International Journal of Modern Physics B, Vol. 16, No. 4 (2002) 631–637

SEQUENCE DESIGN AND FOLDING DYNAMICS OF LATTICE PROTEIN-LIKE MODELS

MENG QIN, JUN WANG, TANPING LI and WEI WANG
National Laboratory of Solid State Microstructure, Nanjing University, Nanjing 210093, China

Received 3 December 2001

A sequence design method based on maximizing the thermodynamic occupying probability of the target structure is investigated. Some model-protein sequences are designed using the occupying-probability-maximized procedure on a $3\times3\times3$ lattice. The thermodynamic and dynamic features of these sequences show their great improvement comparing with those of the sequences designed by an energy-minimized method. A better foldability is achieved for the occupying-probability-maximized sequences. These results suggest that the native occupying probability rather than the energy would be a better judgment for protein-like models.

PACS number(s): 87.15.Cc, 87.15.Aa, 87.14.Ee

Protein folding and protein design are interesting and unsolved problems in the molecular biology. For protein folding, a given sequence can reach its unique native conformation (or structure) with the lowest free energy quickly without scanning all possible conformations with an astronomically large number.[1,2] Differently, protein design is to obtain a sequence (or sequences) which can fold itself into a target structure as its native one.[3] Although big progresses have been made in the study on both problems,[3–6] there are some basic aspects which still need to be clarified. For example, what is the best or most effective method for sequence design, and how good is the folding behavior of the designed sequences?

Previously, the design procedure was developed as follows. For a given target structure Γ, one first fixes the composition of the amino acids (residues), i.e. the numbers of various kinds of residues from 20 natural occurring kinds, in a sequence S, then the sequence is threaded onto the target structure Γ. By exchange the positions of different residues, one obtains many different sequences for the target structure Γ. Among these sequences, it was believed that there is a sequence S_N being the best sequence of the structure Γ when various residues are so properly arranged within the sequence that the energy (as a summation over all contacts) of the sequence S_N is the lowest one than all others. This means that the sequence with a lowest total energy of various contacts between all residue pairs in a lattice space is the native one relating to the target structure Γ. This is called as the energy

minimized method (EMM) for the sequence design.[3] It relates to a probability $P'_S(\Gamma, T) \sim \exp(-E_S(\Gamma)/T)$ where T is the temperature and $E_S(\Gamma)$ is the contact energy of the sequence S on structure Γ. Thus, the task of maximizing $P'_S(\Gamma, T)$ is actually reduced to the one of minimizing $E_S(\Gamma)$. However, it was argued that such a design method may result in some sequences which fold to a state with their energy lower than that of the target Γ.[7] In addition, a Z-score method (ZSM) was proposed for sequence design, which seems to be a good design scheme.[8] However, both the EMM and ZSM are based on the contact energy of the target conformations. The folding of protein actually depends on the whole landscape rather than on a single state only, though the native state takes an essential role in the folding processes. Some more consideration on the whole energy spectrum may be possibly enhance the properties of designed sequences.

Considering the stability and the dynamic behavior of the folding, in this work, we report a study on the sequence design based on the maximization of the occupying probability of the target Γ among all the conformations[9]

$$P_S(\Gamma, T) = \frac{e^{-E_S(\Gamma)/T}}{Z} \tag{1}$$

where $Z = \sum_{\Gamma'} \exp[-E_S(\Gamma')/T]$ is the partition function over all conformations Γ'. At some certain temperature T, maximizing the occupying probability of native structure equals to increasing the difference between native and denatured states and, thus, making the folding more cooperatively. An approximation for the partition function Z is only taking into account the compact conformations since they dominate the summation. Indeed, for the attracting contact potential like the Miyazawa and Jernigan (MJ) potential,[10] the compact structures take an important role in the low energy region of the energy spectrum. As a result, the summation could be a suitable approximation to the partition function of the system only over the compact conformations, at least for low temperature T. This approximation practically makes the computation feasible. This is the case, especially for the lattice models. For a cubic $3 \times 3 \times 3$ lattice, $103,346$ compact conformations are taken for Γ'.[11] Thus, the sequence design procedure is followed by the maximization of $P_S(\Gamma, T)$ for a given composition of residues and a target Γ. By exchanging randomly the positions of two residues in the sequence, i.e. the sequence changing from S to S', we have different values of $P_S(\Gamma, T)$. If an exchange gives a high value of $P_{S'}(\Gamma, T)$, we accept this exchange. Otherwise, we accept the exchange by a possibility $r = \exp[(P^{\text{new}}_{S'}(\Gamma, T) - P^{\text{old}}_{S}(\Gamma, T))/T^*)$. It consists of the optimization of Monte Carlo (MC) procedure. Here the design temperature T^* is an artificial temperature which should be chosen properly, and it decreases slowly as the exchange times increase. S' means a sequence after the residues being exchanged. Clearly, for different sequences, S or S', the values of the partition function Z in Eq. (1) are different. It is worthy to note that the EMM actually treats the partition function Z in Eq. (1) as a constant. During the residues exchange, obviously, taking Z as a constant or not may lead to different results. Let us show an example

to see the difference. In our simulations, we have seen many exchanges which give $E_{S'}^{\text{new}}(\Gamma) > E_S^{\text{old}}(\Gamma)$ and $P_{S'}^{\text{new}}(\Gamma, T) > P_S^{\text{old}}(\Gamma, T)$ at the same time. According to the EMM, these exchanges are basically abandoned (accepted with a Metropolis probability). Nevertheless, these exchanges are absolutely accepted for the occupying probability method (OPM). Moreover, for the EMM, the sequence may be trapped in some local minima, and could not escape easily. Physically, this may be due to sole consideration of the contact energy without the effects of conformations of the system.

Now let us present our method and results. The composition of the sequences is set as 14 hydrophobic residues (randomly selected from residues L, I, V, M, C, F, Y, W) and 13 polar residues (from R, S, T, Q, H, D, E, K, G, A, N, P).[12] Then these 27 residues are randomly assigned onto a chain. For a cubic lattice model, the contact energy of the chain with conformation Γ' is

$$E_S(\Gamma') = \sum_{i<j+3} U_{ij}\delta(r_{ij} - a) \tag{2}$$

where U_{ij} is the contact potential between residues i and j.[10] The function $\delta(r_{ij} - a)$ characterizes the contact between i and j residues with $\delta(0) = 1$ and 0 otherwise, and a is the lattice space. A compact structure with the highest designability in Ref. 13 is used as the native one [see Fig. 1(A) in Ref. 13]. 1000 sequences are optimized using both design methods, respectively. Several examples are listed in Table 1. From these sequences we see that the OPM can reach an energy, for example, as low as $E_{\text{Nat}} = -124.2$. Nevertheless, the EMM could not reach such a low energy even the number of the MC steps in the optimization program goes to 10^8.

In Fig. 1, we plot the occupying probability $P_S(\Gamma, T)$ versus the temperature T. We can see that the sequences designed by the OPM have basically the same behavior. Three curves for three different sequences show very small difference. As the temperature T decreases, the values of $P_S(\Gamma, T)$ increase rapidly, and undergoes

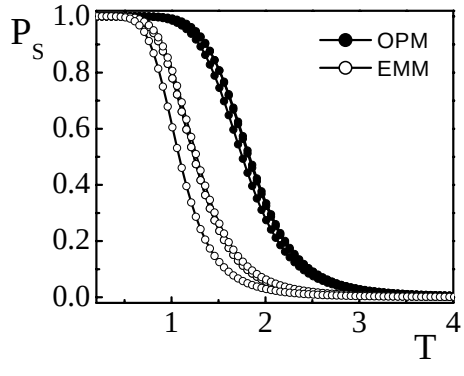

Fig. 1. The occupying probability $P_S(\Gamma, T)$ over the compact conformations versus the temperature T for three sequences for both design methods.

Table 1. The values of T_θ, T_f and $\sigma = |T_\theta - T_f|/T_\theta$ and several sequences for both design methods. $E_{\rm Nat}$ is the native state energy.

	OPM			EMM		
	S_1	S_2	S_3	S_a	S_b	S_c
T_θ	1.20	1.20	1.12	1.35	1.40	1.31
T_f	1.18	1.19	1.11	0.84	0.93	0,79
σ_T	0.016	0.008	0.009	0.378	0.336	0.397

S_1: CLWHQNVEIRMGDAKVQCSYTWEMRYF ($E_{\rm Nat} = -123.26$)
S_2: MFCYGTIEMRVQDWKWSHNCQVEYRAL ($E_{\rm Nat} = -124.20$)
S_3: VFCHQQMRLDIGRVEWNATCSWKYEYM ($E_{\rm Nat} = -122.93$)
S_a: IFMCDYCKVSMRTVERNWEWGYQAQHL ($E_{\rm Nat} = -120.68$)
S_b: FLQMEYWRVKWGNCEVHYSCQMTADRI ($E_{\rm Nat} = -119.17$)
S_c: WFYRTYIECNVQEMRHDWQMGVSCKAL ($E_{\rm Nat} = -119.48$)

a transition quickly to $P_S(\Gamma, T) \simeq 1$ at $T \simeq 1$. This means that when $T < 1.0$ the target Γ becomes dominant for these sequences and the sequences choose the target Γ as their unique native state due to $P_S(\Gamma, T) \approx 1$ when $T \to 0$. Differently, for the sequences designed by the EMM the curves of $P_S(\Gamma, T)$ versus T show some diversity, and $P_S(\Gamma, T)$ reaches the unit at a low temperature $T \approx 0.7$. The larger melting temperatures for the OPM cases indicate the higher thermodynamic stability of the corresponding designs, which shows the improvement of the OPM over the EMM. Note that the summation is only taken over the compact conformations, not all conformations for the chain.

MC simulations are performed to investigate the folding behavior of the designed sequences at various temperatures. As a characteristic, the overlap function

$$\chi = 1 - \frac{1}{N^2 - 3N + 2} \sum_{i \neq j, j \pm 1} \delta(r_{ij} - r_{ij}^N) \tag{3}$$

is calculated to measure a conformation overlap (or degree of similarity) with the native one.[14] Here r_{ij} is the distance between the i and j residues, and r_{ij}^N is the distance between the same residues in the native conformation. Obviously, we have $\chi = 0$ when the conformation is the native one, and $\chi = 1$ when the conformation is completely dissimilar with the native one. The specific heat

$$C = \frac{\langle E_S^2 \rangle - \langle E_S \rangle^2}{k_{\rm B} T^2} \tag{4}$$

is also calculated where E_S is the total contact energy of the system, $\langle \cdots \rangle$ means an canonical ensemble average over different conformations. $k_{\rm B}$ is the Boltzmann factor, and the thermodynamic average are calculated over a number of simulations. It is well believed that the folding of the chains undergo two kinds of transitions. The chains generally first change from random coil states to a compact phase. This transition is known as the collapse transition, and the transition temperature is

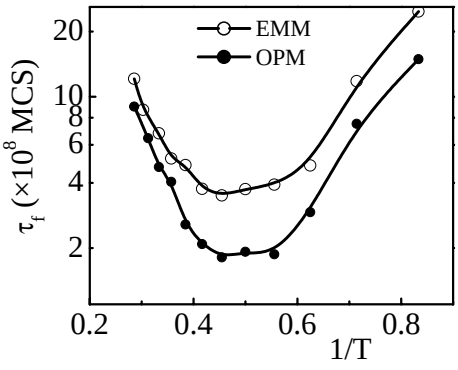

Fig. 2. The folding time τ_f (with logarithmic) versus the inversion of the temperature $1/T$ by averaging over six different sequences.

defined as T_θ which can be estimated by the temperature of peak of the specific heat. The second transition is the folding transition, whose temperature T_f can be identified from the peak in the fluctuation of the overlap function $\Delta\chi = \langle\chi^2\rangle - \langle\chi\rangle^2$. A well defined factor

$$\sigma = \frac{|T_\theta - T_f|}{T_\theta} \tag{5}$$

is used to characterize the foldability of the protein chains.[14] The better the folding behavior of the chain is, the smaller the factor σ is. This means that as T_θ approaches to T_f, the folding transition is close to the collapses, and the folding core of the native state forms early and easily. Consequently, the folding of the chain will be much fast, that is, with a small value of the folding time. In Table 1, we show the calculated values of σ, T_f and T_θ for the sequences designed by both methods, respectively. We can see that for the OPM the values of σ are $\sigma \sim 0.01$ which is about two orders of magnitude smaller than those for the sequences designed by the EMM. This clearly indicates the effectivity of the OPM. Figure 2 shows the folding time τ_f, i.e. the mean first passage time (MFPT), versus the inversion of the temperature. We see that the sequences designed with the OPM fold fast. Especially, around $T = 2.2$ the folding is about two times fast than those of the sequences designed by the EMM. However, for both cases the Arrhenius relationship of the $\tau_f(T)$ is still kept, i.e. $\tau_f \sim \exp(-A/T)$ with A being an energy-barrier-dependent parameter.

Thus, from Figs. 1–2 and Table 1, we know that the foldability of the sequences designed by the OPM is much better than those by the EMM. Now, let us make a physical interpretation for such a feature. The folding of the chain is described usually by the energy landscape theory, and the free energy of the system is characterized by a reaction coordinate Q,[4,5] i.e. a measure of the similarity of a conformation with the native one. For the native state, the value of Q is $Q = 28$ (for the

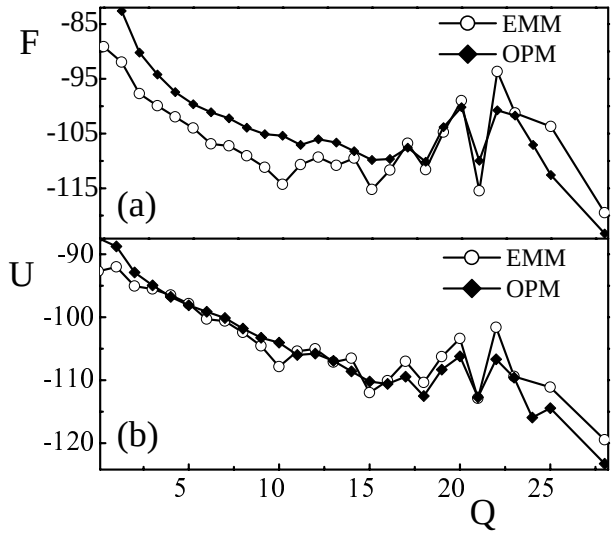

Fig. 3. The free energy (a) and internal energy (b) versus the reaction coordinate Q.

cases of the $3 \times 3 \times 3$ lattice chains). Thus the free energy is

$$F(Q,T) = U(Q,T) - TS(Q,T)\,, \tag{6}$$

where the entropy is $S = \ln(h_Q/h_N)$ with h_Q and h_N being the number of the total possible conformations with their $Q \leq 28$ and at the native state. Here, we have both free energy F and internal energy $U = \langle E_S \rangle$, i.e. one with the conformation entropy and another without the conformation entropy. In Fig. 3(a) and (b), we show several curves of these thermodynamic functions. In Fig. 3(a), the free energy profiles of the sequences designed by OPM have smaller fluctuations than those of sequences designed by EMM. They all have energy barriers for the chain folding to the native state ($Q = 28$). However, for the OPM the barrier is smaller than that for the EMM. That is, the sequences designed by OPM have an energy landscape not so rough compared with those sequences by the EMM. This can also be seen clearly from the internal energy profiles (see Fig. 3(b)). Thus the foldability (or the folding rate) is much better (or faster) for the OPM case. Finally, it is worthy to note that we have basically the same free energy and internal energy profiles for other sequences. This indeed implies the same features of the sequences designed by the OPM.

In conclusion, we have designed some sequences by using two different methods. By studying the folding kinetic and the thermodynamic features, we have seen that the sequences designed by the OPM show good foldability, and a fast folding. The physical origin of such a good foldability is attributed to less roughness of the energy landscape, or energy frustrations. Our design study can be developed for the off-lattice case as long as a set of "compact conformations" can be defined properly and are taken into account.

Acknowledgments

We acknowledge the support by the Foundation of the NNSF (Grant No. 90103031 and No. 10074030) and the Nonlinear Science Project (973) of the NSM.

References

1. C. Anfinsen, *Science* **181**, 223 (1973).
2. C. Levinthal, *J. Chim. Phys. PCB* **65**, 44 (1968).
3. E. I. Shakhnovich, *Folding & Design* **3**, R45 (1998).
4. P. G. Wolynes, J. N. Onuchic and D. Thirumalai, *Science* **267**, 1619 (1995).
5. K. A. Dill and H. S. Chan, *Nature Struct. Biol.* **4**, 10 (1997).
6. J. Wang and W. Wang, *Nature Struct. Biol.* **6**, 1033 (1999).
7. H. W. Hellinga, *Proc. Natl. Acad. Sci. USA* **94**, 10015 (1997).
8. J. U. Bowie, R. Luthy and D. Eisenberg, *Science* **253**, 164 (1991).
9. F. Seno, M. Vendruscolo, A. Maritan and J. R. Banavar, *Phys. Rev. Lett.* **77**, 1901 (1996).
10. S. Miyazawa and R. L. Jernigan, *J. Mol. Biol.* **256**, 623 (1996).
11. E. I. Shakhnovich and A. M. Gutin, *Nature* **346**, 773 (1990).
12. J. Y. Wang, J. Wang and W. Wang, *Chin. Phys. Lett.* **18**, 449 (2001).
13. H. Li, R. Helling, C. Tang and N. Wingreen, *Science* **273**, 666 (1996).
14. D. K. Klimov and D. Thirumalai, *Phys. Rev. Lett.* **76**, 4070 (1996).