Correlations between designability and various structural characteristics of protein lattice models

Jian-Yi Yang
School of Mathematics and Computing Science, Xiangtan University, Hunan 411105, China

Zu-Guo Yu[a)]
School of Mathematics and Computing Science, Xiangtan University, Hunan 411105, China and School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Q4001, Australia

Vo Anh
School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Q4001, Australia

(Received 27 November 2006; accepted 11 April 2007; published online 17 May 2007)

Using six kinds of lattice types (4×4, 5×5, and 6×6 square lattices; $3\times3\times3$ cubic lattice; and $2+3+4+3+2$ and $4+5+6+5+4$ triangular lattices), three different size alphabets (*HP*, *HNUP*, and 20 letters), and two energy functions, the *designability* of protein structures is calculated based on random samplings of structures and common biased sampling (CBS) of protein sequence space. Then three quantities *stability* (average energy gap), *foldability*, and *partnum* of the structure, which are defined to elucidate the designability, are calculated. The authors find that whatever the type of lattice, alphabet size, and energy function used, there will be an emergence of highly designable (*preferred*) structure. For all cases considered, the local interactions reduce degeneracy and make the designability higher. The designability is sensitive to the lattice type, alphabet size, energy function, and sampling method of the sequence space. Compared with the random sampling method, both the CBS and the Metropolis Monte Carlo sampling methods make the designability higher. The correlation coefficients between the designability, stability, and foldability are mostly larger than 0.5, which demonstrate that they have strong correlation relationship. But the correlation relationship between the designability and the partnum is not so strong because the partnum is independent of the energy. The results are useful in practical use of the designability principle, such as to predict the protein tertiary structure. © *2007 American Institute of Physics*.
[DOI: 10.1063/1.2737042]

I. INTRODUCTION

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

The study of protein folding is fundamental in both theory and applications. A variety of models have been proposed to explain the protein folding problem. Based on the *HP* model, the concept of *designability* of protein structure has been introduced.[3] The number of sequences which determines the structure as the unique ground state is called the designability of this structure. When many protein sequences have a common native structure, one can say that the structure is highly designable. The highly designable structures are, on average, thermodynamically more stable than other structures.[3] We will introduce the concept of *stability* to elucidate the designability of protein structure in this paper.

There have been many studies trying to account for the designability of protein structure. Li *et al.*[4] employed a simple solvation model leading to a geometrical formulation of the protein folding problem and they found that the highly designable structures are *atypical*. Wang and Yu[5] proposed a quantity called *partnum* (partition number) of each compact structure to explain the difference of designability of protein structures. The protein sequences that have the highly designable structures as their native structure also fold fast on average which is measured by the Z score Δ/Γ, Δ being the average energy difference between the native and excited states, and Γ being the width in energy of the excited states.[6] Govindarajan and Goldstein[7] proposed a model for sequence foldability, which is a thermodynamic measure characterizing how amenable the free-energy landscape is to successful protein folding, and demonstrated that there are some structures that can be better optimized than others. There are also many investigations about the foldability of protein sequences.[8–10] Buchler and Goldstein[11] studied the effect of the alphabet size and the foldability requirements on the designability of protein structure. The definitions of the Z score and the foldability are identical in form and we will use such definition in sequence to define the *foldability of structure* in this paper in order to elucidate the designability. Based on the square and cubic lattices, Li *et al.*[12] use the Miyazawa-

a)Author to whom correspondence should be addressed. Electronic mail: yuzg@xtu.edu.cn or z.yu@qut.edu.au

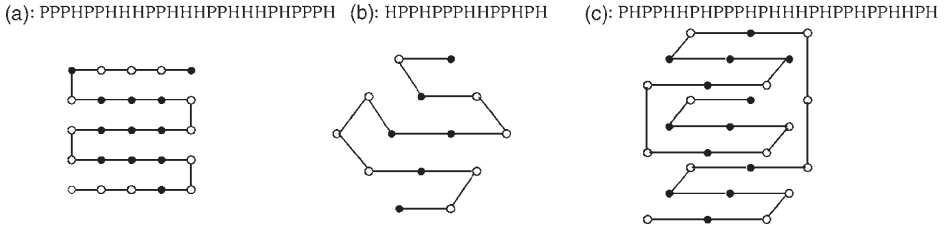

FIG. 1. The three lattice types and the corresponding compact structures of three sequences based on the *HP* model. (a) 5×5 square lattice, (b) $2+3+4+3+2$ triangular lattice, and (c) $3\times3\times3$ cubic lattice. The black beads and the white beads represent the *H* monomer and the *P* monomer, respectively.

Jernigan (MJ) matix[13] to compute the designability and discuss the effect of the alphabet size to the designability.

We may question whether these elucidations are artifacts as (i) they only use two letters (*H* and *P*) to represent the 20 amino acids;[3,5] (ii) they only use square and cubic lattices and do not consider other types of lattices such as triangular lattice as there exists the so-called even-odd problem in the square and cubic lattices;[3–5,11] and (iii) they only consider the contact interactions when calculating the energy of certain structures.[3,4] But the local interactions may contribute to the free energy of the native state of protein.[14] Consequently, we will ask the following:

- Are there highly designable structures which are applicable to other lattice types, alphabet sizes, and energy functions?
- If such structures do exist, what is the relationship between designability, stability, and partnum?
- How to define the foldability of protein structure and what is its relationship with the designability?

In this paper, the definitions of designability, stability, foldability, and partnum of protein structure are given. We calculate these four quantities using six types of lattices (4×4, 5×5, and 6×6 square lattices; $3\times3\times3$ cubic lattice; and $2+3+4+3+2$ and $4+5+6+5+4$ triangular lattices), three different size alphabets (*HP*, *HNUP*, and 20 letters), and two energy functions (36 cases altogether) via the random sampling method when the chain is too long or the number of compact structures is too large to completely enumerate. Then the relationships between designability and the three other quantities are studied by their correlation coefficients. The effect of the alphabet size, lattice type, and energy function on the designability and the relationship between designability and the three other quantities are discussed.

II. METHODS AND MODELS

A. Lattice type

There are many types of lattices; we only consider the square lattice, cubic lattice, and triangular lattice in this paper. They are shown in Fig. 1.

Because of the limitation of computing power,[3,15] we can only use short chains (less than 36) and small lattice size (smaller than 6×6). In this article, we adopt the 4×4, 5×5, and 6×6 square lattices; $3\times3\times3$ cubic lattice; and $2+3+4+3+2$, $4+5+6+5+4$ triangular lattices in the computations. We use *cubic*, *tri.1*, and *tri.2* to represent the $3\times3\times3$ cubic, the $2+3+4+3+2$ triangular, and the $4+5+6+5+4$ triangular lattices, respectively.

B. Alphabet size

The nature proteins are made of 20 amino acids and 20 letters are often used to represent them (A, I, L, M, F, P, W, V, D, E, N, C, Q, G, S, T, Y, R, H, and K).[16] Sometimes, as it is too complex to calculate, the 20 amino acids are designated into two groups by their affinities for water. A well-known model of protein sequences is the *HP* model.[17] In this model, the 20 kinds of amino acids are divided into two groups: hydrophobic (*H*) (or nonpolar) and polar (*P*) (or hydrophilic). But the *HP* model may be too simplistic and lacks sufficient information on the heterogeneity and complexity of the natural set of residues.[18] According to Brown,[16] in the *HP* model, one can divide the polar class into three classes: positive polar, uncharged polar, and negative polar. This model, which is called the detailed *HP* model,[19–21] provides more information than the *HP* model. The eight residues A, I, L, M, F, P, W, and V designate the hydrophobic class; the two residues D and E designate the negative polar class; the seven residues N, C, Q, G, S, T, and Y designate the uncharged polar class; and the remaining three residues R, H, and K designate the positive polar class.[16] We use four letters (*H*, *N*, *U*, and *P*) for their representation: *H* represents the hydrophobic class, *N* the negative polar class, *U* the uncharged polar class, and *P* the positive polar class. The three models of protein sequences (20 letters, 4 letters, and 2 letters) are represented by MJ, *HNUP*, and *HP*, respectively.

C. The formation of compact structure for certain protein sequences

Because of the compact structure of globular proteins and the volume effect,[22,23] we only consider the compact self-avoiding walk[23] which is in fact the well-known Hamiltonian path in the graphic theory. Once we fix the lattice type and the alphabet size of the protein sequence, we can enumerate all the possible maximally compact structures. For example, supposing that we have fixed the 5×5 square lattice and *HP* alphabet size, we can find all the possible compact structures using the computer.[15] A protein sequence is specified by a choice of monomer type at each position on the chains $\{\delta_i\}$, where δ_i can be either *H* or *P*, and *i* is a monomer index. A structure is specified by a set of coordinates for all the monomers $\{\gamma_i\}$. Examples of such compact structures in a 5×5 square lattice, a $2+3+4+3+2$ triangular lattice, and a $3\times3\times3$ cubic lattice are shown in Fig. 1.

TABLE I. The four contact energies in the *HP* model (*HP* matrix).

	H	*P*
H	−2.3	−1
P	−1	0

D. Energy

There are many different methods in measuring the free energy of a protein sequence folding into a particular structure.[3,4,13,14,24,25] The following energy function has been used by Irbäck and Sandelin:[14]

$$E = \kappa E_L + E_G, \tag{1}$$

where

$$E_L = \sum_{i=2}^{N-1} (1 - \cos\theta_i), \tag{2}$$

$$E_G = \sum_{1 \le i < j \le N} E_{\delta_i \delta_j} \Delta(\gamma_i - \gamma_j). \tag{3}$$

Here $\Delta(\gamma_i - \gamma_j) = 1$ if γ_i and γ_j are adjoining lattice sites but i and j are not adjacent in position along the chain and $\Delta(\gamma_i - \gamma_j) = 0$ otherwise. E_L represents the total local interaction energy and E_G represents the total contact energy. The contact energy $E_{\delta_i \delta_j}$ depends on the monomers in contact. For example, if we use the *HP* model, they are E_{HH}, E_{PH}, E_{HP}, and E_{PP} corresponding to *H-H*, *P-H*, *H-P*, and *P-P* contact energies, respectively. The parameter θ_i represents the bend angle formed by sites γ_{i-1}, γ_i, and γ_{i+1}, and it takes a value from the set $\{0, \pi/2\}$ for the square and cubic lattices and from $\{0, \pi/3, 2\pi/3\}$ for the triangular lattices. N is the chain (sequence) length. The remaining parameter κ determines the strength of the local interactions. We will discuss two situations of κ ($\kappa = 0, 0.3$) in this paper to show the effect of local interactions. When we set $\kappa = 0$, it is just the energy function used in Ref. 3.

When we compute the energy using the *HP* model, the four (2×2) contact energies are $E_{HH} = -2.3$, $E_{PH} = E_{HP} = -1$, and $E_{PP} = 0$, which have been used by Li *et al.*[3] For the 20-letter model, the 400 (20×20) contact energies are taken from the MJ matrix.[13] For the detailed *HP* model, we use the MJ matrix to deduce the contact energies. In order to reduce the complexity, we first rearrange the MJ matrix according to the four classes of amino acids and then calculate the average values as the 16 (4×4) interaction energies between the four classes of amino acids. The results are listed in Table II as the *HNUP* matrix. In order to make sure our computing methods are not artifacts, we use the same method to calculate the interaction energies in the *HP* model from the *HNUP* matrix to see whether they satisfy the two conditions proposed by Li *et al.*:[3] (i) $E_{PP} > E_{HP} > E_{HH}$ and (ii) $2E_{HP} > E_{HH} + E_{PP}$. Our calculation results are $E_{HH} = -4.88$, $E_{HP} = E_{PH} = -2.88$, and $E_{PP} = -1.72$. They satisfy conditions (i) and (ii), so we can say that our calculation methods and results are reasonable. The interaction energies we adopt in the calculation of the free energies are all listed in Tables I–III.

TABLE II. The 16 contact energies in the detailed *HP* model (*HNUP* matrix).

	H	*N*	*U*	*P*
H	−4.88	−2.55	−3.25	−2.83
N	−2.55	−1.12	−1.69	−1.90
U	−3.25	−1.69	−2.24	−1.85
P	−2.83	−1.90	−1.85	−1.26

E. The three quantities used to account for designability

In this paper, we make use of the following three quantities to elucidate the designability of protein structure: stability (average energy gap), foldability, and partnum.

Designability. Given a sequence, we compute its energy when it folds into certain structure using Eq. (1). If there are a total of T structures unrelated by rotational and reflection symmetries altogether in a certain lattice (e.g., 1081 in a 5×5 square lattice), we will get T energies. If the lowest energy is unique, we say that the corresponding structure is the native structure for this sequence. Given certain structure S, we denote by N_S the number of sequences that have S as their native structure. Then N_S is defined as the designability of this structure by Li *et al.*[3] There may be some structures with large N_S (i.e., high designability) and some structures with small N_S (even zero) (i.e., low designability).[3] We attempt to account for such difference in the designability of protein structure in this paper.

Stability. Are the highly designable structures, on average, thermodynamically more stable than other structures? For a given sequence, the energy gap δ is defined as the minimum energy required to change the ground-state structure to a different structure.[3] That is to say δ is the difference between the unique lowest energy and the second lowest energy considering the lowest energy is unique. The stability of a structure S can be characterized by the average energy gap ($\overline{\delta_S}$), averaged over the N_S sequences that design the structure.[3] The larger the average energy gap is, the more stable one structure is and we say its stability is higher. Here

$$\delta = E_1 - E_0, \tag{4}$$

$$\overline{\delta_S} = \langle \delta \rangle, \tag{5}$$

where E_0 denotes the lowest energy, E_1 the second lowest energy, and $\langle \cdot \rangle$ the average value over the N_S sequences that have structure S as their common native structure.

Foldability Using the concepts used in the physics of spin glasses, Bryngelson and Wolynes[26,27] considered that two thermodynamic transitions are possible in a protein: one to the fold state at a temperature T_f and the other to a glass state at a temperature T_g. The ratio T_f/T_g is a measure of how fast a given sequence can fold into its native structure.[9,10,26] The larger the ratio is, the faster the sequence will fold into its native structure. In the random energy model,[28] T_f/T_g can be expressed as a monotonically increasing function of the Z score[29]

TABLE III. The 400 contact energies in the 20-letter model (rearranged MJ matrix).

	M	F	I	L	V	W	A	P	E	D
M	−6.06	−6.68	−6.33	−6.01	−5.52	−6.37	−3.99	−4.11	−3.19	−2.90
F	−6.68	−6.85	−6.39	−6.26	−5.75	−6.02	−4.36	−3.73	−3.51	−3.31
I	−6.33	−6.39	−6.22	−6.17	−5.58	−5.64	−4.41	−3.47	−3.23	−2.91
L	−6.01	−6.26	−6.17	−5.79	−5.38	−5.50	−3.96	−3.06	−2.91	−2.59
V	−5.52	−5.75	−5.58	−5.38	−4.94	−5.05	−3.62	−2.96	−2.56	−2.25
W	−6.37	−6.02	−5.64	−5.50	−5.05	−5.42	−3.93	−3.66	−2.94	−2.91
A	−3.99	−4.36	−4.41	−3.96	−3.62	−3.93	−2.51	−2.80	−1.51	−1.57
P	−4.11	−3.73	−3.47	−3.06	−2.96	−3.66	−2.80	−1.18	−1.40	−1.19
E	−3.19	−3.51	−3.23	−2.91	−2.56	−2.94	−1.51	−1.40	−1.18	−1.23
D	−2.90	−3.31	−2.91	−2.59	−2.25	−2.91	−1.57	−1.19	−1.23	−0.96
C	−5.05	−5.63	−5.03	−5.03	−4.46	−4.76	−3.38	−2.92	−2.08	−2.66
Y	−4.92	−4.95	−4.63	−1.26	−4.05	−4.44	−2.85	−2.80	−2.42	−2.25
G	−3.75	−3.72	−3.65	−3.43	−3.06	−3.37	−2.15	−1.72	−1.22	−1.62
T	−3.73	−3.76	−3.74	−3.43	−2.95	−3.31	−2.15	−1.66	−1.45	−1.66
S	−3.55	−3.56	−3.43	−3.16	−2.79	−2.95	−1.89	−1.35	−1.48	−1.46
Q	−3.17	−3.30	−3.22	−3.09	−2.67	−3.16	−1.70	−1.73	−1.33	−1.26
N	−3.50	−3.55	−2.99	−2.99	−2.36	−3.11	−1.44	−1.43	−1.43	−1.33
H	−3.31	−4.61	−3.76	−3.84	−3.38	−4.02	−2.09	−2.17	−2.27	−2.14
R	−3.49	−3.54	−3.33	−3.15	−2.78	−3.56	−1.50	−1.85	−2.07	−1.98
K	−3.11	−2.83	−2.70	−2.63	−1.95	−2.49	−1.10	−0.67	−1.60	−1.32
	C	Y	G	T	S	Q	N	H	R	K
M	−5.05	−4.92	−3.75	−3.73	−3.55	−3.17	−3.50	−3.31	−3.49	−3.11
F	−5.63	−4.95	−3.72	−3.76	−3.56	−3.30	−3.55	−4.61	−3.54	−2.83
I	−5.03	−4.63	−3.65	−3.74	−3.43	−3.22	−2.99	−3.76	−3.33	−2.70
L	−5.03	−1.26	−3.43	−3.43	−3.16	−3.09	−2.99	−3.84	−3.15	−2.63
V	−4.46	−4.05	−3.06	−2.95	−2.79	−2.67	−2.36	−3.38	−2.78	−1.95
W	−4.76	−4.44	−3.37	−3.31	−2.95	−3.16	−3.11	−4.02	−3.56	−2.49
A	−3.38	−2.85	−2.15	−2.15	−1.89	−1.70	−1.44	−2.09	−1.50	−1.10
P	−2.92	−2.80	−1.72	−1.66	−1.35	−1.73	−1.43	−2.17	−1.85	−0.67
E	−2.08	−2.42	−1.22	−1.45	−1.48	−1.33	−1.43	−2.27	−2.07	−1.60
D	−2.66	−2.25	−1.62	−1.66	−1.46	−1.26	−1.33	−2.14	−1.98	−1.32
C	−5.44	−3.89	−3.16	−2.88	−2.86	−2.73	−2.59	−3.63	−2.70	−1.54
Y	−3.89	−3.55	−2.50	−2.48	−2.30	−2.53	−2.47	−3.33	−2.75	−2.01
G	−3.16	−2.50	−2.17	−2.03	−1.70	−1.54	−1.56	−1.94	−1.68	−0.84
T	−2.88	−2.48	−2.03	−1.72	−1.59	−1.59	−1.51	−2.35	−1.97	−1.02
S	−2.86	−2.30	−1.70	−1.59	−1.48	−1.37	−1.31	−1.94	−1.22	−0.83
Q	−2.73	−2.53	−1.54	−1.59	−1.37	−0.89	−1.36	−1.85	−1.85	−1.02
N	−2.59	−2.47	−1.56	−1.51	−1.31	−1.36	−1.59	−2.01	−1.41	−0.91
H	−3.63	−3.33	−1.94	−2.35	−1.94	−1.85	−2.01	−2.78	−2.12	−1.09
R	−2.70	−2.75	−1.68	−1.97	−1.22	−1.85	−1.41	−2.12	−1.39	−0.06
K	−1.54	−2.01	−0.84	−1.02	−0.83	−1.02	−0.91	−1.09	−0.06	−0.13

$$Z\ \text{score} = \frac{\Delta}{\Gamma}, \tag{6}$$

where Δ denotes the average energy difference between the native and the excited states, and Γ denotes the width in energy of the excited states.[6] The Z score for a given sequence is also called foldability.[8–10] The following definitions have been used by Mélin *et al.*:[6]

$$\Delta = \frac{1}{N_C}\sum_{\alpha>0}(E_\alpha - E_0), \tag{7}$$

$$\Gamma^2 = \frac{1}{N_C}\sum_{\alpha>0}(E_\alpha)^2 - \left(\frac{1}{N_C}\sum_{\alpha>0}E_\alpha\right)^2, \tag{8}$$

where $E_\alpha(\alpha>0)$ denotes the energies of the excited compact structures, E_0 is the lowest compact state energy, and N_C is the number of excited compact structures.

Notice that the foldability defined above is used to measure how fast the protein will fold, but our motivation here is to account for the designability of structures. It is necessary to transform such definition for structures. Inspired by the idea of average energy gap characterizing the stability of a structure introduced by Li *et al.*,[3] in this paper we propose the conception of foldability of certain protein structure S. It

TABLE IV. The total number of structures (Hamiltonian paths) unrelated by symmetry, which is denoted by T.

Lattice type	Number of structures (T)
4×4	69
5×5	1081
6×6	57 337
Cubic	103 346
Tri.1	2571
Tri.2	1 475 782

is defined as the average Z score, averaged over the N_S sequences that have the structure S as their common native state.

$$\text{Foldability of certain structure } S = \langle Z \text{ score} \rangle, \tag{9}$$

where $\langle\cdot\rangle$ denotes the average value of Z scores over N_S sequences having structure S as their common native state.

Partnum Wang and Yu[5] proposed a quantity called partnum to explain the difference of designability of protein structure. Given one structure S, if the chain length is N, then there are $N-1$ steps during the self-avoiding walk. If the ith step has a total of C acceptable choices not being symmetrically related, then this step gives a number called partnum of the ith step $p_i=\ln(1/C)$. Adding all the $N-1$ numbers and then dividing the sum by $N-1$, we get the partnum P of the structure S,

$$P(S) = \frac{1}{N-1}\sum_{i=1}^{N-1} p_i. \tag{10}$$

It is sequence independent and is easy to compute as it does not need to consider the interaction detail.

III. RESULTS AND DISCUSSION

First, we find all the compact structures (Hamiltonian paths) of a certain lattice. Using the back-trading algorithm, we can enumerate all the compact structures unrelated by rotational and reflection symmetries. As proteins are produced successively from one end to another,[5] we do not consider the reverse-labeling symmetries of structures here. Such symmetry type has been considered in Refs. 3 and 30 to decrease the total number of compact structures. So there are eight kinds of symmetry in the square lattice, 48 in the cubic lattice, and 4 in the triangular lattice.[27,31] After excluding all the symmetries, the total number of compact structures for these lattices are listed in Table IV. The results have been reported in Refs. 3, 15, 30, and 31.

Second, we calculate the designability, stability, and foldability of protein structure. For a given protein sequence, we compute the energies of the T different compact structures using Eq. (1). If the lowest energy for certain structure S is unique, we say that the native state of the sequence is S and the energy gap δ and Z score of this sequence can be calculated at the same time using Eqs. (4) and (6), respectively. Otherwise, the sequence is said to be degenerate.[3,30] The designability of certain structure can be calculated with the definition in Sec. II E and we denote it by N_S. The stability and the foldability can be calculated from Eqs. (5) and (9), respectively. Then we can discuss the relationship between designability and the two quantities stability and foldability. The correlation coefficient[32,33] is a good tool in studying the relationship between two variables, which has been used by Wang and Yu in Ref. 5,

$$S_{xy} = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{n-1}, \tag{11}$$

$$r = \frac{S_{xy}}{S_x S_y}, \tag{12}$$

where $\Sigma x_i y_i$ is the sum of the products $x_i y_i$ for each of the n pairs of measurements, S_x and S_y are the standard deviations of random variables X and Y, S_{xy} is the covariance between X and Y, and r is then the correlation coefficient between X and Y.[32] The closer the value of r is to 1 or -1, the stronger the linear relationship is between the two variables.[32,33] We will use Eqs. (11) and (12) to calculate the correlation coefficients between the designability and the three other quantities defined in Sec. II E.

As the total structures or the protein sequences are too large to completely enumerate, samplings of structures and sequences are used in the computing. We use the sampling method in Refs. 3 and 14 to sample the structure. As to the sampling of sequence space, there are many kinds of methods we can use.[3,4,12,14,34–37] A general sampling method of the sequence space is the *random sampling* (RS) which has been used in Refs. 3, 4, 12, and 14. It was shown that in order for a sequence to fold into a given target structure, its energy in that structure should follow below a certain threshold E_c.[38–40] In fact, the sequence space is too vast and many sequences will not have a particularly low energy (below E_c) in a given structure. For a given structure, low energy sequences are so rare that they will not be sampled at all unless the search of sequence space is biased toward these sequences. Therefore in order to sample the sequence with low energy, we can set an energy threshold E_c first and then sample the sequences. The problem we must face is how to decide the energy threshold E_c. Shakhnovich[39] gave out an estimation of E_c

$$E_c = E_m - JN(2\ln\gamma)^{1/2}, \tag{13}$$

where E_m and J are the mean and standard deviation of the energies, respectively, and γ is the number of conformations per monomer.[38–40]

Before we start to sample the sequences, we have to estimate the value of E_m and J. For such estimation, we sample 10^3 random sequences first and then use them to estimate the value of E_m and J. After E_c is calculated with Eq. (13), we can start to sample the sequences according to that threshold. The sampling method contains three steps, mainly: (step 1) generate a sequence randomly; (step 2): calculate the energy E_t on the target structure; (step 3): compare E_t with the threshold E_c. If it is smaller than E_c, such sequence will be sampled; otherwise, the sequence will be excluded. (step 4) Repeat from step 1 to step 3 until the total number of sampling sequences is achieved. Such sampling

TABLE V. The sampling number of compact structures and protein sequences are denoted by n_1 and n_2, respectively. Stru. and Seq. represent the sampling structures and sequences, respectively.

	HP		*HNUP*		MJ	
	Stru. (n_1)	Seq. (n_2)	Stru. (n_1)	Seq. (n_2)	Stru. (n_1)	Seq. (n_2)
4×4	69	65 536	69	50 000	69	50 000
5×5	1081	50 000	1081	50 000	1081	50 000
6×6	2500	50 000	2500	50 000	2500	50 000
Cubic	5000	50 000	5000	50 000	5000	50 000
Tri.1	2571	16 384	2571	50 000	2571	50 000
Tri.2	5000	50 000	5000	50 000	5000	50 000

method can be seen as *common biased sampling* (CBS) as some sequences might be excluded during the sampling process. The total number of sampling structures and sequences for each structure are shown in Table V, which are denoted by n_1 and n_2, respectively.

In the calculations there are many conditions to be considered.

- The strength of the local contact energy. We choose $\kappa=0,0.3$ when computing the energy using Eq. (1) to discuss the effect of local energy.
- The lattice types: 4×4, 5×5, and 6×6 square lattices; 3×3×3 cubic lattice; and 2+3+4+3+2 and 4+5+6+5+4 triangular lattices. We use these six kinds of lattices to compute the corresponding quantities to discuss the effect of lattice type and chain length.
- The alphabet size: *HP* model, detailed *HP* (*HNUP*) model, and the 20-letter (MJ) model. We use these three kinds of alphabet sizes to compute the corresponding quantities to discuss the effect of alphabet size.

Of course, the above three conditions can be integrated in the computation, with $2\times6\times3=36$ different situations altogether.

The designability in the 12 situations (detailed *HP* model, $\kappa=0,0.3$, all lattice types) are shown in Fig. 2. The designability in different situations (36 altogether) are shown in Table VI. From Fig. 2 and Table VI we can see that (i) the designability is much larger obviously in the condition $\kappa=0.3$ than in the condition $\kappa=0$ as there are more structures contributing to the longer tail of the distribution,[3] which can be seen from Fig. 2. In addition, the largest N_s (LN) are all larger in the former condition than those in the latter shown in Table VI. (ii) The local interactions reduce the degeneracy [i.e., the ratio of the number of sequences which have nondegenerate ground states increases measured by NDR defined by (14)]

$$\mathrm{NDR}=\frac{LN}{n_2}\times 100, \tag{14}$$

where n_2 is total number of the samplings of the sequence for the structure with the highest designability LN. NDR represents the ration of the sequences that are not degenerate for the target structure with the highest designability. From Table VI, we can see the values of LN and NDR are different for different lattice types, alphabet sizes, and values of κ. So we conclude that the designability is sensitive to lattice type, alphabet size, and energy function.

The designability against the stability in six of the situations (20-letter model, $\kappa=0$, all lattice types) is shown in Fig. 3. The lines in the figures are the fitting results of the data. From the figure we can see that the larger the designability is, the larger the stability is. It indicates that the highly designable structure must be highly stable at the same time. From an evolutionary point of view, Li *et al.*[3] speculate that highly designable structures are more likely to have been chosen through random selection of sequences in the primordial age and such structures are stable against mutations. This explanation is in fact in line with Darwin's evolution theory of species. It is one application of the natural selection principle in protein tertiary structures. The protein sequences

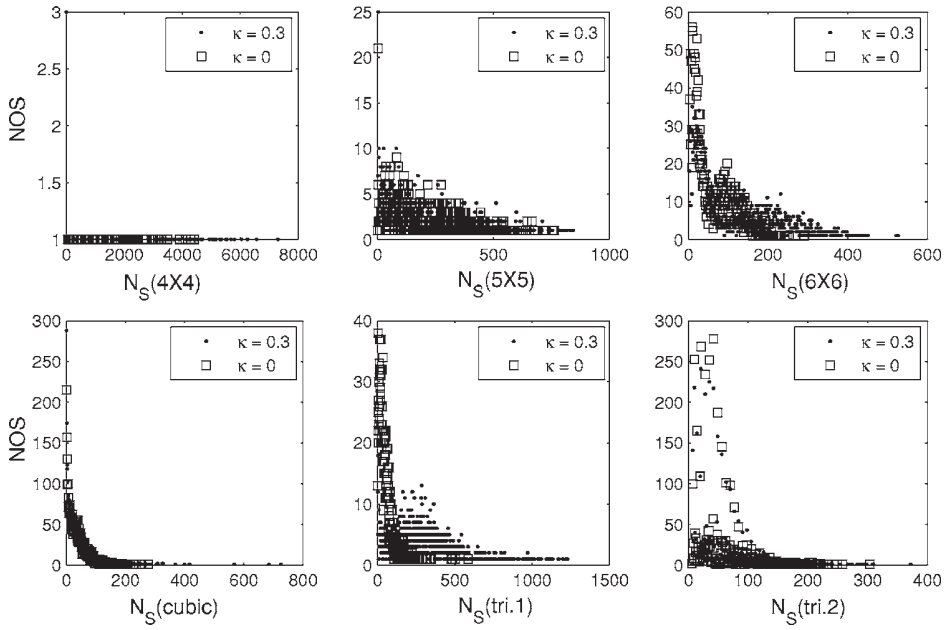

FIG. 2. Histograms of designability N_S for the six kinds of lattices based on the detailed *HP* model when $\kappa=0$ and 0.3. NOS for the y axis represents the number of structure.

TABLE VI. The designability in different lattice types, alphabet sizes, and κ values. The values in the $\kappa=0.3$ column are larger than that in the $\kappa=0$ column given the same alphabet size, which shows that the local interactions reduce degeneracy and make the designability higher. See text for the meaning of the LN and NDR.

		HP		*HNUP*		MJ	
Lattice type	Quantity	$\kappa=0$	$\kappa=0.3$	$\kappa=0$	$\kappa=0.3$	$\kappa=0$	$\kappa=0.3$
4×4	LN	2168	4734	4414	7294	5026	6618
	NDR	3.31	7.22	8.83	14.59	10.05	13.24
5×5	LN	613	785	762	845	514	571
	NDR	1.23	1.57	1.52	1.69	1.03	1.14
6×6	LN	445	1112	291	527	494	785
	NDR	0.89	2.22	0.58	1.05	0.99	1.57
Cubic	LN	303	390	277	725	396	442
	NDR	0.61	0.78	0.55	1.45	0.79	0.88
Tri.1	LN	104	207	586	1230	578	1029
	NDR	0.63	1.26	1.17	2.46	1.16	2.06
Tri.2	LN	242	363	304	372	153	173
	NDR	0.48	0.73	0.61	0.74	0.31	0.35

do not necessarily have the same structure S as their common native state in the primordial age. But other structures are not as stable as S and can be easily replaced by this structure S according to the principle of survival of the fittest. As a result, the structure S is highly designable and thermodynamically stable at the same time. The results in the other situations are listed in Table VII. From Table VII, we can see that the correlation coefficients between designability and stability are clearly larger than zero in most situations. For two variables, the larger their correlation coefficient is the better the linear relationship between them is.[32,33] So we can conclude that the designability and the stability have good linear relationship.

Third, we calculate the partnum using Eq. (10) independently (i.e., without considering the interaction detail) and compare it with the designability. The designability against the partnum in six of the cases (20-letter model, $\kappa=0$, all lattices) is shown in Fig. 4. From the figures we can see that generally the larger the designability is, the larger the partnum is. Notice that the correlation coefficients between the designability and the partnum in the square and cubic lattices are all larger than 0.4 in the *HP* model when $\kappa=0$, which can be seen from Table VII. This is just the situation which has been discussed in Ref. 5 and our results are in line with that. After getting similar results, Wang and Yu[5] constructed one hierarchical tree to demonstrate that the path of a walk meeting with fewer branches has larger partnum. They assumed that there existed a random process which selected out only the structures with the largest partnum. However, when the alphabet size increases, the correlation coefficients decrease.

When we consider the local interactions (i.e., $\kappa=0.3$), the correlation coefficients between the designability and the partnum change. We give a possible reason to explain such a phenomenon as the following. Considering the local interactions means that the information contained in the designability of structure S increases. Because in the calculation of the

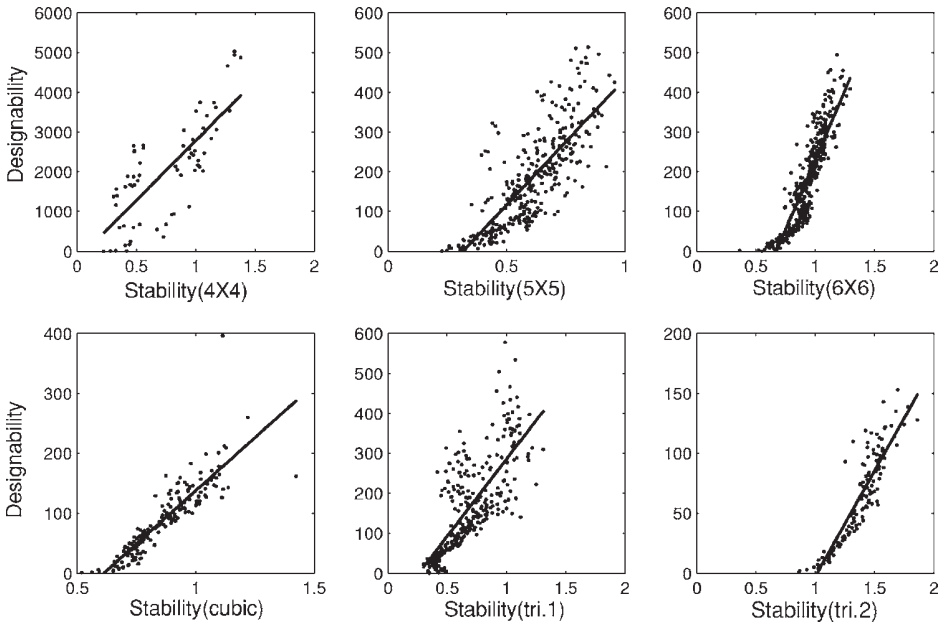

FIG. 3. Plots of designability N_S against the stability for the six kinds of lattices based on the 20-letter model when $\kappa=0$. The line is the fitting result. The correlation coefficients are all listed in Table VII.

TABLE VII. The correlation coefficients in different situations. The C_{NS}, C_{NP}, and C_{NF} represent the correlation coefficients between the designability and the stability, the partnum, and the foldability, respectively.

		$\kappa=0$						$\kappa=0.3$					
		4×4	5×5	6×6	Cubic	Tri.1	Tri.2	4×4	5×5	6×6	Cubic	Tri.1	Tri.2
HP	C_{NS}	0.78	0.60	0.56	0.66	−0.14	0.43	0.63	0.83	0.66	0.84	0.22	0.38
	C_{NP}	0.52	0.54	0.40	0.45	0.72	−0.18	0.19	0.16	−0.08	0.49	0.35	0.08
	C_{NF}	0.79	0.45	0.59	0.64	−0.70	0.54	0.68	0.34	0.69	0.67	−0.53	0.33
HNUP	C_{NS}	0.84	0.79	0.82	0.86	0.68	0.54	0.88	0.80	0.83	0.79	0.89	0.57
	C_{NP}	0.45	0.37	0.33	0.45	0.50	0.21	−0.01	0.24	0.09	0.36	−0.10	0.21
	C_{NF}	0.86	0.57	0.84	0.75	0.53	0.66	0.87	0.50	0.84	0.71	0.76	0.56
MJ	C_{NS}	0.78	0.79	0.89	0.88	0.75	0.90	0.80	0.78	0.91	0.92	0.85	0.78
	C_{NP}	0.23	0.38	0.21	0.25	0.36	0.21	−0.02	0.29	0.07	0.25	−0.03	0.36
	C_{NF}	0.90	0.67	0.87	0.20	0.76	0.91	0.86	0.61	0.89	0.19	0.76	0.80

partnum, we do not need to consider the energy, the partnum is independent of the energy. Hence the information in the partnum is invariable. This leads to the nonsymmetry of information contained in the two quantities from the information theory[41] aspect. As a result, the correlation coefficients between them will change when the local interactions are considered. From Table VII, we can see that the correlation coefficients between the designability and the partnum are smaller than that between the designability and the stability in most cases (except for the two cases: 2+3+4+3+2 triangular lattice, *HP* model, $\kappa=0,0.3$).

The designability against the foldability in six of the situations (detailed *HP* model, $\kappa=0.3$, all lattices) is shown in Fig. 5. From the figures we can see that the larger the designability is, the larger the foldability is as a whole. The physical meaning of the foldability of structure S is to measure how fast the protein sequence will fold into that target structure S. Therefore, this phenomenon can also be explained similarly as that in the stability. The correlation coefficients between the designability and the foldability in most cases are larger than 0.5, as shown in Table VII. Noting that the corresponding correlation coefficients in the 2+3+4+3+2 triangular lattice ($\kappa=0,0.3$) are negative (−0.70 and −0.53) in the *HP* model, it might indicate that the *HP* model is too simple for protein structure in triangular lattice when the chain length is too short.

From Table VII, we can also see that the correlation coefficients for the triangle lattices are not stable with the alphabet size like the square and cubic lattices; this may possibly be caused by their different bending angles. After the comparison of Figs. 3–5, we find that the linear fit between the designability and partnum is much worse than that between the designability and the other two quantities (i.e., the stability and foldability). It is easy to be understood because in the definitions of the designability, stability, and foldability we considered the same energy. But in the definition of the partnum, we do not consider any energy. So the partnum characterizes different aspects of the lattice model compared with the designability, stability, and foldability. This may also be the reason which causes the correlation coefficients C_{NP} between the designability and partnum to be unstable for different lattice types and alphabet sizes in Table VII.

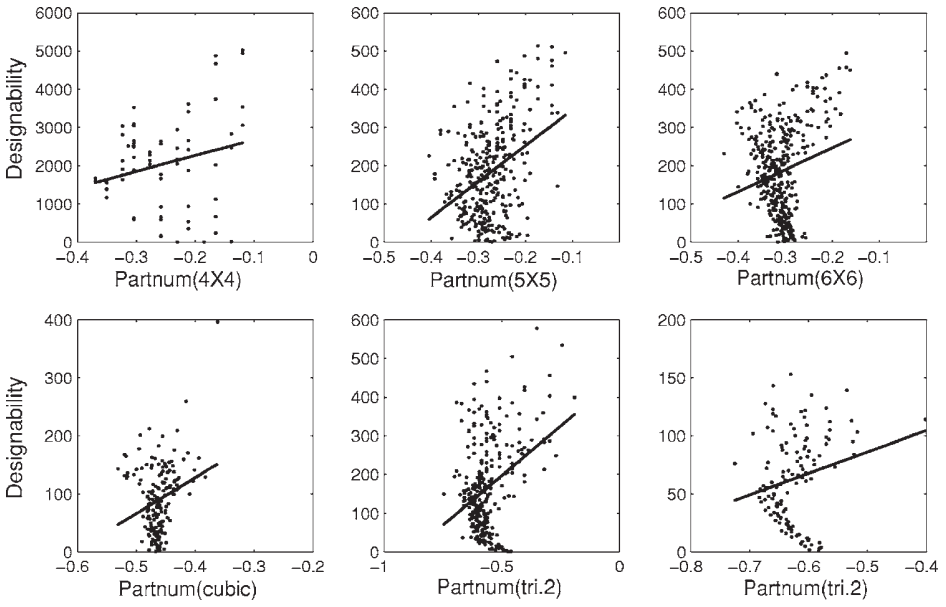

FIG. 4. Plots of designability N_S against the partnum for the six kinds of lattices for the 20-letter model when $\kappa=0$. The line is the fitting result. The correlation coefficients are all listed in Table VII.

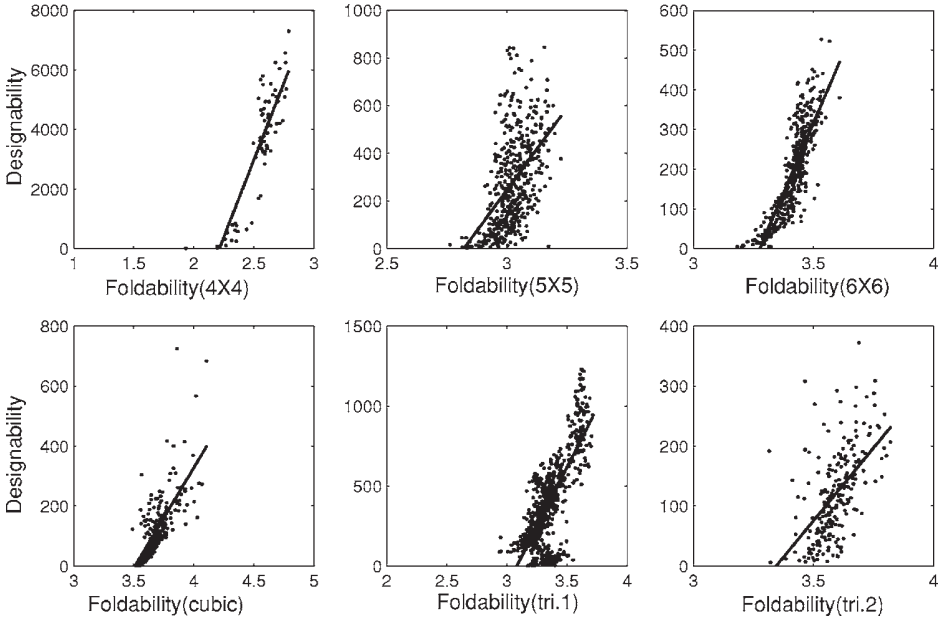

FIG. 5. Plots of designability N_S against the foldability for the six kinds of lattices for the detailed *HP* model when $\kappa=0.3$. The line is the fitting result. The correlation coefficients are all listed in Table VII.

Fourth, we consider more values of κ in Eq. (1) to see the effect of the strength of the local interaction on the designability and the relationship between it and the stability, partnum, and foldability. But because of the limitation of computing power, we only consider here the 5×5 square lattices based on the detailed *HP* model. We discuss 23 values of κ ranging from −2.75 to 2.75. The numbers of sampling structures and protein sequences are the same as that we have used in the previous corresponding calculating (i.e., 1081 structures and 50 000 sequences for each structure). In the sampling of sequence space, we use the CBS methods as above.

The two quantities used in Table VI (*LN* and *NDR*) are plotted against κ in Fig. 6. From the figures, we can see that the above two quantities all attain the smallest values when $k=0$. Therefore, we can conclude that the local interactions make the designability higher and make the degeneracy ratio of protein sequences lower. Similar results have also been reported in Ref. 14. In general the larger the absolute values of κ are (corresponding to a stronger local interaction), the more obvious such effects are, which can be seen from Fig. 6.

The correlation coefficients between the designability and the stability (C_{DS}), partnum (C_{DP}), and foldability (C_{DF}) for different values of κ are shown in Fig. 7. The values of C_{DS} are all larger than 0.5, which shows that the designability and the stability have good linear relationship. Except for one point with $\kappa=0.25$, the values of C_{DF} are also larger than 0.5, which shows that the designability and the foldability also have good linear relationship. But the values of C_{DF} are smaller than the values of C_{DS} in most cases showing that the linear relationship between the designability and the foldability is not as good as that between the designability and the stability. In addition, from Fig. 7 we can see that the correlation coefficients between the designability and the partnum (C_{DP}) are smaller than that between the designability and the stability (C_{DS}) and foldability (C_{DF}). This is in line with the results we have got above. Based on Fig. 7, we can conclude that the stability is the best predictor of the designability, the foldability is the second best, and the partnum is the third.

Although these discussions of local interactions are constrained to the 5×5 square lattice and the detailed *HP* model because of the limitation of computing power, such discussions can be done in other situations and similar results may hold.

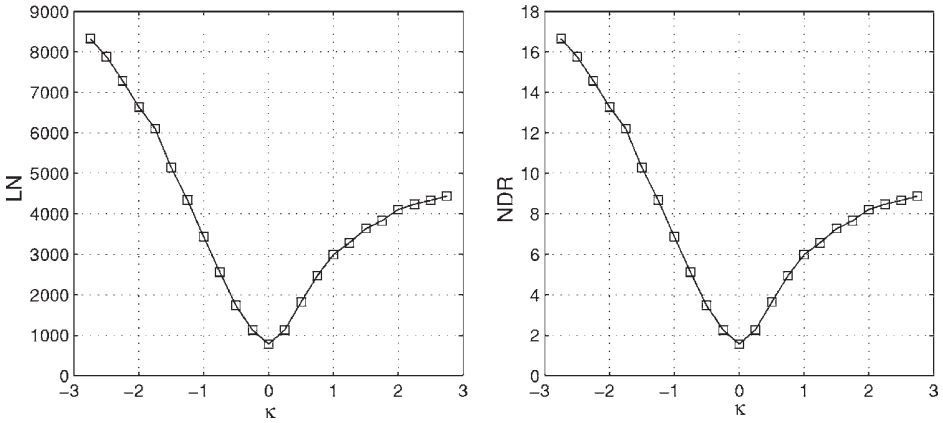

FIG. 6. The effect of the strength of local interactions (κ) to the two quantities (LN and NDR) used in Table VI for the 5×5 square lattice based on the detailed *HP* model. It suggests that the local interactions make the designability higher and reduces degeneracy.

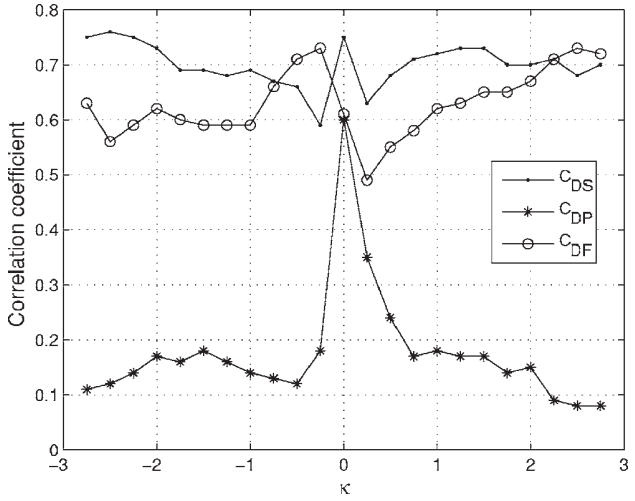

FIG. 7. The effect of the strength of local interactions (κ) to the correlation coefficients between the designability and the stability (C_{DS}), partnum (C_{DP}), and foldability (C_{DF}) for the 5×5 square lattice based on the detailed *HP* model. The correlation coefficients are sensitive to the value of $\kappa=0$. It also suggests that the stability and the foldability are better than the partnum for measuring the designability.

Now, let us give a short summary for the above calculation process. First, we enumerate all the compact structures unrelated by symmetry in certain lattice type. The total number of structures is denoted by T. Then after the alphabet size and the strength of local interaction (i.e., the value of κ) are fixed, we can calculate the four quantities according to the following steps. Step 1: sample n_1 structures from the T structures; step 2: select one target structure S from the n_1 structures and calculate its partnum; step 3: sample (design) n_2 sequences for the target structure S; step 4: calculate the three quantities (designability, stability, and foldability) of structure S with the n_2 designed sequences; step 5: repeat from step 2 to step 4 until the four quantities for all the n_1 structures are calculated.

Lastly, we discuss the effect of the sampling method on the results. That is to say, we will sample (design) sequences for each target structure with other methods (i.e., step 3 above).

The above calculations are all based on the common biased sampling of sequence space. However, in the lattice model of protein structure, a usual sampling method of sequence space is RS.[3,4,12,14] As a result, a question one may ask is the following: If we use other sampling methods, what will the results be like? In order to make a comparison, we also use the RS method in Refs. 3 and 14. We want to compare the corresponding results calculated from such sampling method with those we have got above with the CBS method. For the four cases (4×4 lattice, *HP* model, $\kappa=0,0.3$ and 2 +3+4+3+2 triangular lattice, *HP* model, $\kappa=0,0.3$), the two methods are the same as we can enumerate all the sequences completely considering the sequence space is not too large. As a result, their corresponding results are identical with each other. One of the comparisons is shown in Fig. 8. From Fig. 8, we can see that the designability calculated with the CBS method is much higher than that with the RS method. The corresponding results in all the 36 cases considered above are listed in Tables VIII and IX. Comparing Table VI with Table VIII, we find that the designability in all the considered cases calculated with the CBS method is higher than that with the RS method, and the CBS method reduces degeneracy. From the Tables VII and IX, we see that correlation coefficients calculated with the two methods have some

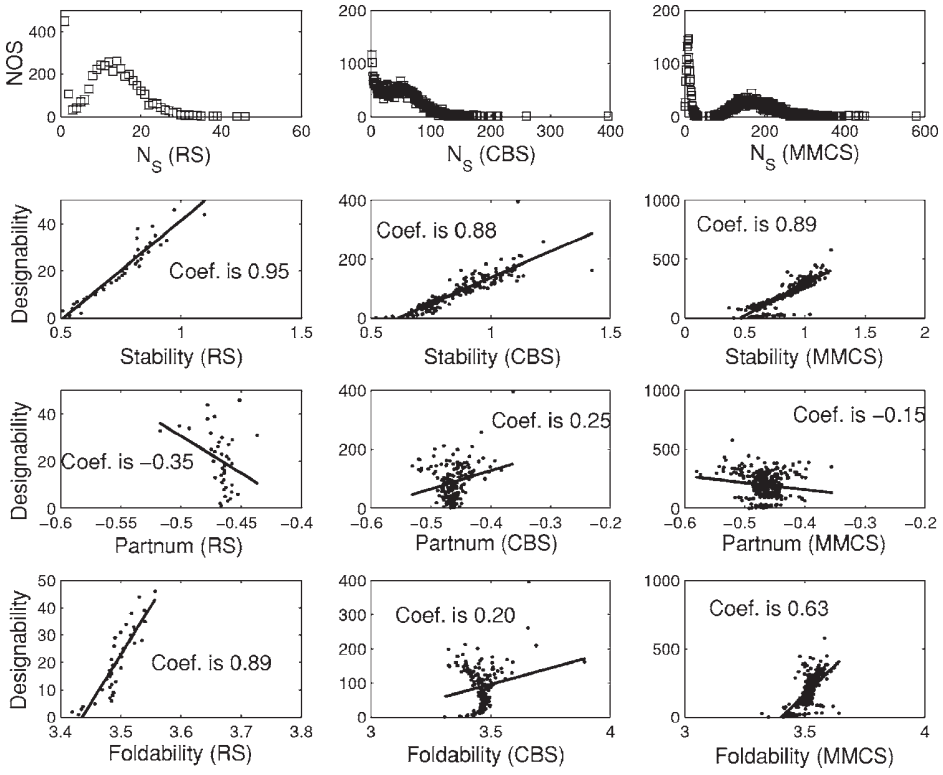

FIG. 8. Comparisons between the three different sampling methods of the sequence space for the case ($3\times3\times3$ cubic lattice, $\kappa=0$, 20 letter). Coef. in the figures and NOS for the y axis represents the correlation coefficient between those two corresponding quantities and the number of structure, respectively. Compared with the RS method, the CBS and the MMCS methods both make the designability higher. The correlation coefficients seem to be sensitive to the sampling method. But for the stability, the three correlation coefficients are all larger than 0.85, which suggests that the stability is the best predictor of the designability.

TABLE VIII. The designability in different lattice types, alphabet sizes, and κ values calculated with the RS method. Similar conclusions can be achieved from the table as that from Table VI. Compared with Table VI, it shows that the CBS method makes the designability higher and reduces the degeneracy.

		HP		*HNUP*		MJ	
Lattice type	Quantity	κ=0	κ=0.3	κ=0	κ=0.3	κ=0	κ=0.3
4×4	LN	2168	4734	1727	2965	1886	2258
	NDR	3.31	7.22	3.45	5.93	3.77	4.52
5×5	LN	202	247	208	246	216	243
	NDR	0.40	0.49	0.42	0.49	0.43	0.49
6×6	LN	103	176	105	149	101	133
	NDR	0.21	0.35	0.21	0.30	0.20	0.27
Cubic	LN	30	68	43	48	46	46
	NDR	0.06	0.14	0.09	0.10	0.09	0.09
Tri.1	LN	104	207	304	897	195	447
	NDR	0.63	1.26	0.61	1.79	0.39	0.89
Tri.2	LN	66	143	102	162	81	114
	NDR	0.13	0.29	0.20	0.32	0.16	0.23

changes. But as a whole, the RS method also suggests that the four quantities (designability, stability, foldability, and partnum) have similar relationships as those discussed for the CBS method above.

In order to reproduce biased samplings for the design of sequences for certain target structure, we can also consider using the Metropolis Monte Carlo sampling[35] (MMCS) method into the search of sequence space which has been used in Refs. 34, 36, and 37. Such method can be described as following: (step 1) fix temperature T_{sel} and amino acid composition; (step 2) generate a random sequence; (step 3) calculate the energy of the sequence E_{old} when it folds into the target structure; (step 4) random switch between two compositions along the amino acid sequence to generate a new sequence (such switch can keep the amino acid composition unchanged); (step 5) calculate the energy of the new sequence E_{new} when it folds into the target structure; (step 6): compare the two energies (E_{old} and E_{new}). If $E_{\mathrm{new}}<E_{\mathrm{old}}$, the switch is accepted and the sequence becomes the new one; otherwise, the Metropolis criterion[35] is used to decide whether the switch is accepted or rejected. (step 7): Repeat from step 3 to step 6 until the number of Monte Carlo steps is arrived. For a fixed temperature and amino acid composition (i.e., step 1), we can generate many different sequences to start (i.e., step 2). The Metropolis criterion is the following: a switch is accepted if $e^{-(E_{\mathrm{new}}-E_{\mathrm{old}})/T_{\mathrm{des}}}$ is larger than a random number uniformly distributed between 0 and 1.

But after trying with such a method, we found that it was time consuming for the computer to fulfill the task of the sequence sampling. Therefore, we only gave out the corresponding results calculated by the MMCS method in one of the cases (20-letter, $3\times3\times3$ cubic lattice, and κ=0). In the sampling process, the temperature we select lies between 0.1 and 1.2 and the number of Monte Carlo steps is 10^5. For a fixed temperature and amino acid composition, the number of different random generated sequences is 10^2. The results are given in Fig. 8 for comparison with the above two different sampling methods. It also shows that the MMCS method makes the designability higher compared with the RS method.

IV. CONCLUSIONS

Let us now summarize our answers to the three questions raised in the Introduction. (i) On introducing the definitions of the designability, stability, and partnum and defin-

TABLE IX. The correlation coefficients in different situations calculated based on random sampling of sequence space. The meanings of C_{NS}, C_{NP}, and C_{NF} are the same as those in Table VII.

		κ=0						κ=0.3					
		4×4	5×5	6×6	Cubic	Tri.1	Tri.2	4×4	5×5	6×6	Cubic	Tri.1	Tri.2
HP	C_{NS}	0.78	0.55	0.76	0.41	−0.14	0.56	0.63	0.86	0.81	0.51	0.22	0.39
	C_{NP}	0.52	0.66	0.54	0.74	0.72	−0.17	0.19	0.25	0.07	−0.08	0.35	0.39
	C_{NF}	0.79	0.40	0.85	0.13	−0.70	0.73	0.68	0.40	0.81	−0.01	−0.53	0.17
HNUP	C_{NS}	0.84	0.80	0.84	0.93	0.46	0.56	0.84	0.79	0.91	0.85	0.82	0.68
	C_{NP}	0.44	0.59	0.48	−0.33	0.65	0.40	−0.02	0.35	0.35	0.02	0.05	0.32
	C_{NF}	0.82	0.56	0.84	0.78	0.26	0.72	0.86	0.42	0.89	0.40	0.66	0.76
MJ	C_{NS}	0.79	0.80	0.89	0.95	0.73	0.79	0.75	0.79	0.90	0.90	0.82	0.69
	C_{NP}	0.33	0.56	0.62	−0.35	0.53	0.40	0.02	0.40	0.39	−0.25	0.07	0.65
	C_{NF}	0.84	0.70	0.89	0.89	0.65	0.81	0.84	0.58	0.89	0.81	0.83	0.58

ing the foldability of protein structure, we calculate these four quantities based on random samplings of structures and common biased sampling of protein sequence space because of the limitation of computing power in 36 different situations. Whatever the type of the lattice, the alphabet size and energy function are used, there will be an emergence of highly designable (preferred) structure. Such similar result has been achieved by Li *et al.*[3] based on the *HP* model and it is enhanced by our conclusions as ours are based on more different lattice types, alphabet sizes, and energy functions. The local interactions reduce degeneracy and make the designability higher.[11] (ii) There are strong correlation relationships between the designability and the two quantities of stability and foldability. This may be the result of competition: nature will select the highly stable and foldable structure and other structures will be eliminated during the evolution contributing to the emergence of preferred (i.e., highly designable) structure.[3,5] This seems to be the application of Darwin's evolution theory in the selection of protein tertiary structure. Compared with this, the correlation relationship between the designability and the partnum is not so strong because the partnum is independent of the energy. (iii) The designability is sensitive to the lattice type and the alphabet size and the energy function (i.e., considering the local interaction or not). Consideration of the local interaction energy makes the designability higher and the corresponding correlation coefficient change. The correlation coefficients between the four quantities suggest that the stability is the best predictor for designability, the foldability stands in the second, and the partnum is relatively worse compared with the first two quantities. But for the square lattices, the partnum is also a good candidate in measuring the designability based on the *HP* model when we do not consider the local interactions, which is just the situation considered in Ref. 5 and our conclusion is in line with that in it.

Comparisons between the results calculated from the two different sampling methods of sequence space (CBS and RS) suggest that the CBS make the designability higher and reduces degeneracy. The correlation coefficients between the four quantities are sensitive to the sampling methods. But as a whole, similar conclusions can be made from the two different sampling methods. The MMCS method also makes the designability higher in the considered case (20-letter, $3\times3\times3$ cubic lattice, and $\kappa=0$).

These conclusions seem to be useful when we put the designability principle into practical use (such as to predict the protein tertiary structure and protein design), as we can consider the three quantities of stability, foldability, and partnum together to get much more meaningful information.

ACKNOWLEDGMENTS

The first author would like to thank Changsha University of Technology for providing facility in the process of revising the manuscript. Financial support was provided by the Chinese National Natural Science Foundation (Grant No. 30570426), Fok Ying Tung Education Foundation (Grant No. 101004) and the Youth Foundation of Educational Department of Hunan Province in China (Grant No. 05B007) (Z.-G. Yu), and the Australian Research Council (Grant No. DP0559807) (V. V. Anh).

[1] *Protein Folding*, edited by T. E. Creighton (Freeman, New York, 1992).
[2] C. Anfinsen, Science **181**, 223 (1973).
[3] H. Li, R. Helling, C. Tang, and N. Wingreen, Science **273**, 666 (1996).
[4] H. Li, C. Tang, and N. S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **95**, 4987 (1998).
[5] B. Wang and Z. G. Yu, J. Chem. Phys. **112**, 6084 (2000).
[6] R. Mélin, H. Li, N. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).
[7] S. Govindarajan and R. A. Goldstein, Biopolymers **36**, 43 (1995).
[8] S. Govindarajan and R. A. Goldstein, Proc. Natl. Acad. Sci. U.S.A. **93**, 3341 (1996).
[9] T. Roy, J. Math. Phys. **42**, 4283 (2001).
[10] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolyness, Proc. Natl. Acad. Sci. U.S.A. **89**, 9029 (1992).
[11] N. E. G. Buchler and R. A. Goldstein, Proteins: Struct., Funct., Genet. **34**, 113 (1999).
[12] H. Li, C. Tang, and N. S. Wingreen, Proteins: Struct., Funct., Genet. **49**, 403 (2002).
[13] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[14] A. Irbäck and E. Sandelin, J. Chem. Phys. **108**, 2245 (1998).
[15] V. S. Pande, C. Joerg, A. Y. Grosberg, and T. Tanaka, J. Phys. A **27**, 6231 (1994).
[16] T. A. Brown, *Genetics*, 3rd ed. (Chapman and Hall, London, 1998).
[17] K. A. Dill, Biochemistry **24**, 1501 (1985).
[18] J. Wang and W. Wang, Phys. Rev. E **61**, 6981 (2000).
[19] Z. G. Yu, V. Anh, and K. S. Lau, Physica A **337**, 171 (1998).
[20] Z. G. Yu, V. Anh, and K. S. Lau, J. Theor. Biol. **226**, 341 (2004).
[21] Z. G. Yu, V. Anh, K. S. Lau, and L. Q. Zhou, Phys. Rev. E **73**, 031920 (2006).
[22] E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[23] A. Kloczkowski, T. Z. Sen, and R. L. Jernigan, Polymer **45**, 707 (2004).
[24] R. Tatsumi and G. Chikenji, Phys. Rev. E **60**, 4696 (1999).
[25] A. Irbäck, C. Peterson, F. Potthast, and O. Sommelius, J. Chem. Phys. **107**, 273 (1997).
[26] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[27] J. D. Bryngelson and P. G. Wolynes, Biopolymers **30**, 171 (1990).
[28] B. Derrida, Phys. Rev. Lett. **45**, 79 (1980).
[29] J. U. Bowie, R. Liithy, and D. Eisenberg, Science **253**, 164 (1991).
[30] A. Kloczkowski and R. L. Jernigan, Comput. Theor. Polym. Sci. **7**, 163 (1997).
[31] X. M. Li and N. C. Wang, Genomics, Peoteomics Bioinf. **2**, 245 (2004).
[32] W. Mendenhall, R. J. Beaver, and B. M. Beaver, *Introduction to Probability and Statistics* (Machine Press, Beijing, 2004).
[33] P. E. Pfeiffer, *Probability for Applications* (Springer-Verlag, New York, 1989).
[34] J. L. England and E. I. Shakhnovich, Phys. Rev. Lett. **90**, 218101 (2003).
[35] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller, J. Chem. Phys. **21**, 1087 (1953).
[36] E. I. Shakhnovich and A. M. Gutin, Protein Eng. **6**, 793 (1993).
[37] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **60**, 7195 (1993).
[38] E. I. Shakhnovich, Folding Des. **3**, R45 (1998).
[39] E. I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994).
[40] E. I. Shakhnovich and A. M. Gutin, Nature (London) **346**, 773 (1990).
[41] M. J. Usher, *Information Theory for Information Technologists* (Macmillan, London, 1984).