Conservation of Hydrophobic and Hydrophilic Residues in Four-Helix Bundle *

QIN Meng(秦猛), WANG Jun(王骏), WANG Wei(王炜)

National Laboratory of Solid State Microstructure, Nanjing University, Nanjing 210093

(Received 19 May 2003)

The conservation of the hydrophobic and the hydrophilic residue sites obtained from 1000 designed sequences with the Z-score method for a four-helix bundle has been studied. The folding dynamic and thermodynamic features of the designed sequences and their different mutations are also studied. It is found that this conservation is related to the stability and the fast folding of the model proteins. Our results are consistent with the experimental results.

PACS: 87.15.Cc, 87.15.Aa, 87.14.Ee

The protein folding is still an unsolved problem in molecular biology.[1-3] Sequence design, the reverse folding problem, is an approach for understanding not only the relationship between amino acid sequences and their corresponding three-dimensional structures, but also the underlying physical principles that govern the folding and functions.[4-7] Experimental studies showed that the natural protein sequences which have been evolved for thousands of million years present some characteristics about the conserved arrangement of the hydrophobic (H) and the hydrophilic (P) residues at some sites in the sequences.[8,9] These conserved sites have a great correlation with their corresponding structural features, which is very important for the folding processes, the stability and biological functions.[10,11] To simplify the complexity of the real proteins, coarse-gained models are used in study of the folding processes and sequence design.[12,13] In this work based on a simple lattice model, we use the so-called Z-score design method[6] to design sequences to mimic the natural sequence evolution. We then identify the positions where the amino acids are most conserved in these designed sequences. We analyse the folding dynamic and thermodynamic features of both the "wild type" and mutated sequences to study the importance of the conserved sites to the folding process and stability. We find that the site conservation is consistent with the preservation of the features of folding and stability, and the results of the natural evolution are affected mutually by these factors. Our results are in agreement with the previous experimental and theoretical conclusions.

The lattice protein models have been argued to provide many insights for real proteins, such as the folding behaviour.[10,14] Two nonconsecutive beads, spatially neighbouring with one lattice spacing a, can form a contact. The energy of the system is considered as the contribution of all the contacts,

$$E_s^{\Gamma} = \sum_{i \geq j+3} U_{s_i,s_j} \delta(r_{ij} - a), \tag{1}$$

where Γ indicates the conformation, U_{s_i,s_j} is the contact potential between residue S_i and S_j and the potential of the modified MJ matrix (MMJ) is used,[15,16] and $\delta(r_{ij} - a)$ characterizes the geometrical requirement of the contact between residues i and j with $\delta(0) = 1$ for a contact or 0 otherwise.

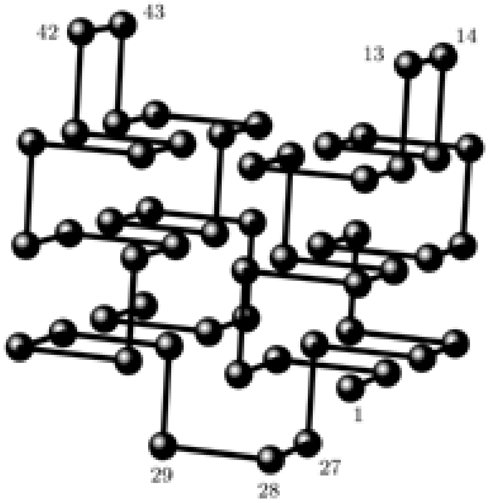

Fig. 1. Native structure of the four-helix bundle lattice model. The chain length is 55 and the number of the native contacts is 61.

The Z-score design method is a popular and successful method among all of the current protein design methods. It enlarges the energy gaps between the native state and the unfolded states to realize a funnel-like landscape. The Z-score is defined as

$$Z_{\text{score}} = \frac{E_{\text{average}} - E_{\text{target}}}{\sqrt{\langle E^2 \rangle - \langle E \rangle^2}}, \tag{2}$$

where E_{target} is the energy for the sequence in the target structure, E_{average} is the average energy of the unfolded conformations for the sequence and is estimated from the contact averages as $E_{\text{average}} = N \times \langle e \rangle$, in which $\langle e \rangle$ being the average energy of all possible

* Supported by the National Natural Science Foundation of China under Grant Nos 90103031, 10074030, and 10204013, and the Nonlinear Science Project (973) of the Ministry of Science and Technology of China.

contacts and N is the number of contacts in the native state or the target state. By fixing the compositions of H-residues (L, I, V, M, C, F, Y, W) and P-residues $(R, S, T, Q, H, D, E, K, G, A, N, P)$ to 22:33,[18] we design more than 1000 sequences (each sequence is designed for 10^7 Monte Carlo steps, and because the amino acids of each sequence are random chosen, each designed sequence is nearly independent) by maximizing the values of the Z-score. A target structure for the design is a lattice model of the four-helix bundle protein which contains only α-helix (see Fig. 1).[19,20]

We obtain the distribution of the hydrophobic and the hydrophilic residues at each site of the model chain by averaging over 1000 designed sequences. The occurrence of the hydrophobic residues (OHR) in each site of the chain is shown in Fig. 2. It is found that there are 11 sites where the probabilities for the hydrophobic residues occurring are more than 90% (or the value of OHR ≥ 0.90). These conserved hydrophobic sites distribute in the interior of four helices, and interact with each other thus form a hydrophobic network. This hydrophobic interaction network performs as the hydrophobic core, which plays an important role on the stability of the native state and the folding process, especially in the beginning of the folding since the hydrophobic interaction is the main driving force.[21] Surprisingly, about 16 sites distribute few or none H-residues among the 1000 sequences, and their values of OHR ≤ 0.1. These conserved hydrophilic sites are located at the surface of the four-helix bundle and are so packed that the waters are excluded from the interior of the structure to make the model chain more stable. We also find that there are six extraordinary conserved sites, namely sites 13, 14, 27, 29, 42, 43 with OHR $= 0$. According to the structure of the four-helix bundle, these sites are located at three turns (or loops) which connect the four helices. The presence of these hydrophilic (polar) residues decreases the opportunity of hydrophobic residues to form other competitive cores, which leads the energy landscape to be smooth. These weak interactions also increase the flexibilities of some local conformations, for example, the loop regions. It is well known that in real proteins most of the residues on the surface are hydrophilic. Our results really imply that this observation and the conserved hydrophilic residues are highly on the surface sites and have interactions with the solvent molecules. The separation of the hydrophobic and hydrophilic residues in the structure can energetically increase the energy gap, which increases the stability of the target structure. Changing the composition of different ratios of the H-residues to the P-residues, we obtain the similar results (data not shown). From our results, we also find that some sites located between the hydrophobic core and the hydrophilic surface are not so sensitive to the types of the residues and the values of OHR are about 0.3–0.5. These sites locate at the interface of the hydrophobic core and the hydrophilic surface, reflecting the diversity of the protein sequences.

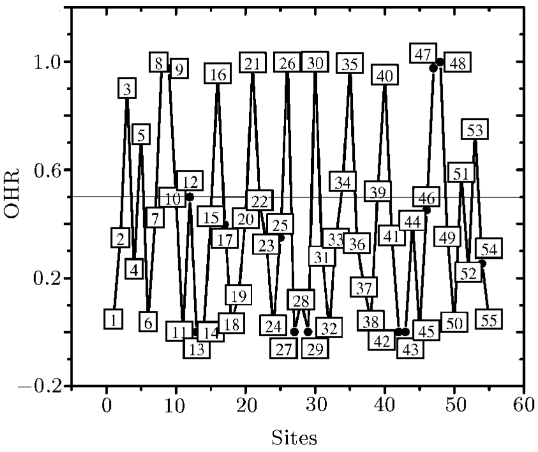

Fig. 2. Occurrence of the hydrophobic residues (OHR) distributions over the site. The number in each pane represents the corresponding sites.

The conservation of the hydrophobic residues and the hydrophilic residues is related to the natural proteins.[11] For the four helix domain in the nature proteins (for example the cytochrome b_{562}, human growth hormone, myohemerythrn and so on), the residues in each helix of the domain are arranged so that the hydrophobic residues are buried between the helices, and the hydrophilic residues are on the outer surface of the domain. For each helix, the arrangement of the residues has a tendency to change from the hydrophobic residues to the hydrophilic residues with a period of about four residues. This feature is also reflected in our designed sequences as shown in Fig. 3. The arrangement of residues for each helix is similar to the shape of the letter "W", and this distribution can make the hydrophobic residues in the interior of the structure to form the hydrophobic core and the hydrophilic residues to expose to the solvents for the folding and function reasons. It is found that each helical segment has a similar distribution of the hydrophobic and the hydrophilic residues and the occurrences of the hydrophobic residues in each helix segment are 47.42%, 46.21%, 46.30%, 42.40%, respectively. This shows that the four helices have similar role in their folding and stability.

The special residue patterns are closely correlated to the stability and the dynamic characters of the proteins. Now we use the mutation method to analyse what can be produced from the folding characters and stability.[10,11,22] We make the mutations at four sites of a designed sequence (the wild-type sequence), namely two sites in the region of the conserved hydrophobic core (site 8 and 26) and the other two in the non-sensitive sites (site 12 and 22). All of the mutations are made simply by replacing the wild H-type residue to a P-type residue or by replacing the wild P-type residue to an H-type residue. For these mutated sequences, we perform the folding simulations, and analyse the dynamic and thermodynamic features with respect to those of the original wild-type sequences.

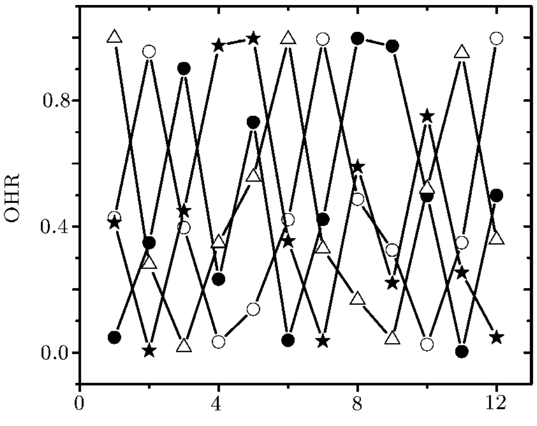

Fig. 3. OHR of each helix of the four-helix bundle versus the chain sites. The symbols •, ∘, △, and ⋆ correspond to the 1, 2, 3, 4 helix of the four-helix bundle.

The thermodynamics of the sequence can be reflected by the folding temperature T_f, which is measured by the peak of the heat capacity versus the temperatures.[23] We find that the folding temperatures of all these mutated sequences are less than the original sequence, especially for those sequences mutated at the hydrophobic core. In details, the T_f values are 0.50, 0.44, 0.42, 0.41, 0.37 for the wild type sequence, sequence mutated at the sites 22, mutated both at 22 and 12, mutated at 8, and mutated both at 8 and 26, respectively. At the same time, the gaps between the $E_{\rm average}$ and $E_{\rm target}$ become small after the mutations are made. This shows that the highly conserved regions of the hydrophobic core are very important to the thermodynamic stability of the native structure and these conserved sites are also the requirement of the structural stability.

We obtain the relationship of the folding time τ_f over a wide range of temperatures as shown in Fig. 4. The shapes of the curves of τ_f versus $1/T$ for four mutated sequences are similar to that of the original wild-type sequence but the fast folding (the minima of τ_f) temperatures $T_{\rm min}$ shift to the low temperatures. When the mutation is made for a conserved hydrophobic residue, e.g., site 8, from a hydrophobic residue M to a hydrophilic residue D, the folding becomes slower than that of the original wild-type sequence (see the curve with the filled circles). When we mutate two sites (sites 8 and 26) in the core region, it folds slower than that of one site mutation and the fast folding temperature shift to much low temperature (see the curve with open diamonds). Thus, the conserved hydrophobic sites affect the folding processes and play an important role in the thermodynamic stability. It is noted that the values of τ_f for the sequences mutated in the non-sensitive regions are similar to that of the wild type one (data not shown). At high temperature we observed that the four mutated sequences fold slower than the wild type sequence, especially for the sequences mutated at the core position. This is due to the fact that at high temperatures, the partially folded conformations of the mutated sequences can be destroyed more easily, leading to a large number of the partially folded conformations. Therefore, it needs longer time to search the folded state among these partially folded states of the mutated sequences than that the wild type sequence does. This effect is more significant for the sequences mutated at the core positions. However, the folding behaviours are more or less the same if the mutations are made for these conserved sites with the residue belonging to the same types, for example an H-type is replaced by another H-type residue. Our simulations are consistent with the experimental results of the mutation of the conserved core of the $SH3$ domain.[24] Thus, these conserved sites have been evolved to fit the requirement of the fast folding. We can conclude that the characteristics of the sequences mutated with different types of residues in the region of the hydrophobic core become worse both in the kinetic folding process or in the stability. The mutations for the insensitive sites almost do not affect the folding process but increase the thermodynamic stability little bit. Such features may result from the evolutionional conservation. For

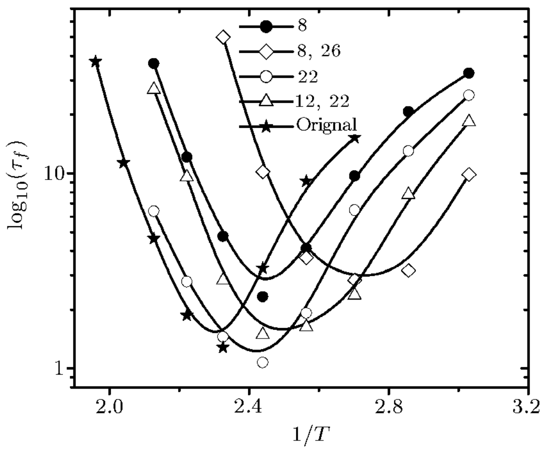

Fig. 4. Logarithm values of folding time $\log_{10}(\tau_f)$ versus the reciprocal of the folding temperature $1/T$. The symbol ⋆ corresponds to the wild-type sequence, • to the sequence mutated at the site 8 from residue M to D, ⋄ to the sequence mutated at the sites 8 and 26 from residues M and I to D and G, ∘ to the sequence mutated at the site 22 from residue I to H, △ to the sequence mutated at the sites 12 and 22 from residues K and I to F and H, respectively.

the case of the two-site mutations related to pairs of the H-residue and P-residue, the effects depends on its place in the four-helix structure, and the detail results would be presented in our further study. For the case of the two-site mutations related to pairs of the H-type and P-type residues, the effects depend on sites in the four-helix structure, and the detail results would be presented in our further study.

In summary, we show that the conserved residues from 1000 designed model protein sequences have a tight relationship with the important regions of the native structure. From the analysis of the dynamics and thermodynamics for these sequences (wild and mutated), we show that the conserved residues in the evolution of the protein sequences are very important to the folding process and the thermodynamic stability. Protein sequences selected by the natural environment for thousands of years conserve important residues in some special regions. These particular patterns are the requirement of the fast folding process, the stability of the native structures and also the functional activities. Conclusively, the sequence and the folding/stability cooperate harmoniously to realize the conservation of proteins. Our observation is really consistent with some experimental studies of the four-helix bundle protein design.[4] These can also provide us some insights into protein folding and the function prediction.

References

[1] Anfinsen C 1973 *Science* **181** 223
[2] Levinthal C 1968 *J. Chim. Phys.* **65** 44
[3] Wolynes P G et al 1995 *Science* **267** 1619
[4] Kamtekar S et al 1993 *Science* **262** 1680
[5] Shakhnovich E I 1994 *Phys. Rev. Lett.* **72** 3907
[6] Bowie J U, Luthy R and Eisenberg D 1991 *Science* **253** 164
[7] Pabo C 1983 *Nature* **301** 200
[8] Branden C and Tooze J 1999 *Introduction to Protein Structure* (New York: Garland Publishing)
[9] Huang Y Z and Xiao Y 2002 *Chin. Phys. Lett.* **19** 434
[10] Shakhnovich E I, Abkevich V I and Ptitsyn O 1996 *Nature* **379** 96
[11] Plaxco K W et al 2000 *J. Mol. Biol.* **298** 303
[12] Leach A R 1996 *Molecular Modelling Principles and Applications* (London: Addison Wesley Longman)
[13] Wang J and Wang W 1999 *Nat. Struct. Biol.* **6** 1033
[14] Nymeyer H, Garcia A E and Onuchic J N 1998 *Proc. Natl. Acad. Sci. USA* **95** 5921
[15] Betancourt M R and Thirumalai D 1999 *Protein Sci.* **8** 361
[16] Qin M, Wang J and Wang W 2003 *Phys. Rev.* E **67** 061905
[17] Cordes M H, Davidson A R and Sauer R T 1996 *Curr. Opin. Struct. Biol.* **6** 3
[18] Wang J Y, Wang J and Wang W 2001 *Chin. Phys. Lett.* **18** 449
[19] Kaya H and Chan H S 2000 *Phys. Rev. Lett.* **85** 4823
[20] Kaya H and Chan H S 2002 *J. Mol. Biol.* **315** 899
[21] Dill K A 1991 *Biochemistry* **29** 7133
[22] Matouschek A et al 1989 *Nature* **340** 122
[23] Klimov D K and Thirumalai D 1996 *Phys. Rev. Lett.* **76** 4070
[24] Riddle D S et al 1999 *Nat. Struct. Biol.* **6** 1016