Medium effects on the selection of sequences folding into stable proteins in a simple model

You-Quan Li,[1] Yong-Yun Ji,[1] Jun-Wen Mao,[1,2] and Xiao-Wei Tang[1]

[1]*Department of Physics, Zhejiang University, Hangzhou 310027, People's Republic of China.*

[2]*Department of Physics, Huzhou Teachers College, Huzhou 313000, People's Republic of China*

(Received 6 December 2004; published 16 August 2005)

We study the medium effects on the selection of sequences in protein folding by taking into account surface potential in hydrophobic-polar model. Our numerical calculation demonstrates that the surface potential enhances the average gap for the highly designable structures. It also shows that the most stable structure may be no longer the most stable one if the medium is changed.

DOI: 10.1103/PhysRevE.72.021904 PACS number(s): 87.10.+e, 87.14.Ee, 87.15.−v

Protein folding has been one of the long-standing problems in biology. Composed of a specific sequence of amino acids, each protein folds into a unique structure as its native state. It is believed that the native structure is the global minimum of free energy [1] for most single-domain proteins. The one-dimensional amino acid sequences encode sufficient information to determine the three-dimensional conformations that play an important role in the biological function of proteins. There has been great interest in the study of protein folding by molecular dynamical simulation [2] and lattice model [3–5].

From the coarse-grained point of view, the 20 amino acids can be classified [6] as hydrophobic (H) and polar (P) groups according to their contact interaction. Thus, a so-called *HP* lattice model whose structures are defined on a lattice and whose sequences take only two kinds of amino acids (either "*H*" or "*P*") was presented [4]. In terms of the three-dimensional *HP* model [5], Li *et al.* proposed [7] a "designability principle" to interpret nature's selection mechanism for protein structure, where the designability is defined as the number of sequences possessing the structure as the unique lowest energy state. They found that structures differ drastically in the designability. A small number of structures possess high designability and large energy gaps with more thermodynamic and mutational stability [7,8]. Although this simple model is still far from applications to real proteins, it does provide a qualitative understanding of the essential points of real proteins on the basis of the current computers. Studies on the designability for other lattice models [9] and for off-lattice models [10] presented similar results. For many-letter models, the different parameters gave different results: Buchler *et al.* [11] reported that the designability of the structure depends sensitively on the size of the alphabet, and Li *et al.* [12] found that the designability of the structure is not sensitive to the alphabet size when a realistic interaction potential (MJ matrix) is employed. Ejtehadi *et al.* found that if the strength of the nonadditive part of the interaction potential becomes larger than a critical value, the degree of structure designability will depend on the parameters of the potential [13].

Since useful features concerning protein folding and their stability can be explored on the basis of the lattice model, it will be worthwhile to study effects of the medium. In this paper, we consider effects of the medium on structure stability by introducing different parameters characterizing various concentrations of medium solution. Our results give some answers to the following questions: Are those sequences associated with highly designable structures universally good? How do they vary depending on media [14] where the protein is placed?

We investigate the effects of media upon the category of highly designable protein sequences, which will undoubtedly provide a clue to understand the variations in the natural selection of protein species caused by media where the protein lives. For this purpose, we must reconstruct the original *HP* model by introducing potential parameters to the monomers at protein's surface. As protein is figured as a chain of beads occupying the sites of a lattice in a self-avoiding way, the energy of a sequence folded into a particular structure in our model is given by

$$H = \sum_{i<j} E_{\sigma_i\sigma_j}\delta_{|r_i-r_j|,1}(1-\delta_{|i-j|,1}) + \sum_{r_j\in S} U_{r_j}\delta_{\sigma_j,P}, \qquad (1)$$

where i, j denote for the successive labels of monomers in a sequence, r_i for the position (of the ith monomer) on lattice sites, and σ_i refers to H or P corresponding to the hydrophobic or polar monomer. Here the Kronecker delta notation is adopted, i.e., $\delta_{a,b}=1$ if $a=b$, and $\delta_{a,b}=0$ if $a\neq b$. As the hydrophobic force [6] drives protein to fold into a compact shape with more hydrophobic monomers inside as possible, the *HH* contacts are more favorable in this model. This property can be characterized by choosing $E_{PP}=0$, $E_{HP}=-1$, and $E_{HH}=-2.3$, which were derived in Ref. [7] from their analysis of the real-protein data contained in the Miyazawa-Jernigan matrix for the inter-residuum contact energies between different types of amino acids. As the effects caused by the protein's surrounding medium are relevant to salt concentration [14] of a solution where the protein is placed, we introduce U_V, U_E, and U_F to represent the attractive potentials in the protein surface for polar (hydrophilic) monomers at vertices, edges, or face centers, respectively. These attractive forces arise from the interaction between the medium (solution) and the hydrophilic monomers. Since the lattice model is not appropriate for a spherical surface, we consider different weight coefficients at the surface, say $U_\tau=-\gamma_\tau V$, to imitate protein analogs. The coefficients γ_τ are the degree of burial of the monomer at surface. For example, it is 1/2 for a face contact, 3/4 for an edge contact, and 7/8 for a vertex

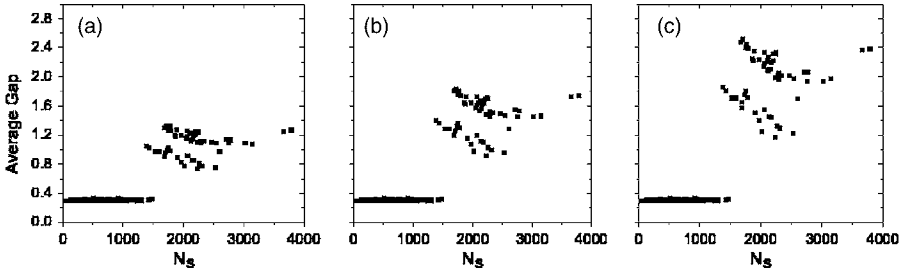

FIG. 1. Average gap of structures vs N_s of the structures in the case of γ_V=7/8, γ_E=6/8, γ_F=0 for (a) V=0.0001, (b) V=0.9, and (c) V=2.1, respectively.

contact. If $\gamma_V = \gamma_E = \gamma_F \neq 0$, no new results occur in comparison to the result that Li *et al.* had studied. This is because the core in the cubic of the 27-site model is always hydrophobic, which implies that the surface potentials merely cause a global shift in energy spectrum of the 27-site model if we impose equal weights on a vertex, edge, as well as the center of a face. We investigate several cases of nonvanishing γ_τ later on.

It has been noticed that some structures can be designed by a large number of sequences, while others can be designed by only few sequences. The designability of a structure is measured by the number (N_s) of sequences that take the given structure as their unique ground state, as was reported by Li *et al.* [7]. And the structures differ drastically according to their designability, i.e., highly designable structures emerge with a number of associated sequences much larger than the average ones. Additionally, the energy gap δ_s is the minimum energy for a particular sequence to change its ground-state structure into a different compact structure. The average energy gap $\bar{\delta}_s$ for a given structure is evaluated by averaging the gaps over all the N_s sequences that design that structure. The structures with large N_s have much larger average gap than those with small N_s, and there is an apparent jump around N_s=1400 in the average energy gap. This feature was noticed by Li *et al.* [7] in the medium-independent *HP* model, thus these highly designable structures are thermodynamically more stable.

Although the choices of E_{PP}=0, E_{HP}=−1, and E_{HH}=−2.3 adopted in Ref. [7] fulfil the principle that the major driving force for protein folding is the hydrophobic force, the difference between the contact energy for the monomers inside the protein and those at the surface was disregarded. Therefore, to explore the designability affected by the medium surrounding the protein, the application of surface potential in our model becomes inevitable. We have pointed out in the above that the 26 monomers are on the surface for 27-site model, which gives trivial results for uniform weights to the surface potential. On the other hand, increasing the number of the lattice sites will make the model beyond the calculation capacity of current computers. However, for the other parametrization of surface potential, we are able to obtain nontrivial and interesting results. First, we consider a "cubic-shape approximation" by different potential weights: γ_V=7/8, γ_E=6/8, and γ_F=4/8, which represent the degree of burial of the monomers at the surface. For this parameter choice, we find there are 17 more sequences possessing unique ground states, regardless of the magnitudes of V (ranging from 0.1 to 2.1), while they do not possess unique ground states in the model studied by Li *et al.*, where the effect of the medium was neglected [7]. Our calculation further exposes that 14 of those 17 sequences mainly belong to the highly designable structures, and have relatively larger energy gaps. We analyze all 17 sequences, and find that 14 can be related to each other by a single mutation, which implies that they belong to the "neutral island" suggested by Trinquier *et al.* [15]. These results confirm that protein structures are selected in nature because they are readily designed and stable against mutations, and that such a selection simultaneously leads to thermodynamic stability and foldability. Thus, a key point to understand the protein-folding problem is to understand the emergence and the properties of highly designable structures.

The second parametrization is to consider γ_V=7/8, γ_E=6/8, and γ_F=0, which models a quasicubic-shaped protein with seven monomers inside and 20 ones at the surface. In this case, we find there are 48 more sequences possessing unique ground states for a wider range of magnitudes of V (from 0.0001 to 2.1), which, however, have no unique ground states in the case of Ref. [7]. Only one sequence designs the highly designable structure while the other 47 sequences design lowly designable structures. All the energy

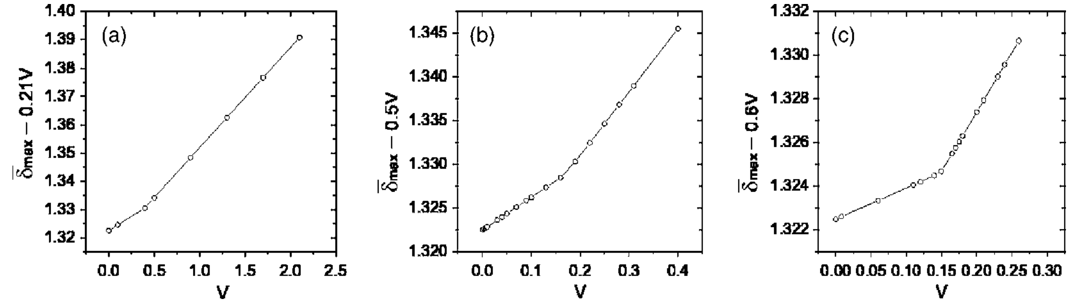

FIG. 2. The largest average gap $\bar{\delta}_{max}$ vs the parameter V: (a) for γ_V=7/8, γ_E=6/8, and γ_F=4/8; (b) for γ_V=7/8, γ_E=6/8, and γ_F=0; (c) for γ_V=1, γ_E=1, and γ_F=0.

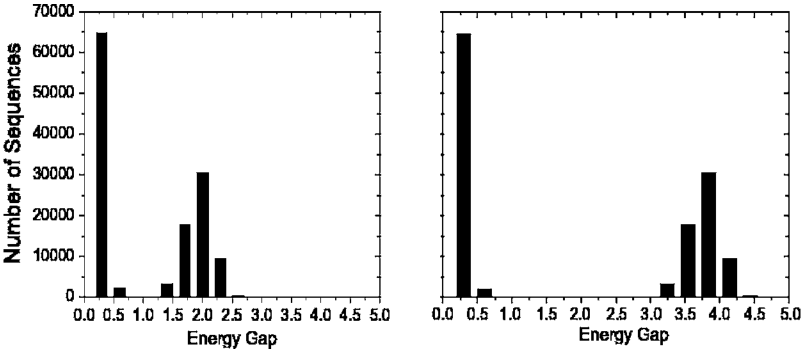

FIG. 3. The histogram for the number of sequences vs the energy gap for the 60 highly designable structures in the absence of medium (left panel) and in the presence of medium $\gamma_V=7/8$, $\gamma_E=6/8$, $\gamma_F=0$, $V=2.1$ (right panel).

gaps of those new sequences are found to be $V/8$. Since the ratio of the numbers of the monomers at the surface to that inside is of the order 1 in natural proteins [8], and the ratio in our model is 26:1 in the first case, but 20:7 in the second case, the latter case ought to be closer to the usual natural proteins. Figure 1 shows the average energy gap for different potential parameters. Clearly, the protein-medium interaction enhances the average gap of highly designable structures, which illustrates that the highly designable structures selected by nature are more stable in proper media than in "vacuum." Thus our theoretical results may evoke more attention to the dependence of stability on medium effects in further model studies.

We calculate the case by assuming the potentials at the vertices and at edges with the same weights, i.e., $\gamma_V=1$, $\gamma_E=1$, and $\gamma_F=0$. This is an analog of a ball shape with seven monomers inside and 20 monomers on the surface, which corresponds to a case when the protein conformation is not cube shaped. We find that there is no sequence beyond those of Ref. [7] to take the highly designable structures. Just like the result in Ref. [15], there are also 60 structures that possess large average gaps. When we take into account the effects of the medium, the average gap for highly designable structures increases apparently as the potential parameter increases, but the average gap of lowly designable structures does not change much. In all the aforementioned cases, the average gap of a single highly designable structure increases linearly with respect to the increase of potential V. Furthermore, we find the structure with largest average gap is not fixed for all potential parameters. Crossings between energy levels always take place when the potential parameter changes. It is therefore worthwhile to point out that the gains of stability for distinct structures vary, and the most stable protein structure in one surrounding medium maybe no more the most stable one in another medium. The plots of the largest energy gap versus the parameter V are shown in Fig. 2, respectively, for the three cases discussed above. In order to show an apparent change for eye's view, we have set the value of the vertical axis in Fig. 2 to be the largest average gap minus $0.21V$, $0.5V$, and $0.6V$ for the cases (a), (b), and (c), respectively. In each case is there a critical value of V across which the plot transits from a straight line to another straight line. The critical values of V differ in different cases, but the largest average gaps at the transition point take the same value $\bar{\delta}_s=1.4137$.

We analyze all the sequences that design the 60 highly designable structures, respectively. In the absence of medium, $\gamma_V=\gamma_E=\gamma_F=0$, the energy gaps δ_s of those sequences range from 0.3 to 2.6 (see Fig. 3). Almost half of them have small energy gaps (around 0.3). In the presence of medium, the energy gaps for most of the sequences with larger (over 1) gap rise as the parameter increases while those for the sequences with small gap does not rise apparently. For the cases (a) $\gamma_V=7/8$, $\gamma_E=6/8$, $\gamma_F=4/8$, (b) $\gamma_V=7/8$, $\gamma_E=6/8$, $\gamma_F=0$, and (c) $\gamma_V=\gamma_E=1$, $\gamma_F=0$, the increments in energy gaps are mainly $3V/8$, $7V/8$, and V, respectively. There is also a small portion of sequences whose energy gaps decrease in the medium, e.g., 276 sequences in the case $\gamma_V=7/8$, $\gamma_E=6/8$, $\gamma_F=4/8$. Considering some particular structures among the 60 highly designable ones, we analyze the sequences that design them. The energy gap of the structure with the larger gap will mostly increase when the sequence is placed in the medium, which leads to the linear increment of the average gap. Our results also illustrate that the distribution shapes emerge similarly for those three structures. In addition, the total number of sequence in (b) is less than in (c), but there are many more sequences possessing large energy gaps in (b) than in (c).

We introduce a distribution function versus the number of polar monomers,

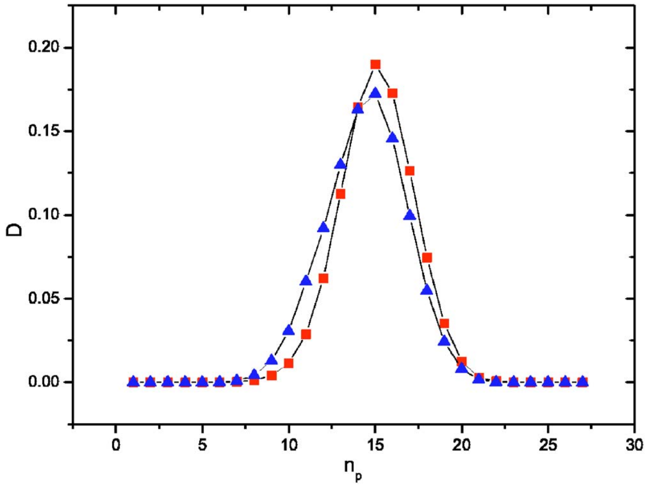

FIG. 4. (Color online) Comparison of the average distributions $\bar{D}(n_P)$ versus the number of polar monomers n_P for the lowly designable structures (▲), and the highly designable structures (■), respectively.

$$D(n_P)=\frac{N_s(n_P)}{\sum_{n_P}N_s(n_P)},\qquad(2)$$

where $N_s(n_P)$ stands for the number of sequences that contain n_P polar monomers and design the given structure. Clearly, the designability of the structure is given by $N_s=\Sigma_{n_P}N_s(n_P)$. We calculate the average distribution $\bar{D}(n_P)$ for the highly designable structures and for the lowly designable ones, respectively, which are plotted in Fig. 4. Clearly, the distribution shape for highly designable structures shifts toward the larger number of polar monomers, while that for lowly designable structures shift toward the small ones. The more polar monomers there are, the lower their energy caused by the surface potential will be. This may provide an interpretation that the protein-medium interaction enhances the average gap more for the highly designable structures than for the lowly designable ones.

In summary, our model exhibits that the surface potential that represents protein-medium interaction enhances the average gap of highly designable structures, which implies that the highly designable structures selected by nature are more stable in proper media than in "vacuum." We obtained that the energy gap of the sequences with larger energy gap will mostly increase when the sequence is placed in a medium, which leads to the linear increment of the average gap. We noticed that there is a critical value for the parameter of the surface potential, which means that a most stable structure may be no longer the most stable one if the medium parameters changed. We analyzed the average distribution of the number of hydrophilic (polar) monomers that provides a qualitative interpretation of the medium effect we obtained. One should note that this is for a 27-site lattice model; the thermodynamic stabilities and designability may not be necessarily correlate for the system with large number of monomers [16]. Recently, the medium effect was noticed in Ref. [17], by molecular dynamic simulation, that the protein is stable at neutral pH solution while undergoes a conformation change at low pH. Since a lot of studies have shown that several properties of natural proteins can be captured by simple models, our discussion in above may motivate people to model the effect of medium on all theoretical studies where the medium potential was ignored.

This work is supported by NSF China Grant No. 10225419 and the Key project of Chinese Ministry of Education Grant No. 02046.

[1] C. Anfinsen, Science **181**, 223 (1973).
[2] T. Lazaridis and M. Karplus, Science **278**, 1928 (1997).
[3] H. Taketomi, Y. Ueda, and N. Go, Int. J. Protein Res. **7**, 445 (1975).
[4] K. A. Dill, Biochemistry **24**, 1501 (1985).
[5] M. E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[6] W. Kauzmann, Adv. Protein Chem. **14**, 1 (1959).
[7] H. Li, R. Helling, C. Tang, and N. S. Wingreen, Science **273**, 666 (1996).
[8] H. Li, C. Tang, and N. S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **95**, 4987 (1998).
[9] H. Cejtin, J. Edler, A. Gottlieb, R. Helling, H. Li, J. Philbin, N. Wingreen, and C. Tang, J. Chem. Phys. **116**, 352 (2002).
[10] J. Miller, C. Zeng, N. S. Wingreen, and C. Tang, Proteins **47**, 506 (2002).
[11] N. E. G. Buchler and R. A. Goldstein, Proteins **34**, 113 (1999).
[12] H. Li, C. Tang, and N. S. Wingreen, Proteins **49**, 403 (2002).
[13] M. R. Ejtehadi, N. Hamedani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad, J. Phys. A **31**, 6141 (1998).
[14] B. N. Dominy, D. Perl, F. X. Schmid, and C. L. Brooks, III, J. Mol. Biol. **319**, 541 (2002).
[15] G. Trinquier and Y. H. Sanejouand, Phys. Rev. E **59**, 942 (1999).
[16] H. J. Bussemaker, D. Thirumalai, and J. K. Bhattacharjee, Phys. Rev. Lett. **79**, 3530 (1997).
[17] D. O. V. Alonso, S. J. DeArmond, F. E. Cohen, and V. Daggett, Proc. Natl. Acad. Sci. U.S.A. **98**, 2985 (2001).