Composition Preference of Amino Acids in Model-Proteins *

WANG Jian-Yong(王建勇), WANG Jun(王骏), WANG Wei(王炜)
National Laboratory of Solid State Microstructure and Department of Physics, Nanjing University, Nanjing 210093

(Received 14 August 2000)

The folding behaviour is investigated of some sequences of 36 monomers with different proportions of hydrophobic residues in a three-dimensional lattice based on the Miyazawa–Jernigen interaction matrix. It is found that the sequences with good folding properties are those with an optimal number of hydrophobic residues, neither too many nor too few. The reason for the deterioration of folding properties of sequences out of this range has also been analysed.

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

As the basic 'alphabet' of proteins, the composition of amino acid residues is one of the important ingredients in proteins.[1,2] Naturally, proteins are composed of 20 kinds of residue, and the ratio of various residues is sometimes regarded as the zeroth structure of proteins. According to their different affinities to water, protein residues are classified into two types: hydrophobic (H-type) residues (including the residues C, M, F, I, L, V, W, Y, A, G) and polar (P-type) residues (including the residues T, S, N, Q, D, E, H, R, K, P). As a coarse-grained consideration, the compositional feature then is described by the ratio of the number of the H-type residues to the total number of residues in a protein. It is well known that the average ratio for natural proteins approaches 51%, and that proteins with too many or too few H-type residues are rarely found in nature. Generally, there is a combinatorial preference to the sequences with a different proportion of residues. The number of possible sequences with equal H-type and P-type residues is much larger than other cases. Are the compositional properties dominated by this combinatorial preference? Some recent studies based on lattice models showed that there is a preference to the composition due to the foldability requirement or the structural reasons.[3–6] This implies that there might be some connection between the composition and the foldability or the structure. How does the composition preference affect the folding behaviour? What ratio of the H-type residues in proteins is the best for proteins to have good folding behaviours?

In fact, a main structural characteristic of natural proteins is the segregation of the H-type and P-type residues. Therefore, a proper arrangement of the residue composition is necessary to fulfil this feature. From a coarse-grained view, the composition may influence the realization of this structural property. Using the landscape language, too few H-type residues often make the energy landscape of the resulted proteins fairly flat, due to the weak attraction between polar residues. On the other hand, too many H-type residues make the landscape rather rugged since there may be many non-native contacts between the H-type residues, which makes the transition from one minimum to another difficult (see Fig. 1). Therefore, there should be a suitable proportion of the H-type residues which optimizes the mobility of the system on the energy landscape and the stability of the native state simultaneously. The corresponding composition may be the favourite during the nature selection.

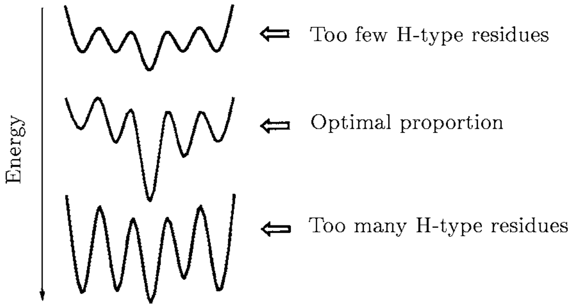


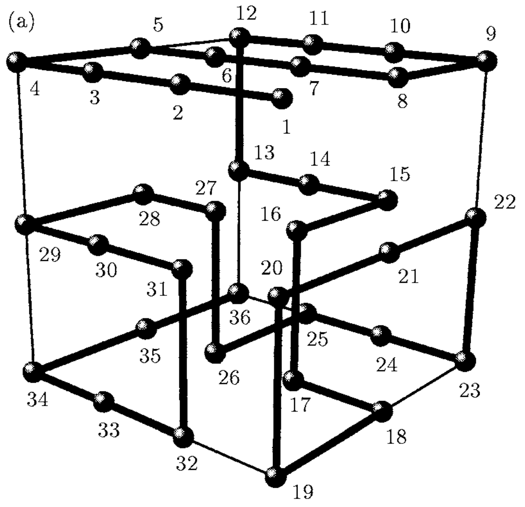
Fig. 1. Illustration of free energy profile with small, median and large numbers of the H-type residues. Different features of the profiles (flat, protein-like or rugged) are sketchily demonstrated.

In this letter, based on the foldability of model-proteins, we study the composition preference of the H-type and the P-type residues in the sequences. A number of sequences are designed by a Monte Carlo (MC) procedure in the sequence space to maximize their foldability (characterized by the Z score[7]). We find that some sequences with a proper composition have a better foldability. The structural reason for the composition preference is also discussed.

In our modelling, 'proteins' are represented as self-avoiding chains on a cubic lattice.[8] The model chains with 36 monomers are studied in this work. All the concerned model-proteins have the same native state as used in other studies[9] (see Fig. 2(a)). For the lattice chains, the energy of a conformation is the sum of the energies of pairwise contacts between monomers.

*Supported by the National Natural Science Foundation of China for Outstanding Young Researchers under Grant Nos. 19625409 and 10074030.

Two monomers are regarded to be in contact with each other when they are neighbours on the lattice (not connected by a covalent bond). In our simulation, we take the Miyazawa–Jernigan (MJ) statistical potential matrix[10] in practice, which is a widely used experimental interaction in the protein design and prediction studies.



(b)

Ratio of H/P	Sequences
11/25	dqnkemhqkgrwevtlgpkgptksnyiadchsnder
12/24	sqpkgyiedrtkphclvskdvndrtalgnhmnqkge
13/23	dqqdriytkgrdnmalhpkshgkgnvlcnvatpees
14/22	egtdricrkggdnvalhsknaskpnmlwdvyqtepq
15/21	ghakpylsegtknamliqdrvqdpsvlgrcfnhkte
16/20	napdniwqdhpervafcskgasktgllvemyrgdtq
17/19	natdnfwpkateriylvgdgaqkhsilcemvgpdsr
18/18	ntwepfcrdtanglyliskrvskpamladivghegq

Fig. 2. (a) Native structure studied in this work. (b) Representative sequences with different ratio of H/P residues.

We employ a sequence design procedure[11] to create optimal sequences. A quantity based on the statistical landscape paradigm, the Z score, is used as an estimation of the foldability of the selected model chains.[7] The quantity Z has the form of $(\langle E\rangle - E_{\rm nat})/\sigma$, which quantifies the degree of the frustration of the free energy landscape. Here $\langle E\rangle$ is the average energy and $E_{\rm nat}$ is the energy of the native conformation. A large value of Z means less frustration in the free energy surface of the model-protein. As an efficient computational approach, we calculate $N\langle e\rangle$ as $\langle E\rangle$ where N is the number of contacts in the compact conformation and $\langle e\rangle$ is the average energy of all topologically possible contacts between all residue pairs, and σ is the corresponding variation. With these settlement of models, an MC procedure in the sequence space is used to find the possible sequences with a maximal value of Z score. The sequence space is sampled by randomly exchanging two residues at different sites with a fixed composition. A series of sequences with different compositions are designed. For each composition, a sequence with the best foldability is picked out as a representative as listed in Fig. 2(b). These designed sequences are verified to have protein-like characteristics and are used as the objects of our further studies.

To test the foldability of these sequences, their folding behaviours are studied with dynamic simulations by an MC procedure.[12] Each simulation starts from a random coil conformation. To measure the progress of the folding, we use the structural similarity between a given conformation and the native state (Fig. 2(a)), Q, as an order parameter, which is the number of native contacts in the present conformation. For our 36-monomer case, Q has the largest value of 40 for the chain in its native state.

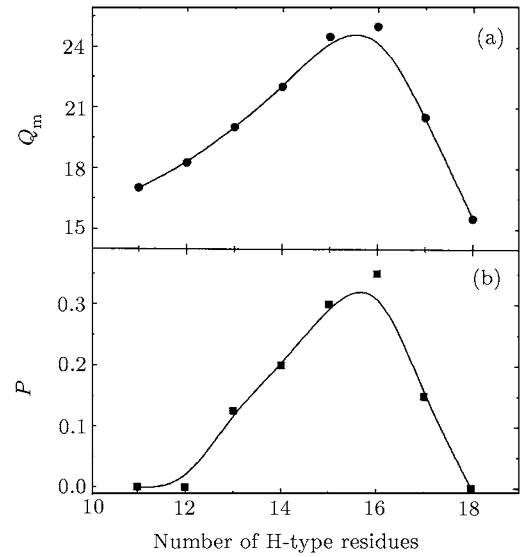


Fig. 3. (a) $Q_{\rm m}$ and (b) P versus the number of the H-type residues.

The foldability of a sequence is measured by both the average of the contact overlap $Q_{\rm m}$ between the native structure and the lowest energy structure with energy $E_{\rm min}$ sampled during each trial, and the probability P for this sequence to reach its native state within the sampling time.[13] It is noted that the mean first passage time (MFPT) is not used to quantify the foldability which is computationally infeasible for such large chains. However, the quantities $Q_{\rm m}$ and P can identify sequences that can move close to their native state. It is tested that the qualitative aspects of foldability with MFPT and the present characteristics are similar. In our work, the sampling time is 4×10^8 MC steps, and $Q_{\rm m}$ and P are averaged over 10 times for each sequence.

The foldability (characterized by $Q_{\rm m}$ and P) versus the number of the H-type residues is shown in Fig. 3. In this figure, the sequences with the median composition have larger $Q_{\rm m}$ and the probability P is also higher than others, thus they have better foldability. The corresponding composition is regarded as 'good' from this foldability view. The value of $Q_{\rm m}$ of the sequence outside this optimal composition region is generally smaller (lower than 20), and it is nearly impossible for them to find their native states within the sampling time. It shows that it is difficult to find the fast-folding protein sequences out of the optimal composition region. As a result, nature may seldom choose the 'bad' composition because they are not favoured to implement the basic fast-folding feature of natural proteins. This may offer a restriction on the composition based on the foldability feature.

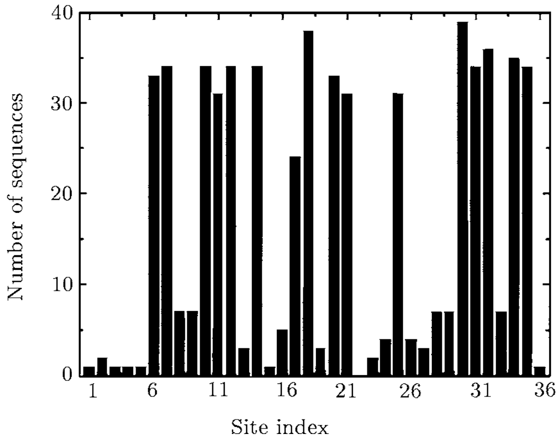


Fig. 4. Count of the number of sequences with the H-type residues at a site, with 40 sequences in statistics.

To obtain a detailed understanding of the composition preference in our lattice model, we count the number of sequences with the H-type residues at a site, which reflects the probability of the occupation by the H-type residues at site i in the sequences (see Fig. 4). We also find that the H-type residue prefers to be located at some sites with larger probability than that of the P-type residues. This indicates a certain pattern of the foldable sequences, which relates to the structural feature of the model-proteins. We find that the maximal number of the sites which the H-type residues favour (with a count larger than 20) is 16, which coincides with the number of the most favourite H-type residues from the foldability examination. These sites form a suitable hydrophobic core in the native conformation. When the number of the H-type residues decreases, the flattening of the energy surface accompanies a weak combination of the inner residues. The decrease of the energy gap $\langle E\rangle - E_{\rm nat}$ leads to the decrease in the stability of model-proteins. It is not easy for the residues to form a stable core to help a protein find its final structure. On the other hand, the increase of the number of the H-type residues from the optimal condition may induce some other strong non-native contacts between the H-type residues, which generally stabilizes other low-energy local minima that are quite dissimilar from the native state. This makes the diffusion of the system on the energy surface rather difficult, and seriously affects the dynamics. This factor plays an important role in the low-temperature folding dynamics, which makes the increase of H-type residues from the optimal condition affect the foldability more intensively than the case with fewer H-type residues. It is consistent with the change of foldability versus the number of H-type residues shown in Fig. 3.

In conclusion, the folding properties of a series of designed model-proteins in a three-dimensional lattice with the MJ interaction matrix are studied. It is found that the good sequences must have an optimal number of residues, which is consistent with the compositional properties of natural proteins. The composition preference as well as the detailed arrangement of residues has its connection with the feature of the native structure, which may be helpful for further understanding on folding processes and sequence designs. The composition preference also suggests the conservation of a suitable hydrophobic core. The certain arrangement of the hydrophobic core or the topology of the native structure are considered as the underlying reason of the favouritism of the composition of the amino acids.[14]

References

[1] Wang J and Wang W 1999 *Nat. Struct. Biol.* **6** 1033

[2] Plaxco K W, Riddle D S, Grantcharova V and Baker D 1998 *Curr. Opin. Struct. Biol.* **8** 80

[3] Wang J and Wang W 2000 *Phys. Rev.* E **61** 6981

[4] Yue K & Dill K 1992 *Proc. Natl. Acad. Sci.* USA **89** 4163

[5] Shakhnovich E, Abkevich V and Ptitsyn O 1996 *Nature* **379** 96

[6] Micheletti C, Seno F, Maritan A and Seno F 1998 *Phys. Rev. Lett.* **80** 2237

[7] Bowie J U, Luthy R and Eisenberg J D 1991 *Science* **253** 164

[8] Chan H S and Dill K A 1991 *Annu. Rev. Biophys. Biophys. Chem.* **20** 447

[9] Klimov D and Thirumalai D 1998 *J. Chem. Phys.* **109** 4119

[10] Miyazawa S and Jernigan R 1996 *J. Mol. Biol.* **256** 623

[11] Shakhnovich E I 1994 *Phys. Rev. Lett.* **72** 3907

[12] Socci N D and Onuchic J 1994 *J. Chem. Phys.* **101** 1519

[13] Dinner A R, Sali A and Karplus M 1996 *Proc. Natl. Acad. Sci. USA* **93** 8356

[14] Kuhlman B and Baker D 2000 *Proc. Natl. Acad. Sci. USA* **97** 10383