, i.e., a residue replaced by 19 kinds of naturally occurring residues, the stability of native conformation and folding behavior of mutated sequences are studied. The $3\times3\times3$ lattice protein model with two kinds of interaction potentials between the residues, namely the original Miyazawa and Jernigan (MJ) potentials and the modified MJ potentials (MMJ), is used. Effects of various sites in the mutated sequences on the stability and foldability are characterized through the Z-score and the folding time. It is found that the sites can be divided into three types, namely the hydrophobic-type (*H*-type), the hydrophilic-type (*P*-type) and the neutral-type (*N*-type). These three types of sites relate to the hydrophobic core, the hydrophilic surface and the parts between them. The stability of the native conformation for the serial mutated sequences increases (or decreases) as the increasing in the hydrophobicity of the mutated residues for the *H*-type sites (or the *P*-type sites), while varies randomly for the *N*-type sites. However, the foldability of the mutated sequences is not always consistent with the thermodynamic stability, and their relationship depends on the site types. Since the hydrophobic tendency of the MJ potentials is strong, the ratio between the number of the *H*-type sites and the number of the *P*-type sites is found to be 1:2. Differently, for the MJJ potentials it is found that such a ratio is about 1:1 which is relevant to that of real proteins. This suggests that the modification of the MJ potentials is rational in the aspect of thermodynamic stability. The folding of model proteins with the MMJ potentials is fast. However, the relationship between the foldability and the thermodynamic stability of the mutated sequences is complex.
 [DOI: 10.1063/1.1651053]

I. INTRODUCTION

Proteins are elementary blocks to execute biological functions in the living organisms. There are many kinds of proteins in nature which carry out various complicated activities. Proteins are composed of 20 kinds of naturally occurring amino acids, and are encoded by complex patterns of these 20 kinds of amino acids. That is, 20 kinds of amino acids introduce diversity and complexity into proteins due to various specific propensities.[1–8] In general, a protein contains about 10 000 atoms and also interacts with a huge number of solvent molecules. Even using the fastest computers, it is still quite difficult to simulate the folding process for an amino acid sequence with a reasonable size when all the interactions between the atoms are included.[9] Presently, protein folding is a fundamental and unsolved problem in molecular biology. The folding resembles a diffusion process on a rugged funnel-like energy landscape.[10–13]

It is well known that some kinds of amino acids or residues have similar physicochemical properties while some other kinds of residues are different. Thus, it is possible that some residues in a protein can be substituted by some kinds of similar residues to simplify the complexity in protein. Nevertheless, some residues can be substituted by quite different kinds of residues to check or to study the importance and the role of these residues and their corresponding positions. Such a way of substitution of residues is termed as mutation.[14–23] Through the mutation method, there are many experimental and theoretical studies for various features of proteins, such as for the stability, the folding rate, the biological functions, and the complexity simplification by grouping the residues.[24–26] In experiments, mutations for proteins are often done by replacing the residues of a protein sequence with other kinds of residues via gene engineering method,[14–17,27,28] while in theoretical studies, the mutations are usually realized by changing the interactions between the replaced residue and its related residues, sometime by cutting or adding directly the native contacts.[18,19] According to the number of mutated residues, the mutation can be divided into single-site mutations and double-site mutations, as well as multisite mutations.[20,21]

Experimental and theoretical studies showed that there are some ''hot,'' and ''cold,'' as well as ''warm'' sites in a protein according to their contribution to the thermodynamic stability of the native structure.[25,26] For the ''hot'' sites, the mutations are likely to cause the protein to change largely in their folding properties and for the ''cold'' sites, the mutations have no relevant effects on thermodynamics, while for the ''warm'' sites, the effect of the mutations is in the intermediate situation between those of the ''hot'' sites and those of the ''cold'' ones.[25,26] Further studies also indicated that the distribution of such three kinds of sites relates to the interaction between the residues.[26] The ''hot'' sites and ''cold'' sites do not straightly correspond to the hydrophobic core sites or the hydrophilic surface sites in real proteins because

a)Electronic mail: wangwei@nju.edu.cn

it is found that the ''hot'' sites can be found on the surface as well.[26]

Different sites play different roles in folding kinetics as well as in thermodynamic stability. It is well known that the rate-limiting transition state is critical to protein folding. To study the property of the transition state, the so-called ϕ value is introduced. The definition of ϕ value is mutation-based, and it is regarded as the ratio (upon mutation) of the difference in the free energies between the transition state and the unfolded state with respect to the difference in the free energies between the folded state and the unfolded state.[27,28] The normal ϕ value of the mutated site is between 0 to 1. The larger the ϕ value of the mutated site, the more important the mutated site for the folding kinetics is.[29–35] It is found that the application of ϕ value is not suitable for the systems with large frustration.[33,36–40] The abnormal ϕ values out of the range between 0 to 1 shown either in experiments or simulations have also been argued to be useful on studying the speciality of sites during protein folding.[35] It has been suggested that corresponding to experiments, ϕ value is a good parameter to characterize the transition states and the folding nucleus.[40] The study of protein evolution indicates the conservation or the importance of the folding nucleus during the evolution.[41–44] These all are realized based on the techniques of mutations at different sites in proteins.

Many questions come out. How do the effects of various sites on the thermodynamic and kinetic relate to their spatial arrangements? How do different sites affect the folder? What are the effects of the hydrophobic or the hydrophilic nature of residues even at the same site of a protein? These questions have been attracting the researchers for many years, and much work has been done recently.[25–35,40–44]

In this work, we study the relationship between the thermodynamic and kinetic properties of proteins based on various mutations at different sites in model proteins. The model proteins consist of 27 monomers onto a cubic lattice[45–47] with the MJ potentials[48] and the modified MJ potentials (MMJ) raised by Thirumalai *et al.*[49] Considering the ''continuity principle'' of the mutations,[20] single-site mutation method is used. A serial of mutations which includes a group of mutations for one site from the original residue to the other 19 kinds of residues is made. By observing the correlation between the thermodynamic stability and the hydrophobicity of mutated residues in the mutation serial, we find that different sites contribute to the stability of the native structure differently. Accordingly the sites are classified into three types, which are the hydrophobic sites (*H*-type sites), the hydrophilic sites (*P*-type sites), and the neutral sites (*N*-type sites) based on their preferences. For the case of the MJ potentials, the stability of the mutated sequence is increasing (or decreasing) as the increase in the hydrophobicity of the mutated residues for the *H*-type sites (or *P*-type sites), while is random for the *N*-type sites. These three types of sites are shown to be related to the hydrophobic core, the hydrophilic surface and the connected sites between the core and the surface in the model proteins. It seems that the *H*-type and *P*-type sites are generally more thermodynamic important than the *N*-type sites. Further observation of three types of sites in determining the native structure, the serial mutations introduce an ''edge effect,'' indicating the mutations on the residues with weak energetics would change the native structure unexpectedly. Contrasted to the thermodynamic study, the kinetic properties of the mutated sequences are also discussed. Note that the hydrophobic tendency of the MJ potentials[48] is strong. As a comparison, we further make a study using the modified MJ potentials (MMJ).[49] Our results show that the classification of three types of sites is independent of the potentials. It is found that the ratio between the *H*-type sites and the *P*-type sites is about 1:1, which is more reasonable than that of the MJ potentials, i.e., 1:2. This suggests that the modification of the MJ potentials is rational.

The organization of this paper is as follows. In Sec. II and Sec. III, we present the model and the methods used in this work. In Sec. IV, we present the results and discussions for the case of the MJ potentials first, and then for the case of the MMJ potentials. In the final section, we give a conclusion.

II. MODEL

The protein is modelled as a self-avoiding chain on a cubic lattice [see Fig. 1(a)]. The energy of a conformation for the chain is the sum of energies of nonsequential neighboring pairwise contacts with a lattice unit *a*. The nonsequential neighboring interaction energy depends on the identities of the two residues (or monomers) involved. Thus, the energy of a conformation can be represented as

$$E=\sum_{i,j,i<j}^{N} B(\xi_i,\xi_j)\Delta_{ij}, \tag{1}$$

where $\Delta=1$ if monomer i and monomer j are two nearest nonsequential neighbors and $\Delta=0$ otherwise, and ξ_i and ξ_j define the kinds of residues at positions i and j, respectively. $B(\xi_i,\xi_j)$ is the interaction energy between two residues of kinds ξ_i and ξ_j. Here we use two kinds of potentials, i.e., the MJ and MMJ potentials which are the Table 3 in Ref. 48 by Miyazawa and Jernigan and the Table 2 in Ref. 49 by Thirumalai *et al.* for $B(\xi_i,\xi_j)$. N is the number of the total residues or the length of the model chain, and it is set as 27 in this work.

It is noted that such a lattice model is a classical example for the model proteins, and has been used intensively for the studies on some basic and general features of the proteins and protein folding.[45–47] It is found that such a model protein shows some basic features in common with respect to real proteins. Therefore, the lattice model is used to address questions of some general aspects of proteins rather than of atomic details. In the aspect of thermodynamics, a model protein chain has a unique compact native structure if its amino acid sequence is well designed. At the same time, in the aspect of kinetics, the model protein behaves as a two-state folding and a collapse of folding nucleus.[45–47] The cubic native conformation shows surface sites and internal sites, which correspond to the corner sites and the center sites (including both face centers and body center) in the compact three-dimensional structures, respectively.

III. METHOD

A. Z-score for sequence design

As we know, the Z-score is a good parameter to measure the stability of the native structure for a protein sequence and to characterize the energy gap between the energy of the native conformation and an averaged energy over those of the misfolded or unfolded conformations.[50] The Z-score is defined as

$$Z=\frac{|E_{\mathrm{nat}}-E_{\mathrm{av}}|}{\sigma}, \tag{2}$$

where E_{nat} is the energy of the native conformation, and E_{av} is the averaged energy of the non-native compact conformations. σ is the corresponding dispersion of energies of the non-native compact conformations. Unlike the estimation of the averaged energy E_{av} in previous work,[50] here we calculate E_{av} by averaging the energies of the protein sequence over all the compact structures of the $3\times3\times3$ cubic lattice except the native one, i.e., $103\,346-1$ compact structures.

Obviously, for a certain sequence, the larger the energy gap $|E_{\mathrm{nat}}-E_{\mathrm{av}}|$, the more stable the native conformation is, and the smaller the value of σ, the larger the Z-score is. Thus, a large value of Z-score means that the native conformation is stable, and the sequence may have good folding features, such as the thermodynamic stability and the kinetic accessibility.[50] Actually, the Z-score is often used to design the foldable model protein sequences. With different arrangements of residues in a sequence, the values of Z-score are different. For a fixed composition of various kinds of residues, a sequence with the largest value of Z-score is argued to be the best sequence for the target structure. It is shown that the Z-score is a very useful factor for the study on some properties of proteins.

B. Serial point mutations

In this work we use a serial of point mutations for protein sequences, i.e., we make 19 mutations at a certain site by replacing the original or the ''wild type'' residue with other 19 kinds of naturally occurring residues. This kind of mutations is called as the serial mutations. By such serial mutations at one site for a protein sequence, we obtain 19 mutated sequences. To study the effects of mutations on the stability of the native structure for the 19 mutated sequences, a serial values of Z-score for the related mutated sequences are worked out. These values, called as the serial map of Z-score, are used to characterize the thermodynamic stability of various mutated sequences. By making the mutations at different sites in a sequence, the contribution of each site to the stability of the native structure can be described by the values of Z-score. For comparison, the value of Z-score of the ''wild type'' sequence is also added to the serial map of Z-score.

C. Monte Carlo simulations

In order to study the kinetics of the folding for the mutated chains, we make many Monte Carlo simulations. Similar to the previous work with lattice model, during the folding process, there are three types of moves used in the simulations, namely the end move, the corner move, and the crankshaft move.[45] Starting from an extended conformation, the folding is inspected both by the Monte Carlo steps (MCs) or the Q value. For obtaining the mean first passage time (MFPT) for each ''wild'' sequence, 200 runs are made from different initial states for the chains. To characterize the accessibility of the native conformation for different mutated sequences, a serial map of folding ratios, P_f, is also used. Here the folding ratio P_f is defined as the ratio of the number of runs N_1 for which the sequences fold to their native state to the total number of runs N_0 within a threshold time, i.e.,

TABLE I. The residue composition, the Z-score and the MFPT for 16 well-designed sequences by the Z-score method for the lattice model protein with the native structure shown in Fig. 1(a). The MJ interaction potentials are used. The sequence indexes follow their related values of Z-score.

	Sequence	Z-score	MFPT ($\times10^7$ MCs)
1	ITGARPSPNHDYGCKQWEVGMRGTYLF	6.12	3.76
2	ITGARPSPNHEYGCKQWDVGMRGTYLF	6.11	4.56
3	LQKCGVSAGPDPRHTGIGMEYRNTYFW	6.06	5.65
4	IERYGASGNHTPRYGQVGMDWPKTCLF	6.00	6.62
5	IGNTKASPQPGHRYRYVEMDCGGTWLF	5.96	3.91
6	IGRADCGPNHSPKYQTVRMEWTGGYLF	5.86	7.76
7	IRKAGTSPQPNHRCGYMGVDYGETWLF	5.73	4.99
8	WTGPKPEAGCQYGNRHVDYGMTSRILF	5.72	3.35
9	ITKAEPNPQHGGRYGYWRMDVGSTCLF	5.69	6.79
10	GTRAGPNPQHDMKCGYWGVEIRSTYLF	5.63	2.10
11	LGNAGCSHKPTRRTGEMGVDYPQYFIW	5.30	10.69
12	MDYQRTAPGTPGYSRNCEMKIGHGWVL	5.14	16.63
13	LWSPKCNGHPAQGGRDVEYGMTRTYIF	5.01	30.48
14	FTPHRNGPEGRCGAKSMTVGWDQYYLI	4.97	12.61
15	YGNPSAGYTWKVQCDTMRGPFEGHRLI	4.85	20.70
16	YGNGSRTHDPKTQCPAMGIYVEGRWLF	4.71	7.58

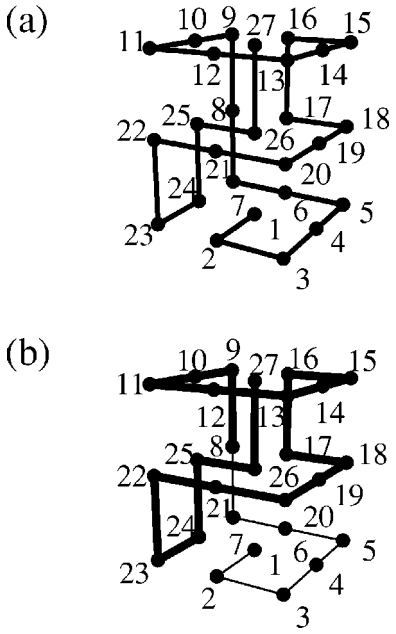

FIG. 1. The lattice model protein. (a) The native conformation (or the target conformation for design). (b) The folding nuclei for the first 10 sequences in Table I. They are plotted overlapped together. The sites of the folding nuclei are connected with thick bonds.

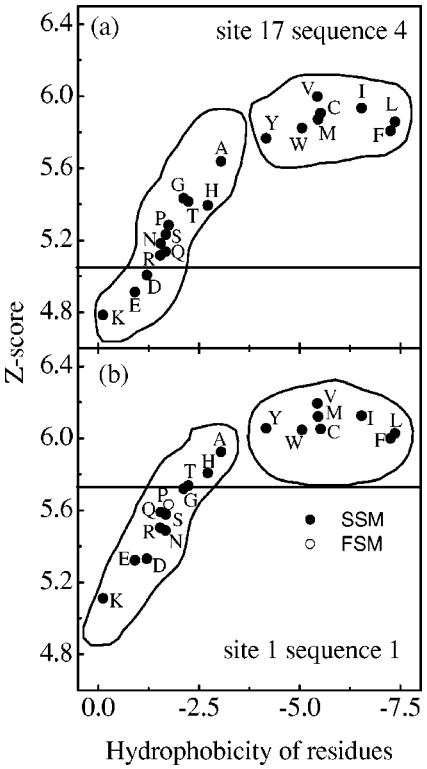

FIG. 2. Serial map of Z-score vs the strength of hydrophobicity of residues for *H*-type site. (a) Plot for site 17 in sequence 4 of Table I. The correlation coefficient is 0.9592. The serial mutations at site 17 in sequence 4 are the successful separating mutations. (b) Plot for site 1 in sequence 1 of Table I. The correlation coefficient is 0.9367. In both plots, the circles enclose the hydrophobic and hydrophilic residues, respectively. The straight line separates the conserved mutations for the native conformation from the nonconserved mutations. The solid circles represent the mutations keeping the native conformation or the conserved mutations, and open circles represent the mutations not keeping the native conformation or nonconserved mutations [for residues *P* in (b)].

$P_f = N_1/N_0$. Here, for the lattice model with two kinds of potentials, we generally choose $N_0 = 200$. For each mutated sequence, the threshold time or the total number of MCs of the simulation are taken as $0.9\times$MFPT with MFPT being the folding time of the ''wild'' sequence. Considering that the single site mutation will not change the folding temperature significantly, a temperature $T = T_f$ is set in all simulations for different mutated sequences with T_f being the folding temperature of the ''wild'' sequence. Actually, considering the coarse-grained description based on the lattice model of proteins, the folding ratio P_f describes well the accessibility or the foldability of a model protein to its native conformation at a biologically relevant time scale.

D. Folding nuclei

Here, we determine the folding nuclei of the sequences by a simple way. Taking the number of the native contacts (Q value) as the coordinate of the folding process, we account the occurring frequency of each native contact when the Q value is between 16 and 22.[51] Then the folding nucleus is defined as the set of the native contacts with the highest frequencies. The folding nuclei for the first 10 sequences in Table I are plotted overlapped together in Fig. 1(b).

IV. RESULT AND DISCUSSION

A. The thermodynamic stability

To study the role of the mutations at different sites in various sequences, using the MJ potentials 16 sequences have been designed[50] as the ''wild'' type sequences for a native target as shown in Fig. 1(a). These 16 sequences are selected as a pool for the serial mutations as listed in Table I. Their thermodynamic and kinetic parameters, such as the values of Z-score and the folding time MFPT are also listed. The sequence index is arranged following the magnitude order of their related values of Z-score. These sequences are selected with the values of Z-score about 4–6, and their MFPT from 10^8 to 10^7 MCs. The first 10 sequences can be regarded as good sequences with high values of Z-score and fast folding time around 10^7 MCs, while the last six sequences fold to the native conformation comparatively slowly with about 10^8 MCs, as well as with low values of Z-score.

Then the serial mutations are carried out for all the sites of these 16 sequences. In order to track the variation of the stability of the native structure during the serial mutations, the serial map of Z-score at each mutated site is plotted (see Figs. 2–4). In the serial map of Z-score, the mutated residues are classified into the hydrophobic and the hydrophilic division by two circles as shown in these figures. All the naturally occurring 20 kinds of residues are arranged following their strengths of the hydrophobicity. Here the strengths of the hydrophobicity of various residues are taken as their values of the components of the principal eigenvector for the MJ matrix.[22] This implies that the more negative value of the hydrophobicity of the residue, the larger hydrophobicity of the residue is. It is noted that we have basically similar fea-

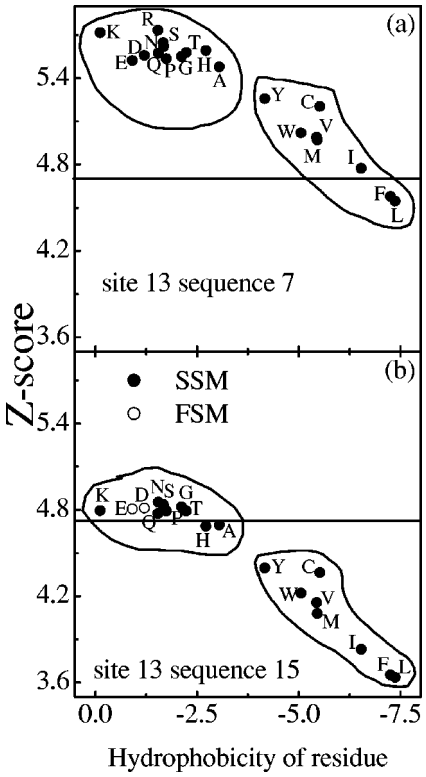

FIG. 3. Serial map of Z-score vs the strength of hydrophobicity of residues for *P*-type site. (a) Plot for site 13 in sequence 7 of Table I. The correlation coefficient is −0.8821. The serial mutations are successful separating mutations. (b) Plot for site 13 in sequence 15 of Table I. The correlation coefficient is −0.8787. The serial mutations are failure separating mutations. The open circles (for residues *E* and *D*) are close to the boundary line. The boundary lines are the same as for Fig. 2.

tures in the serial maps of the Z-score if the values of the hydrophobicities of the natural residues taken from the textbook are used. In the serial maps of Z-score, the value of Z-score is used to characterize the thermodynamic stability of the native structure. By considering the correlation between the Z-score and the hydrophobicity, it is found that the protein sites tend to choose certain kinds of residues under the stable pressure.

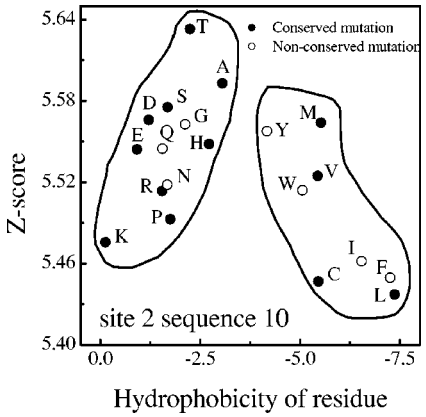

FIG. 4. The serial map of the Z-score for *N*-type site 2 in sequence 10 of Table I. The open circles present the nonconserved mutations for the native conformation. The distribution of the points is dispersive, and no boundary line can be defined.

Collecting the results from all 27×16 serial maps of Z-score of 27 sites for 16 sequences, it is found that there are three different features for these maps. According to these features, the sites in a protein sequence can be classified into three types: (1) the hydrophobic sites (the *H*-type sites) in which the values of Z-score increase as the increase in the hydrophobicity of the mutated residues as shown in Fig. 2(a) and Fig. 2(b); (2) the hydrophilic sites (the *P*-type sites), in which the values of Z-score decrease following the increase in the hydrophobicity of the mutated residues as shown in Fig. 3(a) and Fig. 3(b); (3) the neutral sites (the *N*-type sites), in which the relationship between the values of Z-score and the strengths of the hydrophobicity of the mutated residues is basically random as shown in Fig. 4. These results are listed in Table II, where the letter *H* represents the *H*-type sites, the letter *P* the *P*-type sites, and the letter *N* the *N*-type sites, respectively. From Fig. 2 to Fig. 3, strong correlations between the values of Z-score and the strengths of the hydrophobicity of the mutated residues are observed, indicating that there is an obvious tendency of these sites to some kinds of residues due to stability requirement. For the *H*-type sites, the more hydrophobic they are, the more stable the native structure is. Thus, the *H*-type sites tend to accept the hydrophobic residues rather than the hydrophilic residues. While for the *P*-type sites, the case is reversed. The hydrophilic residues are more suitable than the hydrophobic residues in the *P*-type sites. Note that for both types of sites, the residues are clearly divided into two groups and the values of Z-score show a roughly linear relationship to the hydrophobicity (see Fig. 2 and Fig. 3). However, for the *N*-type sites, the dispersive distribution as shown in Fig. 4 implies that there is an obscure tendency of the *N*-type sites to both the hydrophobic and the hydrophilic residues.

These three types of sites for mutations reflect different roles of various sites in a protein to the stability of the native structure. As shown, the contribution of the sites of *H*-type and *P*-type to the stability of the native structure under the same condition is reversed. Increasing the hydrophobicity of the residues at the *H*-type sites will result in the increase in the stability of the native structure (see Fig. 2), while at the *P*-type sites will result in the decrease in the stability of the native structure (see Fig. 3). Differently, for the *N*-type sites the contribution of the mutations to the stability of the native structure by both the hydrophobic and hydrophilic residues is not relevant to the hydrophobicity (see Fig. 4).

It is well known that the main driving force in the protein folding is the hydrophobic force which is an important effect of solvent. Therefore, in general, real proteins are constructed by three parts. One is the hydrophobic core, one is the hydrophilic surface, and the other is the neutral part between the core and surface. In the hydrophobic core, the hydrophobic residues are energetically favorable, in the hydrophilic surface, the hydrophilic residues are energetically favorable, while in the neutral part the favorable residues are uncertain. What is the relationship of the mentioned above site types to the hydrophobic core? In Fig. 5, a statistical histogram of occupying probabilities for three types of sites at four kinds of structural positions of the cubic native structure is shown. It is clearly seen that the *H*-type sites are

TABLE II. The distribution of the *H*-type, *P*-type, and *N*-type sites at 27 positions of 16 sequences in Table I.

	Sequence															
Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*
2	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*N*	*P*	*P*	*P*	*P*	*N*	*N*
3	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
4	*N*	*N*	*H*	*N*	*H*	*N*	*H*	*N*	*H*	*N*	*N*	*P*	*N*	*N*	*N*	*N*
5	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
6	*N*	*N*	*N*	*N*	*H*	*N*	*N*	*H*	*N*	*N*	*N*	*N*	*H*	*N*	*N*	*N*
7	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
8	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*H*	*N*	*P*	*N*	*N*	*N*
9	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
10	*N*	*N*	*N*	*N*	*N*	*N*	*H*	*N*	*N*	*N*	*N*	*N*	*P*	*N*	*N*	*H*
11	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
12	*N*	*N*	*N*	*N*	*N*	*N*	*P*	*N*	*N*	*H*	*P*	*P*	*N*	*N*	*H*	*H*
13	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
14	*H*	*H*	*N*	*N*	*N*	*H*	*N*	*P*	*N*	*H*	*N*	*P*	*N*	*H*	*N*	*N*
15	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*N*	*P*	*P*	*P*	*P*
16	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
17	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*
18	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
19	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*
20	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
21	*H*	*H*	*H*	*H*	*H*	*H*	*P*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*
22	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
23	*P*	*P*	*P*	*P*	*N*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
24	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*	*P*
25	*N*	*N*	*N*	*H*	*H*	*P*	*H*	*H*	*H*	*N*	*N*	*H*	*H*	*N*	*N*	*N*
26	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*
27	*H*	*H*	*N*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*N*	*H*	*H*	*H*	*H*	*H*

mainly at the positions of the body center and the face centers. Differently, the *P*-type sites are mainly distributed at the positions of the corners and the edge centers. It is also noted that most of the *N*-type sites locate at the edge centers and some at the face centers.

In the 27 lattice model, the body and face centers have been argued to be the internal positions of the cubic structure, and these positions clearly relate to the hydrophobic core. While the corners have been considered as the surface positions of the cubic lattice structure of the model proteins. The high occupying probabilities of the *H*-type sites at positions of body and face centers, and of the *P*-type sites at corners are relevant to the correspondence of the *H*-type, *P*-type, and *N*-type sites to real proteins mentioned above, i.e., the *H*-type sites relate to the hydrophobic core sites, while the *P*-type sites relate to the hydrophilic surface sites. The edge centers are the intermediate positions connecting the face centers and the corners, thus it is difficult to tell whether they belong to the surface positions or the internal positions in the cubic lattice structure. As a result, from Fig. 5 it is observed that for the edge centers, the occupying probabilities of both the *P*-type and the *N*-type sites are more or less the same. These also provide a support on the validity of the 27 lattice model in describing proteins.

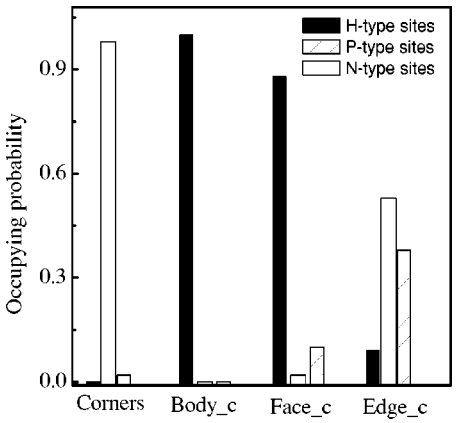

FIG. 5. The histogram of the occupying probability of the *H*-type, the *P*-type, and the *N*-type sites at four kinds of positions of the cubic lattice model with the MJ potentials. These positions are the corners, body center, face centers, and edge centers.

Table III lists the ratio between the numbers of the *H*-type and the *P*-type sites for 16 sequences listed in Table I. It is found that the numbers of the *P*-type sites are about 2 times the numbers of *H*-type sites for most of the sequences. This is not consistent with ratio between the sites of *H*-type and *P*-type in real proteins known as 1:1. This unreasonable ratio may be due to the strong tendency of the hydrophobic-

TABLE III. The ratio of the number of *H*-type sites and the number of *P*-type sites for 16 sequences in Table II.

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

TABLE IV. The statistics on the features of mutations. The total number (TM) of the *H*-type sites, the *P*-type sites, and the *N*-type sites. The number of the full serial mutations with conserved native conformation for three types of sites (FCM). The number of the successful separating mutations for three types of sites (SSM). The number of the failure separating mutations for three types of sites (FSM). R_c presents the ratio of FCM to TM. R_s and R_f show the ratios of SSM and FSM to TM, respectively.

	TM	FCM	SSM	FSM	R_c	R_s	R_f
H	128	70	47	11	54.6%	36.7%	8.7%
P	217	121	79	17	55.8%	36.4%	7.8%
N	87	78	1	8	89.6%	1.1%	9.3%

ity in the MJ potentials (see following results for the MMJ potentials).

In many previous works, the studies on the effect of mutations focus on whether the site mutation results in a change in the native conformation of a sequence. Here, we also discuss such effect for the serial mutations at the *H*-type, the *P*-type, and the *N*-type sites in sequences listed in Table I, respectively. It is noted that our discussion is from the aspect of the thermodynamic stability of the native structure. We use all the $3\times3\times3$ cubic compact conformations, i.e., 103 346 conformations, to judge whether the original native conformation [shown in Fig. 1(a)] is still the lowest state of every mutated sequence. Thus, the conservation of the native conformation for a mutated sequence relates to the fact that the mutated sequence still takes the original conformation as its native conformation. It is known that the hot site is of sensitive site at which mutations often result in changes of the native conformation, while the cold site is of insensitive site at which mutations often result in the conservation of the native conformation.[25,26] Then, is there any correspondence from the hot or the cold site to the *H*-type, the *P*-type or the *N*-type site? To answer this question, we calculate the ratio of the conserved mutated sequences to the total number of the mutated sequences. This ratio, defined as R_c, for all the 27×16 serial mutations is listed in Table IV.

From Table IV, we can see that the values of R_c for both the *H*-type and the *P*-type sites are more or less the same, i.e., $R_c\simeq55\%$, but are smaller than $R_c\simeq89\%$ for the *N*-type sites obviously. This indicates that both the *H*-type and the *P*-type sites are more sensitive to the mutations than the *N*-type sites. This result is reasonable because the sites in the hydrophobic core or those on the hydrophilic surface apparently contribute more to the stability of the proteins than the sites between them.

Considering the distribution of the conserved and nonconserved mutations in the Z-score serial maps, it is found that in most mutations for the *H*-type and the *P*-type sites, there is a boundary for separating the mutations into conserved or nonconserved mutations for the native conformation. As shown in Fig. 2, and Fig. 3, such a boundary has been indicated by a line. Below the line the mutations will change the native conformation from the original one to a new one, and above the line there is no changes of the native conformation for various mutations. Differently, for the *N*-type sites, the disordered distribution of the points relating to various mutations makes it impossible to identify the conserved or nonconserved mutations simply by a boundary line (see Fig. 4). However, there are a few exceptions for the mutations at the *H*-type and the *P*-type sites, for which the mutated sequences have different native conformations. Several examples are indicated as open circles in Fig. 2(b) and Fig. 3(b), respectively. Interestingly, these excepted points of the mutated residues are very close to the boundary line [see Fig. 2(b) and Fig. 3(b)].

When the conserved mutations and the nonconserved mutations can be separated by a boundary line, these mutations are defined as the successful separating mutations (SSM), otherwise the failure separating mutations (FSM). A statistics on the SSM and the FSM can be obtained for the *H*-type, the *P*-type, and the *N*-type sites. Thus, the ratios, R_s and R_f, of the SSM and the FSM to the total number of mutations are also listed in Table IV, respectively. One can see that the values of R_s for both the *H*-type and the *P*-type sites are higher than that for the *N*-type sites, implying that the regularity of the conservation of the native conformation for the serial mutations for both the *H*-type and the *P*-type sites. Therefore, to the serial mutations for the *H*-type and the *P*-type sites, we can determine in general what kinds of mutations conserve the native conformation simply by finding out the boundary line rather than detecting the mutated sequences throughout the compact conformation space. The separation of the conserved and nonconserved mutations relates to the monotonic variation of the values of Z-score versus the strength of the hydrophobicity of the mutated residues at both the *H*-type and the *P*-type sites, and reflects the special role of these two type sites in the thermodynamic stability of proteins. As mentioned before, the *H*-type sites relate to the hydrophobic core in proteins. For each single *H*-type site, the mutations prefer to the hydrophobic residues rather than to the hydrophilic ones. As a matter of fact, the preference of the *H*-type sites to the hydrophobic residues is a cooperative result between the *H*-type site itself and other hydrophobic sites all around it. It is this cooperativity of all the hydrophobic residues, which consists the hydrophobic core, that contributes to the stability of the native conformation. Therefore, to a small protein described by the 27 lattice model, if the mutation occurs in the hydrophobic core by a hydrophilic residue, it is apparently unfavorable in energetics for the original native conformation [shown as the decrease of values of the Z-score like Fig. 2(a)]. Since the energetic constraint from the hydrophobic residues topologically around the mutated residues, it is difficult to induce another hydrophobic core to contribute another native conformation from the compact conformation space until the mutation makes the stability of the native conformation low enough. Thus, during the serial mutations for the *H*-type sites, the increasing in the values of Z-score of the mutated sequences is nearly linear with the increase in the hydrophobicity of the mutated residues and also indicates the increasing in the stability of the native conformation. A similar result can also be obtained for the *P*-type sites. Because the hydrophilic constraint of the *P*-type sites, the favorable residues are of the hydrophilic ones. Thus, the values of Z-score of the mutated sequences decrease as the hydrophobicity of the mutated

TABLE V. The correlation coefficients between P_f serial maps and the hydrophobicity for nine sites of the first 10 sequences in Table I. The negative coefficient relates to the inverse proportion in the P_f serial maps to the hydrophobicity.

	Sequence									
Site	1	2	3	4	5	6	7	8	9	10
1	−0.874	−0.832	−0.781	−0.755	−0.856	−0.864	−0.882	−0.817	−0.885	−0.818
13	0.783	0.792	0.878	0.781	0.706	0.900	0.759	0.586	0.686	0.728
15	−0.778	−0.754	−0.891	−0.470	−0.812	−0.751	−0.836	−0.664	−0.652	−0.716
17	0.544	0.423	0.230	0.088	0.834	−0.223	0.744	0.608	0.883	0.848
19	0.952	0.955	0.926	0.932	0.901	0.916	0.944	0.918	0.935	0.575
21	0.790	0.641	0.658	0.131	0.376	0.426	0.502	0.618	0.488	0.634
25	0.751	0.868	0.861	0.942	0.827	0.878	0.896	0.916	0.865	0.255
26	0.947	0.903	0.917	0.920	0.939	0.921	0.927	0.897	0.924	0.872
27	0.927	0.958	0.783	0.893	0.878	0.879	0.923	0.938	0.914	0.905

residues increases, and so does the stability of the native conformation.

However, for a *N*-type site, the situation is not the same as for both cases of the *H*-type and the *P*-type sites. The favor of the *N*-type sites to the hydrophilic residues is obscure, and because it is often between the *H*-type and *P*-type sites, the cooperative constraint from the residues around the site is not consistent. The effect of the stability of the mutated sequences depends on the balance of the contribution of the hydrophobic residues and the hydrophilic residues all around the mutated residues. Because the interactions between 20 naturally occurring residues are different, the variation of the stability with the hydrophobicity of the mutated residues presents disordered distribution as shown in Fig. 4. For the serial mutations at the *N*-type sites, the decrease in the hydrophobicity of the mutated residues may result in the increase in the stability of the native conformation. However, at the same time, such decrease in the hydrophobicity combined with the plastic surroundings will cause another hydrophobic core, and then contributes another more energetic favorable conformation than the original native conformation. Residues *L* to *F* in Fig. 4 relate to such examples. A reversed case happens for residues *S* to *N* in Fig. 4 also.

The role of the *H*-type, the *P*-type, and the *N*-type sites in determining the stability of protein native conformation can be termed as an ''edge effect.'' The hydrophobic core and the hydrophilic surface can be regarded as two strongly cooperative blocks in proteins, while the loose part between them is the intermediate area with less constraint. Because of strong cooperation in the hydrophobic and the hydrophilic blocks, the effect of single mutation in these two blocks on the native conformation is consistent. Until the single mutation with some kinds of residues makes the stability of the original native conformation low enough, a new native conformation becomes stable. Because of less constraint of the intermediate area, any a single mutation in this part may changes the native conformation of proteins unexpectedly.

Therefore, the instability of the original native conformation related to the hydrophobicity of the mutated residues suggests that the native conformation of a protein is determined by the relative order of the hydrophobic and the hydrophilic residues in the protein sequences. Another point which should be addressed is the limitation of the Z-score in describing the stability of proteins. In the Z-score design, it always happens that a sequence with a high value of Z-score is accepted as more native for the target conformation. From the above discussion on the effects of the *H*-type, the *P*-type, and the *N*-type sites, the existence of the *N*-type sites accounts for such limitation of the Z-score method since the stability is not well related to the value of the Z-score when a mutation in made at the *N*-type sites.

B. Relationship between thermodynamic stability and kinetic foldability

The study on the serial maps of Z-score suggests that the *H*-type, the *P*-type, and the *N*-type sites contribute differently to the stability of the native conformation due to different fitness of the hydrophobicity of the mutated residues. How do these three types of sites contribute to the kinetic features of the proteins, and are their thermodynamic features and the kinetic features consistent to the serial mutations at three types of sites? To explore these questions and avoid overweight computer simulations, we select nine sites (listed in Table V and VI) from the first 10 sequences (listed in Table I) to calculate the P_f serial maps (see method section). We then plot P_f versus the Z-score of each serial mutations at nine selected sites for describing the relationship between thermodynamic and kinetic features of the proteins. It is found that there are three kinds of relationship between P_f and the Z-score: (1) the positive correlation, (2) the negative correlation, and (3) the dispersive correlation. Some examples are plotted in Fig. 6.

Figure 6(a) shows one case of the positive correlation for site 26 in sequence 2 in Table I. Site 26 belongs to the *H*-type sites and its serial map of Z-score is direct proportional to the increasing in the hydrophobicity of the mutated residues. Since site 26 is of the *H*-type site, the stronger the hydrophobicity of the residues at site 26, the more stable the native conformation of the sequence is. Strong hydrophobic interaction leads to fast folding. Thus, the relation of the P_f serial map versus the value of Z-score is direct proportional. As a result, this means that the more stable native conformation, the faster the model protein folds.

However, for the folding nuclei as shown in Fig. 1(b), we find that some sites also show the positive correlation but

TABLE VI. The correlation coefficients of the P_f serial maps versus the values of Z-score for nine sites of the first 10 sequences in Table I. The negative coefficient represents the inverse proportion of the values of P_f and the Z-scores.

	Sequence									
Site	1	2	3	4	5	6	7	8	9	10
1	−0.680	−0.598	−0.685	−0.726	−0.749	−0.729	−0.875	−0.594	−0.791	−0.665
13	−0.945	−0.942	−0.937	−0.942	−0.883	−0.902	−0.846	−0.886	−0.828	−0.768
15	0.955	0.936	0.946	0.748	0.947	0.939	0.883	0.915	0.871	0.907
17	0.524	0.502	0.158	0.008	0.806	−0.222	0.723	0.596	0.744	0.938
19	0.961	0.948	0.952	0.919	0.933	0.913	0.933	0.902	0.945	0.710
21	0.788	0.655	−0.575	0.321	0.521	0.500	0.708	0.708	0.475	0.659
25	0.727	0.583	0.595	0.665	0.757	0.658	0.781	0.865	0.711	0.610
26	0.964	0.959	0.955	0.971	0.965	0.967	0.975	0.975	0.976	0.949
27	0.842	0.822	0.852	0.776	0.818	0.722	0.888	0.8789	0.776	0.865

due to different reasons, e.g., the case of site 15 shown in Fig. 6(b). Site 15 locates at the corner of the folding nucleus, and it contributes to the folding nucleus not constructively, but stably. At such kinds of sites, the increase in the hydrophobicity stabilizes the folding nucleus, but may result in high probability of forming frustrated contacts during the formation of the folding nucleus. It is apparent that for site 15 the frustration effect during the folding process is more important than its stable effect on the folding nucleus. Therefore, the P_f serial map for site 15 is in inverse proportion to the hydrophobicity of the mutated residues. In addition, since site 15 belongs to the *P*-type site, its value of Z-score is inversely proportional to the hydrophobicity of the mutated residues, too. Thus, the relation of P_f versus the Z-score is directly proportional. Noting that in Fig. 6(a) and Fig. 6(b) the directly proportional relationship between P_f and Z-score indicates the thermodynamic stability and kinetic foldability being consistent for these two sites.

The case of negative correlation between P_f versus the Z-score is found for site 1 as shown in Fig. 6(c). Considering the behavior of this site in the folding process, we find that although it is not in the folding nuclei, the non-native con-

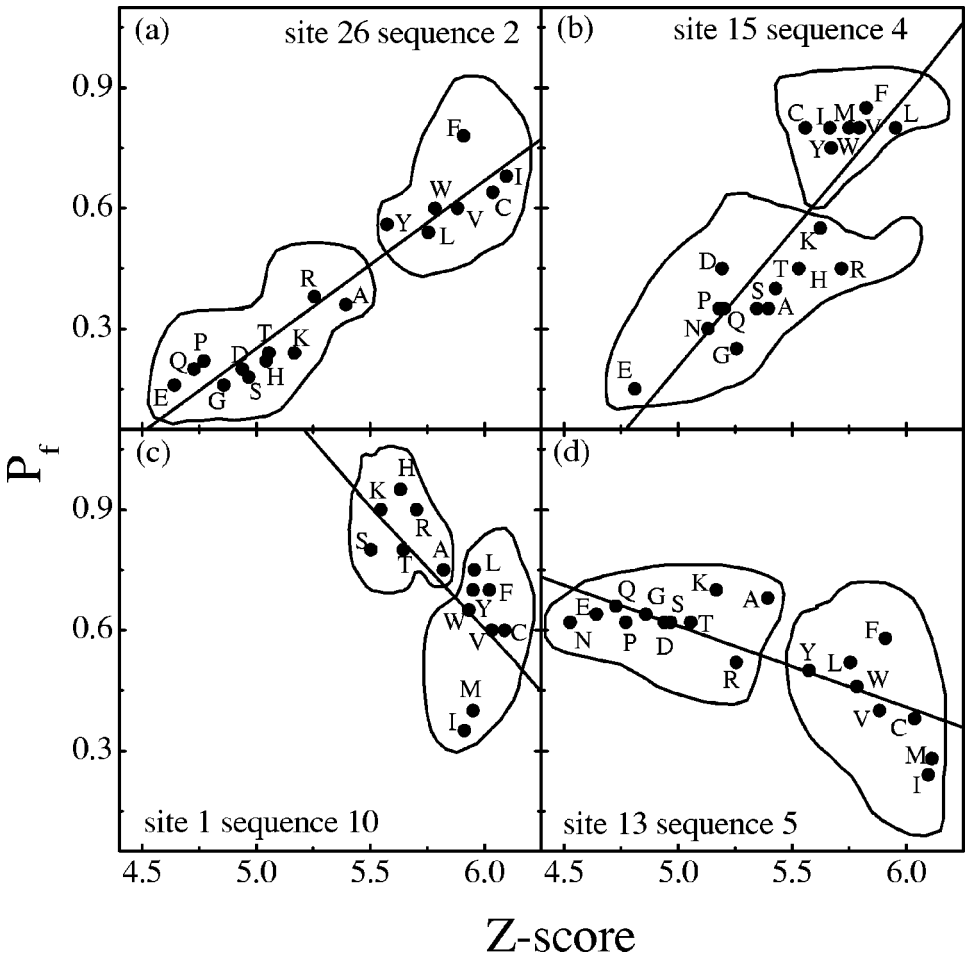

FIG. 6. The value of P_f versus the value of Z-score for sequences in Table I. (a) Plot for site 26 in sequence 2. The correlation coefficient is 0.9588. (b) Plot for site 15 in sequence 4. The correlation coefficient is 0.7479. (c) Plot for site 1 in sequence 10. The correlation coefficient is −0.6650. (d) Plot for site 13 in sequence 5. The correlation coefficient is −0.8827.

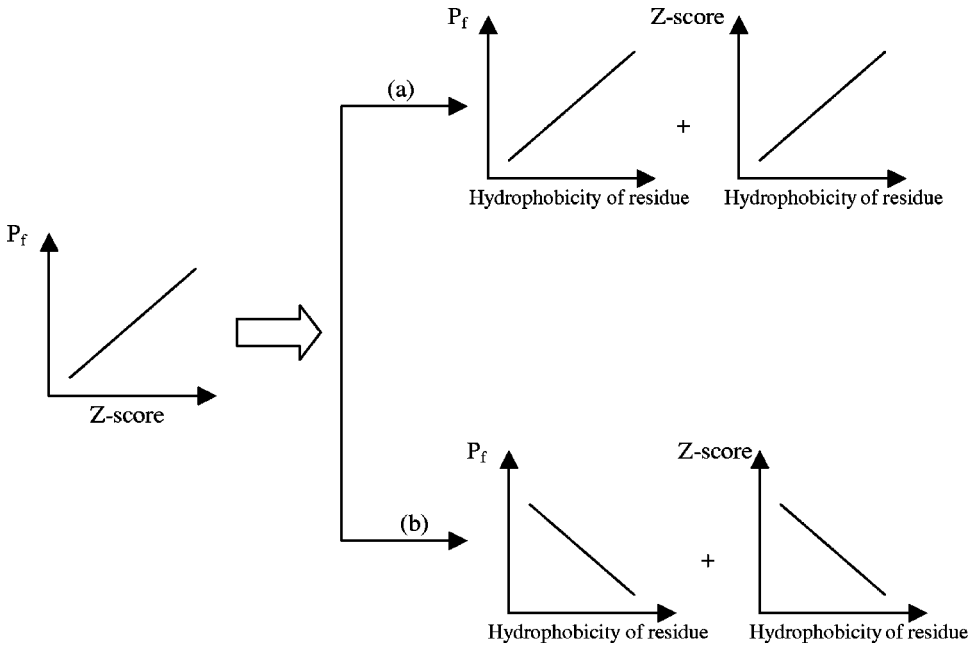

FIG. 7. The sketch plots for the direct proportional relationship between the values of P_f and the values of Z-score for the model proteins with the MJ potentials. Case (a) relates to the core site as shown in Fig. 6(a), while case (b) relates to the surface site as shown in Fig. 6(b).

tacts between site 1 and site 10 and 14 is highly populated in the folding process. The formation of these non-native contacts may be a necessary condition to form the folding nuclei, but too strong non-native contacts may block the formation of the native contacts. Thus for site 1, the weaker the hydrophobicity of the mutated residues in this site, the higher the corresponding P_f value is, that is, the folding becomes fast. Site 13 is also a corner site in the folding nuclei, and contrasting to the site 15, the stable effect of site 13 is dominated by the frustrated effect. Thus, its P_f serial maps are inversely proportional to the hydrophobicity of the mutated residues. Furthermore, P_f versus the Z-score serial map of site 13 is inversely proportional to the hydrophobicity of the mutated residues as shown in the Fig. 6(d).

Following the discussion above, it is found that some sites can take completely opposite role in the thermodynamic stability and kinetic foldability as others do. This indicates that the contribution of the sites to both the thermodynamic stability and the kinetic folding of the protein is not always consistent. Table V and Table VI list the 9×10 correlation coefficient between the values of P_f and Z-score. From Table V, it is found that the variation of the values of P_f still shows three types just like those shown for the serial maps of Z-score as the increase in the hydrophobicity of the mutated residues.

More clearly, two sketch plots are used for understanding the above-mentioned correlation between the values of P_f and the Z-scores. In Fig. 7, the sketch map presents two cases in which P_f versus the Z-score varies with a positive correlation. For the first case in Fig. 7, the values of P_f and the values of Z-score are both direct proportional to the increasing in the hydrophobicity of the mutated residues. This second case plotted in Fig. 7 represents the values of P_f and the values of Z-score are both inversely proportional to the increasing in the hydrophobicity. The sketch plot in Fig. 8 shows the two cases which induce to the inverse proportion of the values of P_f to the values of Z-scores. In the first case, the sites belong to the *H*-type sites, and the P_f serial map is inversely proportional to increasing in the hydrophobicity of the mutated residues. The second case in Fig. 8 corresponds to a *P*-type site such as site 13. As a result, the final P_f versus the Z-score shows an inversely proportional relation.

It is worthy to note that except the linear behavior in P_f versus the Z-score serial maps, there is some disordered relationship between P_f and the values of Z-score. This not only corresponds to the *N*-type sites, but also corresponds to the sites of which the P_f serial map is random. It suggests that the kinetic role of these sites also include three types just like the situation for the thermodynamic role as discussed above.

C. Effects of modified MJ interaction potentials

Since all the elements in the MJ matrix are negative, there exists a strong tendency of hydrophobicity between all pairs of the contacted residues. Thus, it is important to check whether this affects our mentioned above results or not. Similarly, by using the MMJ potentials, six sequences are designed as the ''wild'' sequences for the target conformation shown in Fig. 1(a) (see Table VII). Following the schemes mentioned above, all features, such as the Z-scores and MFPT and so on, have been worked out. It is found that the values of Z-score of these six sequences are generally higher than those of the sequences in Table I and the MFPTs are also apparently small, namely around 10^5 MCs. However, these does not mean that these sequences are better designed than the sequences listed in Table I. Such difference is due to different properties between two kinds of interaction potentials. Actually, the MMJ potentials reduce the overestimated hydrophobic force in the MJ potentials.[49] This makes that the model proteins have less energy deviation and frustration than those with the MJ potentials.

Similarly, the serial mutations are made for these ''wild'' sequences and the Z-score serial maps are obtained (data not

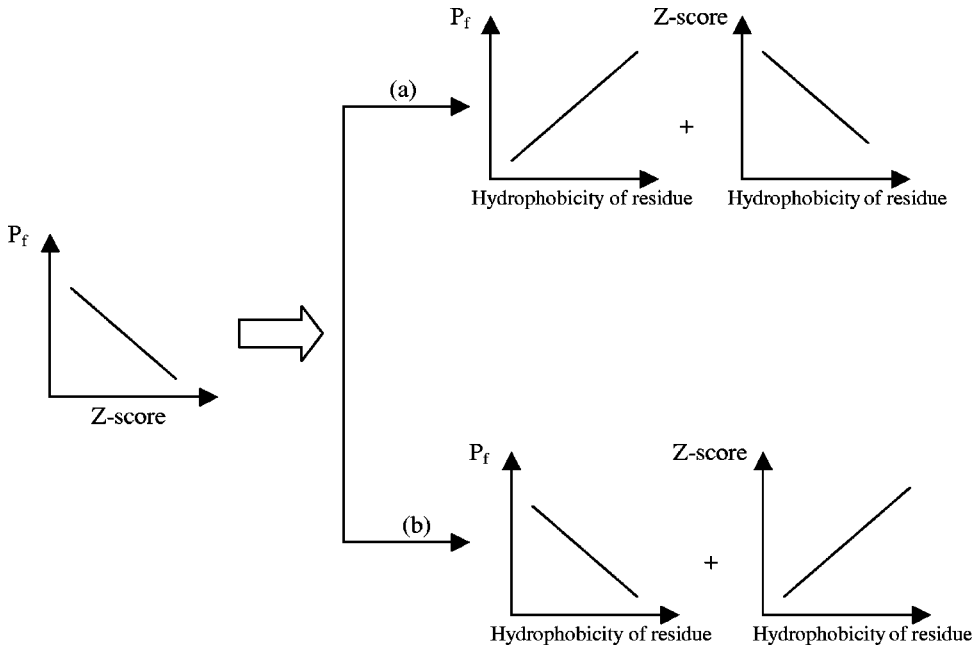

FIG. 8. The sketch plots for the inverse proportional relationship between the values of P_f and the values of Z-score for the model proteins with the MJ potentials. Case (a) relates to the sites as shown in Fig. 6(c) and Fig. 6(d). Case (b) relates to the *P*-type sites.

shown). From our results, the classification for the *H*-type, the *P*-type and the *N*-type sites still exists, and almost the same features as Fig. 2 to Fig. 4 for these three type sites have been seen (see Table VIII). This indicates that the main natures of the MJ potentials and the MMJ potentials are the same, and provides that the types of sites in a sequence may depend on the topology of its native conformation. However, the number of the *P*-type sites in the sequences is small and the number of the *N*-type sites is large compared with the case of the MJ potentials. The change of the number of the *P*-type sites induces a different ratio between the *H*-type sites and the *P*-type sites. As mentioned before, in Table III, the ratio between the *H*-type sites and the *P*-type sites is about 1:2, which is not consistent with the ratio of $H:P$ in real proteins. Differently, for the case of the MMJ potentials the ratio between the *H*-type sites and the *P*-type sites is about 1:1, which is consistent with the ratio of $H:P$ in real proteins. Clearly, the unreasonable ratio between the $H:P$ in the model proteins with the MJ potentials is due to the overestimation of the hydrophobic effect for some *P*-type sites. Because of the overestimation, more *P*-type sites in a sequence are needed under the stable pressure of the model proteins to balance the strong hydrophobic force. While in the model proteins with the MMJ potentials, the hydrophobic force is reduced, thus some of the *P*-type sites convert into the

TABLE VII. The residue composition, the Z-score and the MFPT of six well-designed sequences by the Z-score method for the lattice protein model [Fig. 1(a)] with the modified MJ potentials. The sequence indexes follow their related values of Z-score.

	Sequence	Z-score	MFPT ($\times 10^5$ MCs)
1	WVGCSPGKNRDIGHTQYTFGLRAPEMY	7.41	1.39
2	WVTCGPGRNKDISHGQYGFTLRAPEMY	7.41	2.02
3	WVGCSPGRNKDIGHTQYTFGLRAPEMY	7.38	2.97
4	WVGCTPGRNKDIGHSQYTFGLRAPEMY	7.34	2.21
5	WVSCGPGRNKDITHGQYGFTLRAPEMY	7.33	1.86
6	VQGISAHRGRDPGMTPYTFGYKNCELW	7.07	2.48

TABLE VIII. The distribution of the *H*-type, *P*-type, and *N*-type sites in six sequences of Table VII.

	Sequence					
Site	1	2	3	4	5	6
1	*H*	*H*	*H*	*H*	*H*	*H*
2	*H*	*H*	*H*	*H*	*H*	*H*
3	*N*	*N*	*N*	*N*	*N*	*P*
4	*H*	*H*	*H*	*H*	*H*	*H*
5	*N*	*N*	*N*	*N*	*N*	*N*
6	*N*	*N*	*N*	*N*	*N*	*H*
7	*P*	*P*	*P*	*P*	*P*	*N*
8	*P*	*P*	*P*	*P*	*P*	*P*
9	*P*	*P*	*P*	*P*	*P*	*P*
10	*P*	*P*	*P*	*P*	*P*	*N*
11	*P*	*P*	*P*	*P*	*P*	*P*
12	*H*	*H*	*H*	*H*	*N*	*N*
13	*N*	*N*	*N*	*N*	*N*	*N*
14	*N*	*N*	*N*	*N*	*N*	*H*
15	*N*	*N*	*N*	*N*	*N*	*P*
16	*P*	*P*	*P*	*P*	*P*	*N*
17	*N*	*N*	*N*	*N*	*N*	*N*
18	*P*	*N*	*P*	*P*	*N*	*N*
19	*H*	*H*	*H*	*H*	*H*	*H*
20	*N*	*N*	*N*	*N*	*N*	*N*
21	*H*	*H*	*H*	*H*	*H*	*N*
22	*P*	*P*	*P*	*P*	*P*	*P*
23	*N*	*N*	*N*	*N*	*N*	*N*
24	*N*	*N*	*N*	*N*	*N*	*N*
25	*N*	*N*	*N*	*N*	*N*	*N*
26	*H*	*H*	*H*	*H*	*H*	*H*
27	*N*	*N*	*N*	*N*	*N*	*N*

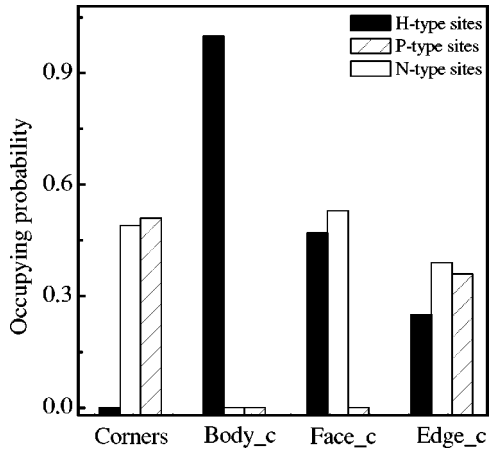

FIG. 9. The histogram of the occupying probabilities of the *H*-type, the *P*-type, and the *N*-type sites at four kinds of positions of lattice model proteins with the MMJ potentials. These positions are the corners, body center, face centers, and edge centers.

N-type sites. However, this change from the *P*-type sites to the *N*-type sites makes the number of the *N*-type sites is near to half of the total sites. The large number of the *N*-type sites may come from the speciality of the cubic lattice model itself. As is known, the completely inside site in the cubic lattice model is only one, i.e., the body center site. Although the face center sites is often regarded as the hydrophobic core sites in the cubic model, it is a surface part of the cubic. Therefore, the irrational body and surface ratio of the 27 lattice model makes the number of the *N*-type sites be more than the numbers of the *H*-type sites and the *P*-type sites.

Figure 9 shows the histogram of occupying probabilities for three types of sites for the case of the MMJ potentials. From Fig. 9, it can be seen that at the face center sites the ratio between the *H*-type sites and the *P*-type sites is 0.47:0, at the corner sites the ratio between the *H*-type sites and the *P*-type sites is 0:0.51, while at the edge center sites, the ratio between the *H*-type sites and the *P*-type sites is 0.25:0.36. These three ratios indicate that in the model proteins with the MMJ potentials, the face center sites tend to accept the *H*-type sites, the corner sites tend to accept the *P*-type sites, while the edge center sites tend to accept the *N*-type sites. This is consistent with that in the model proteins with the MJ potentials, and also indicates that the *H*-type sites are related to the hydrophobic core sites and the *P*-type sites to the hydrophilic surface sites in real proteins, while the *N*-type sites are related to the connected sites between the hydrophobic core sites and the hydrophilic surface sites. The same conclusions from the model proteins with two kinds of interaction potentials suggest that the specific tendency of three types of sites to the sites of real proteins does relate to the topological characteristic of the protein structures.

In addition, we have checked the conservation of the native conformation for different serial mutations at different sites. Our results show that for the model proteins with the MMJ potentials, the ''edge effect'' of the three types of site is still effective, i.e., the boundary line lies in the Z-score serial maps of the *H*-type and *P*-type sites to separate the conserved and nonconserved mutations. We have also studied the kinetic role of three types of sites. We find that the correlations between the P_f serial maps and the hydrophobicity, between the values of P_f and the values of Z-score are somewhat low. This may result from the positive elements in the MMJ potentials. Thus, a single site mutation with reversed sign of interaction potentials for the model proteins may change the kinetic features of folding significantly. This deserves further detailed studies.

V. CONCLUSION

The consistence of the thermodynamic stability and kinetic foldability of proteins has been argued during the study on protein folding for several years.[52–54] Although it has been discussed that the profound stability would induce fast folding of protein,[55] and the fast folding also would induce good stability for protein,[56] the correlation between the thermodynamic stability and the kinetic folding is not corresponded accurately.[56] Then, are the kinetic constraints more important than the thermodynamic ones during the protein folding? Similarly, is the kinetic pressure more important than the stable pressure during the protein evolution?[57,58]

In this paper, based on the lattice protein model, we study the effects of various sites and mutations with various residues on the thermodynamic stability and kinetic foldability of protein model chains with two kinds of potentials, namely the original MJ potentials by Miyazawa and Jernigan and the modified MJ potentials by Thirumalai *et al.* The mutations at a site is realized by replacing a residue with other 19 kinds of naturally occurring residues, and such a kind of mutations is termed as the serial mutations. It is found that the various sites can be classified into three types, namely the hydrophobic (*H*) type, the hydrophilic (*P*) type and the neutral (*N*) type. For the *H*-type (or the *P*-type) sites, the thermodynamic stability increases (or decreases) as the hydrophobicity of the mutated residues increases. In general, the *H*-type and the *P*-type sites relate to the hydrophobic core and the hydrophilic surface. According to the dominant effect of hydrophobic force in constructing protein structures, the *H*-type sites tend to accept hydrophobic residues, while the *P*-type sites tend to accept hydrophilic residues. Comparing to those of the *N*-type sites, these tendencies of the *H*-type sites and the *P*-type sites are relevant to the thermodynamic stability of proteins.

However, from the aspect of folding kinetics, the results with the MJ potentials and the MMJ potentials are some what different. The model proteins with the MMJ potentials fold faster in general than those with the MJ potentials. These difference mainly come from the strengths of the hydrophobic forces in the MJ potentials and the MMJ potentials. It is known that, the energetic effect and the entropic effect are two primary effects controlling the folding process of proteins. Small energetic change in folding may cause the changes in the shape or the roughness of the folding landscape, while the entropic change may cause large changes in the folding landscape or make some trips or barriers, resulting in the loss of some folding pathways. In the case of the MJ potentials, the strong hydrophobic force dominates the shapes of the folding funnel, and makes the energetic effect take the leading role in the folding. Thus, in the P_f serial

maps, the values of P_f vary as the hydrophobicity of residues linearly. In the case of the MMJ potentials, the hydrophobic force is reduced, and even some positive terms appear. Therefore, due to the existence of repulsive force, single site mutations may increase the complexity of the folding process. However, a qualitatively correlation, depending on the sites for mutations, between the thermodynamic stability and kinetic foldability can still be found.

Furthermore, for the MJ potentials, although the values of P_f often varies with the Z-score linearly, the increase in the foldability for the *H*-type and the *P*-type is not always consistent with the increase in the thermodynamic stability. Actually, whether the *H*-type sites or the *P*-sites contribute to the thermodynamic and kinetic properties of proteins consistently depends on their relative positions in the folding nucleus of the protein. As we know, protein folding is argued as a nucleation-condensation process, and how the folding nucleus forms rapidly and stably is crucial to the folding rate. The increase in the thermodynamic stability for certain *H*-type or *P*-type sites which favors to the stabilization and rapid formation of folding nucleus will induce to a consistent increase in the foldability, otherwise, an inconsistence in thermodynamic stability and foldability will be found. Therefore, there are some *H*-type sites and *P*-type sites for which the increase both in thermodynamic stability and in foldability is consistent, such as site 26 of sequence 2, site 15 of sequence 4. Differently, there are some *H*-type sites and *P*-type sites for which the increase in thermodynamic stability and foldability is inconsistent, such as site 1 of sequence 1, site 13 of sequence 5. This reflects the complexity in protein systems. In addition, there are many ingredients relating to the folding properties, such as the topology of the native structures. Due to the difference between various properties, the consistence between them may not be always realized, especially some features after mutations. Here, different effects of *H*-type sites or *P*-type sites on P_f illustrate an example.

Proteins specify their thermodynamic features and kinetic features by natural evolution which carries out selections to satisfy certain requirements as possible as they can. Consequently, the correlation between the thermodynamic features and the kinetic features is an important topic. Our study shows some relevant results although simple lattice model proteins with two kinds of interaction potentials are used. Our classification for the sites makes the role of the sites on the thermodynamic and kinetic features to be clear in some degree. This observation suggests that there are some further selection rules for sequences and native structures of proteins. In addition, the inconsistence of the folding nucleus and the distribution of the *H*-type sites in a protein may result from that the thermodynamic stability is different from the kinetic foldability in some extent. This gives good understanding on the diversity of the thermodynamic and roperties of proteins discussed previously.[56–58]

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 90103031, 10074030, 10021001, and 10204013), and the Nonlinear Project (973) of the NSM.

[1] E. Ron, *Recent Developments in Theoretical Studies of Proteins* (World Scientific, Singapore, 1996).
[2] L. Regan and W. F. Degrado, Science **241**, 976 (1988).
[3] C. D. Sfatos and E. I. Shakhnovich, Phys. Rep. **288**, 77 (1997).
[4] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[5] D. S. Riddle, J. V. Santiago, S. T. BrayHall, N. Doshi, V. P. Grantcharova, Q. Yi, and D. Baker, Nat. Struct. Biol. **4**, 805 (1997).
[6] P. G. Wolynes, Nat. Struct. Biol. **4**, 871 (1997).
[7] H. S. Chan and K. A. Dill, Proteins: Struct., Funct., Genet. **30**, 2 (1998).
[8] J. Wang and W. Wang, Nat. Struct. Biol. **6**, 1033 (1999).
[9] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[10] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **21**, 167 (1995).
[11] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan, Protein Sci. **4**, 561 (1995).
[12] M. Karplus and A. Sali, Curr. Opin. Struct. Biol. **5**, 58 (1995).
[13] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[14] A. Matouschek, J. T. Kellis, Jr., L. Serrano, and A. R. Fersht, Nature (London) **340**, 122 (1989).
[15] D. P. Goldenberg, R. W. Frieden, J. A. Haack, and T. B. Morrison, Nature (London) **338**, 127 (1989).
[16] M. E. Milla, B. M. Brown, C. D. Waldburger, and R. T. Sauer, Biochemistry **34**, 13914 (1995).
[17] V. P. Grantcharova, D. S. Riddle, J. V. Santiago, and D. Baker, Nat. Struct. Biol. **5**, 714 (1998).
[18] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).
[19] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[20] K. F. Lau and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 638 (1990).
[21] R. A. Broglia, G. Tiana, H. E. Roman, E. Vigezzi, and E. I. Shakhnovich, Phys. Rev. Lett. **82**, 4727 (1999).
[22] H. Li, C. Tang, and N. S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[23] G. Tiana, R. A. Broglia, and E. I. Shakhnovich, Proteins: Struct., Funct., Genet. **39**, 244 (2000).
[24] J. Wang and W. Wang, Phys. Rev. E **65**, 041911 (2002).
[25] G. Tiana, R. A. Broglia, H. E. Roman, E. Vigezzi, and E. I. Shakhnovich, J. Chem. Phys. **108**, 757 (1998).
[26] M. Skorobogatiy and G. Tiana, Phys. Rev. E **58**, 3572 (1998).
[27] A. R. Fersht, R. J. Leatherbarrow, and T. N. C. Wells, Nature (London) **322**, 2840 (1986).
[28] A. R. Fersht, R. J. Leatherbarrow, and T. N. C. Wells, Biochemistry **26**, 6030 (1987).
[29] A. R. Fersht, A. Matouschek, and L. Serrano, J. Mol. Biol. **224**, 771 (1992).
[30] A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **92**, 10869 (1995).
[31] E. Alm and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **96**, 11305 (1999).
[32] J. E. Shea, J. N. Onuchic, and C. L. Brooks III, Proc. Natl. Acad. Sci. U.S.A. **96**, 12512 (1999).
[33] H. Nymeyer, N. D. Socci, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **97**, 634 (2000).
[34] A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **97**, 1525 (2000).
[35] S. B. Ozkan, I. Bahar, and K. A. Dill, Nat. Struct. Biol. **8**, 765 (2001).
[36] T. Veitshans, D. K. Klimov, and D. Thirumalai, Folding Des. **2**, 1 (1997).
[37] D. Thirumalai, D. K. Klimov, and S. A. Woodson, Theor. Chem. Acc. **96**, 14 (1997).
[38] H. Nymeyer, A. E. Garacia, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **95**, 5921 (1998).
[39] D. Thirumalai and D. K. Klimov, Curr. Opin. Struct. Biol. **9**, 197 (1999).
[40] L. Lewyn and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **98**, 13014 (2001).
[41] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **291**, 177 (1999).
[42] K. W. Plaxco, S. Larson, I. Ruczinski, D. S. Riddle, E. C. Thayer, B. Buchwitz, A. R. Davidson, and D. Baker, J. Mol. Biol. **298**, 303 (2000).
[43] L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **308**, 123 (2001).
[44] S. M. Larson, I. Ruczinski, A. R. Davidson, D. Baker, and K. W. Plaxco, J. Mol. Biol. **316**, 225 (2002).
[45] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
[46] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, J. Mol. Biol. **252**, 460 (1995).
[47] D. K. Klimov and D. Thirumalai, J. Mol. Biol. **282**, 471 (1998).
[48] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).

[49] M. R. Betancourt and D. Thirumalai, Protein Sci. **8**, 361 (1999).
[50] E. I. Shakhnovich and A. M. Gutin, Protein Eng. **6**, 793 (1993).
[51] J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, and P. G. Wolynes, Folding Des. **1**, 441 (1996).
[52] E. I. Shakhnovich and A. M. Gutin, Nature (London) **346**, 773 (1990).
[53] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).
[54] C. Camacho and D. Thirumalai, Phys. Rev. Lett. **71**, 2505 (1993).
[55] A. Sail, E. I. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[56] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **92**, 1282 (1995).
[57] L. A. Mirny, V. I. Abkevich, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **95**, 4976 (1998).
[58] T. C. Wood and W. R. Pearson, J. Mol. Biol. **291**, 977 (1999).