CHIN. PHYS. LETT. Vol. 27, No. 3 (2010) 038702

Dissection of the Zipping-and-Assembly Mechanism for Folding of Model Proteins *

SUN Li(孙理), WANG Jun(王骏)**, WANG Wei(王炜)

Nanjing National Laboratory of Microstructure and Department of Physics, Nanjing University, Nanjing 210093

(Received 26 November 2009)

Zipping-and-assembly mechanism (ZAM) is a new mechanism describing the kinetics of protein folding. To dissect the validity of this mechanism for various protein-like systems, a prediction test based on three-dimensional HP lattice models is carried out. It is found that only the native structures of a part of protein-like models could be predicted with a ZAM-based method. The detailed comparisons between the model proteins which are predicted or failed with the ZAM-based method suggest that the ZAM is likely to be applicable for the model proteins with the weak hydrophobicity, the low contact order for native conformations, and the large separation between the energies of native state and denatured states. These observations bring us more information about the protein-like systems for which the ZAM could be applied.

PACS: 87. 15. Cc, 87. 15. B−, 87. 15. Hm DOI: 10.1088/0256-307X/27/3/038702

The folding of proteins is a fundamental and complicated process in molecular biophysics.[1−3] Due to the complexity of protein systems, presently there are no unified mechanisms for various protein systems. Various mechanisms are proposed and discussed to describe the underlying physics for such a kind of self-organization behavior of proteins, including the nucleation-condensation mechanism, the diffusion-collision mechanism, the framework model and so on.[4−6] The introduction and dissection of various folding mechanisms have become one of the basic themes in protein studies.[7−11]

Recently, a new mechanism, the zipping-and-assembly mechanism (ZAM), was proposed as a new description for the folding processes of proteins.[11−15] Based on this mechanism, the folding is believed to be realized through the growth of local structures (zipping) and the installation of formed motifs (assembly). This mechanism is an extension for sequential stabilization mechanism in the microscopic level. The most interesting thing is that the zipping behavior may limit the range of conformational searches and speed up the folding. Consequently, an efficient prediction method based on this mechanism was proposed and has been successfully applied in the prediction practice for the native conformations of some proteins.[11,15] Taking account of these facts, the ZAM is believed to be a valuable progress to describe the folding processes. For this new mechanism, it would be interesting to ask: is the ZAM a unified mechanism for all proteins, and what kind of protein-like systems would be suitable to be described with the ZAM mechanism? The answers to these questions would be valuable to understand the role of the ZAM for the folding of proteins, and may be helpful to build an integrated understanding for natural protein systems.

In this work, three-dimensional HP lattice models composed of hydrophobic (H) and polar (P) monomers are used to analyze the validity of the ZAM. By enumerating over all compact conformations, a series of 27-mer sequences with protein-like features are generated. It is found that some model proteins cannot fold to their native structures with the ZAM-based greedy searching algorithm. This implies that the proteins which fold following the ZAM would be a certain subset of all protein-like models. Various features of model proteins which can be predicted with the ZAM-based method are compared with those of the other protein-like systems, including the sequential patterns, the contact order of native conformations, and the property of energy spectra. It is found that the ZAM-based method would be more successful for the proteins with weaker hydrophobicity, better accessibility to native conformation and higher cooperativity. This is consistent with the basic assumption in the ZAM. All these results demonstrate that the ZAM proposes a refinement for the general landscape theory for proteins and would be reasonable for the optimal folders.

To model natural protein systems, three-dimensional (3-D) lattice polymer models composed of the hydrophobic (H) and polar (P) monomers are used. In our simulations, the typical 27-mer chains are used, and the classical HP interaction is employed, that is, the interaction ϵ between two neighboring H residues takes the value $\epsilon_{\rm HH} = -1$, and all the other interactions (between H-type and P-type monomers or

*Supported by the National Basic Research Program under Grant Nos 2006CB910302 and 2007CB814806, the National Natural Science Foundation under Grant Nos 10834002 and 10774069, the Foundation for the Author of National Excellent Doctoral Dissertation of China, and the Fund of Jiangsu Province (BK2009008).

**To whom correspondence should be addressed. Email: wangj@nju.edu.cn

between P-type monomers) are zero ($\epsilon_{\rm HP} = \epsilon_{\rm PP} = 0$). This kind of models have been widely used.[16–19] It is worth noting that, different from previous analyses with the 2-D lattice models,[12] the 3-D lattice models would be more suitable to simulate natural proteins since the 3-D cubic lattice has a larger number of coordinates and could form more interactions with long sequential separations. For this kind of model, the energy of a chain at a conformation $\mathcal{C}$ would be the summation over all the pairs of neighboring monomers, $E(\mathcal{C}) = \sum_{i<j} \Gamma_{ij}\epsilon_{s_i s_j}$, in which $s_{i(j)}$ represents the type (H or P) of the residue $i(j)$, and $\Gamma_{ij} = 1$ when the residues i and j are spatial neighbors or 0 otherwise. This kind of interactions reflect the basic feature of protein systems, and are also widely used in protein studies.

It is well known that not all the sequences composed of H/P monomers behave like proteins. In this work, 1128 HP sequences with protein-like properties are selected as a sample from the whole universe of protein sequences. Each selected sequence has a unique ground state (namely the degeneracy of the lowest energy level equals to 1), which mimics the basic requirement for natural proteins. Practically, the degeneracy of the ground state (namely the lowest energy level) could be determined by enumerating all compact cubic conformations for 27-mer chains. This implementation is related to the attractive feature of HP interaction. As a result, these protein-like sequences build up a collection of objects for the following analysis.

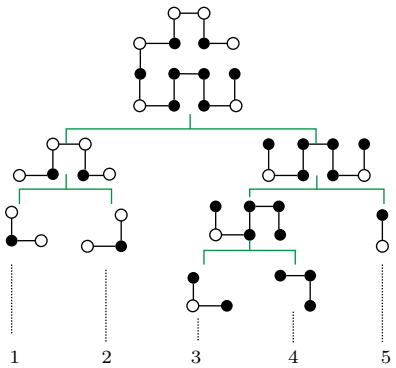

Fig. 1. An example of folding process based on ZAM mechanism, which demonstrates the coalescences of neighboring segments as represented by a binary tree.

To exemplify the effect of the ZAM in folding processes, a ZAM-based method for folding prediction is realized for 3-D lattice models with a CKY graph-searching algorithm, similar to that used in Refs. [11,12]. In detail, by dividing a chain into several segments with three contiguous monomers, the folding process is described as a series of coalescences of the adjacent segments. All the possible combinations of local segments could be examined with dynamic programming. The whole folding process could be represented as binary trees (as shown in Fig. 1). To highlight the effect of the ZAM, a greedy searching algorithm is employed in the present work. That is, only the local structures with minimal energy are kept as the seeds for the next step of mergence. This kind of operations maximizes the greedy feature of sequential stabilization.

With the ZAM-based greedy searching algorithm, the landscape of all the selected protein-like sequences are studied. The conformations with the minimal energy for each sequence are recorded and compared with the corresponding native conformations. It is found that the native conformations could be successfully identified with our ZAM-based algorithm for 363 sequences (32.2% of the sequences we used). While for the others (765 sequences), the ZAM-based method generally fails to observe the native conformations. These two kinds of model proteins are noted as successful ZAM folders (SZF) and the unsuccessful ones (UZF), respectively. The proportions of these two kinds of sequences are not sensitive to the number of protein-like sequences or the implementations for the ZAM-based method (data are not shown). The result of this prediction test demonstrates that the ZAM may be suitable to describe some kinds of protein-like chains, but is definitely not a unified mechanism for various folding behaviors. In this sense, the ZAM is likely to act as a refinement for general landscape theory for proteins.

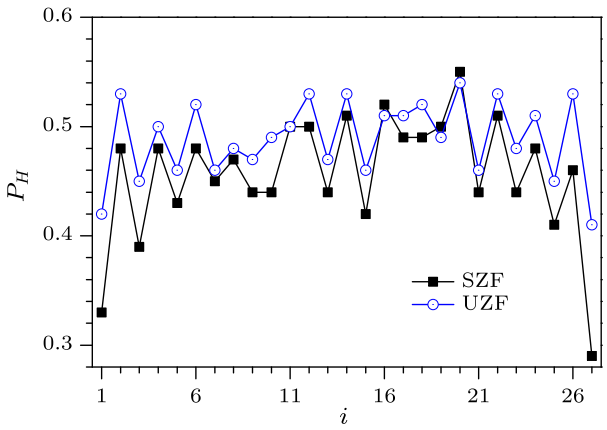

Fig. 2. The probability P_H for the site i to take H-type residues. Solid squares and open circles are used for SZF and UZF cases, respectively.

As a description for some specific folding behaviors, the ZAM would work for some certain subset of protein-like sequences. What kinds of model proteins would like to fold following the ZAM? The applicability of the ZAM is important to bring some insights on the connection between the properties of protein systems and the folding mechanisms. In this work, this question is answered by comparing the SZF and UZF from various features related to sequential patterns, native conformations and energy spectra. These

analyses from multiple aspects may provide a comprehensive picture for the ZAM-related folding.

Firstly, the sequential HP patterns are compared for two kinds of sequences. The sequential pattern is characterized with the probability of the ith site taking H-type monomers, $P_H(i) = \sum_{s_i=H} 1/N_{\rm seq}$, in which $N_{\rm seq}$ is the number of all concerned sequences in statistics. The probability P_H could be calculated for the SZF and UZF. They are marked as $P_H^{\rm SZF}$ and $P_H^{\rm UZF}$, respectively. These probabilities are observed to fluctuate around 0.5 with a similar trend (as shown in Fig. 2). More specifically, the probability $P_H^{\rm SZF}$ is smaller than $P_H^{\rm UZF}$, which indicates that the SZF model proteins have few number of H-type monomers for almost all the sites, especially for those at the head or tail. The decrease of the hydrophobic composition for SZF may physically reduce the chances to form unexpected hydrophobic clusters and to avoid the trap formations during ZAM-related folding. This phenomenon implies that weak hydrophobicity would be compatible with the ZAM.

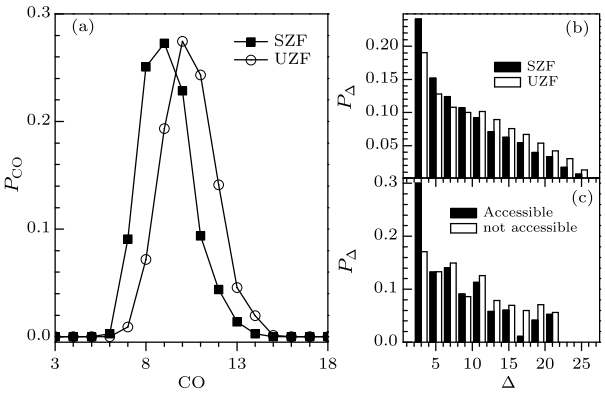

Fig. 3. Comparisons of native conformations for SZF and UZF cases. (a) The probability $P_{\rm CO}$ to take a certain contact order CO. (b) The probability P_Δ for residues to take a certain sequence separation Δ for SZF (solid) and UZF (open). (c) The probability P_Δ for the conformations accessible (solid) and not accessible (open) by ZAM.

The native conformation of a protein-like system is another important feature related to foldability.[20–22] In comparison between the SZF and UZF, the contact order (CO) of the native conformation is used to describe the conformational feature of protein-like model. Here the contact order is defined as CO $= \sum_c |j_c - i_c|/N_{\mathcal{C}}$, where i_c and j_c are the monomer indices related to the contact c, and $N_{\mathcal{C}}$ gives the number of contacts in the conformation $\mathcal{C}$. The distributions of contact orders for the SZF and UZF, $P_{\rm CO}$ are obtained as the histograms in Fig. 3(a). It is found that the contact orders for the SZF are statistically smaller than those for the UZF. The detailed statistics on the sequential separations $\Delta = |j - i|$ for various contacts are also presented. There are more contacts with short sequential separations for SZF (as shown in Fig. 3(b)). For this kind of native conformations with more short-separation contacts, the conformational searches with local zipping operations would be easier to find out the native contacts. The ZAM would kinetically favor this kind of conformational feature. This kind of feature could also be observed from the accessibility in conformational space with the ZAM-base method. For example, for a non-protein-like sequence (PHHHPHPPPHPHPHHPHHPHPHHHPPP), the degeneracy of its ground state is 106. The search with the ZAM-base method only can find out 33 of them. The ratio of accessible space is similar to that of the SZF in all protein-like models. The statistics on the contacts in the accessible 33 conformations and the others are given in Fig. 3(c). The conformations accessed by the ZAM-based method are also have more contacts with short separations. This further supports the intrinsic connection between the ZAM and low contact orders of conformations. As a result, the ZAM would be more suitable for the models with small contact orders for their native conformations.

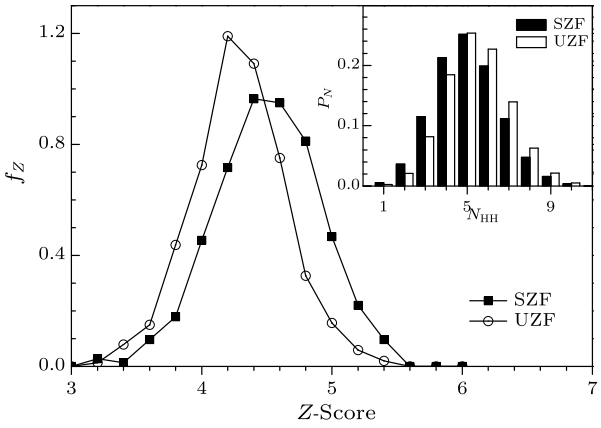

Fig. 4. Probability density f_Z to take a certain Z-score for SZF and UZF. The inset gives the probability P_N to have $N_{\rm HH}$ contacts between H-type residues for two sequences, PHPHPHHPHPPHHPPPHPPPPHPHPHP (SZF) and PHPHPPHPHPPHHPPPHPPPPHPHPHP (UZF).

Besides the sequential and conformational features, the energy spectra of protein-like models contains more direct information about the landscape related to folding. Based on the landscape theory, a Z-score-like quantity $Z = (\overline{E} - E_N)/\sigma_E$ is used to measure the separation between the native state and the denatured states on the landscape, where $\overline{E}$ and σ_E are the average and the variance of the corresponding energy spectrum, and E_N is the energy of the concerned model protein at its own native conformation. A model protein with a large Z-score would generally have an optimal foldability. In this work, a statistics on the Z-scores for all the model proteins is carried out. The probability density f_Z for the SZF and UZF to have a certain Z-score is calculated (as shown in Fig. 4). The results demonstrate that the SZF generally have larger Z-scores (as shown in Fig. 4). This

indicates that the SZF has more conformations with fewer contacts between H-type monomers and thus higher energy. This kind of variation of spectrum of two similar model proteins is exemplified as the distribution of the number of contacts between H-type monomers in the inset of Fig. 4. The energy spectrum for the SZF apparently has larger population at the region with fewer contacts between H-type monomers (namely smaller $N_{\rm HH}$) compared with that for the UZF even though their sequences are very similar (with only one monomer being different). This kind of characteristics of SZF makes the corresponding folding cooperative and efficient, which is consistent with the basic idea of the ZAM. The ZAM-compatible model proteins seem to be more consistent with the minimal frustration principle. These analyses outline another feature of the model proteins which may adopt the ZAM during their folding.

Through the prediction test for the ZAM-based method and the systematic comparisons for various model proteins, it could be concluded that the ZAM is not a unified mechanism for protein-like models. The sequences with weaker hydrophobicity, low complexity for native conformation, and optimal energy landscape would be more likely to fold correctly with the ZAM. These results outline the connections between the properties of protein-like chains and the ZAM, and may stimulate further consistent understanding for protein folding.

References

[1] Onuchic J N, Luthey-Schulten Z and Wolynes P G 1997 *Annu. Rev. Phys. Chem.* **48** 545
[2] Udgaonkar J B 2008 *Annu. Rev. Biophys.* **37** 489
[3] Zuo G H, Hu J and Fang H P 2007 *Chin. Phys. Lett.* **24** 2426
[4] Dagget V and Fersht A R 2002 *Trend Biochem. Sci.* **28** 18
[5] Barrick D 2009 *Phys. Biol.* **6** 015001
[6] Sheehan D 2009 *Physical Biochemistry: Principles and Applications* 2nd edn (New York: John Wiley & Sons Ltd)
[7] Levinthal C 1968 *J. Chim. Phys.* **65** 44
[8] Gianni S, GuyDosh N R, Khan F et al 2003 *Proc. Natl. Acad. Sci.* USA **100** 13286
[9] Gianni S, Geierhaas C D, Calosci N, Jemth P et al 2007 *Proc. Natl. Acad. Sci.* USA **104** 128
[10] Rose G D, Flaeming P J, Banavar J R and Maritan A 2006 *Proc. Natl. Acad. Sci.* USA **103** 16623
[11] Okzen S B, Wu G A, Chodera J D and Dill K A 2007 *Proc. Natl. Acad. Sci.* USA **104** 11987
[12] Voelz V A and Dill K A 2007 *Proteins* **66** 877
[13] Dill K A, Ozkan S B, Weikl T R, Chodera J D and Voelz V A 2007 *Curr. Opin. Struct. Biol.* **17** 342
[14] Dill K A, Ozkan S B, Shell M S and Weikl T R 2008 *Annu. Rev. Biophys.* **37** 289
[15] Shell M S, Ozkan S B, Voelz V, Wu G A and Dill K A 2009 *Biophys. J.* **96** 917
[16] Dill K A 1990 *Biochemistry* **29** 7133
[17] Shih C T, Su Z Y, Gwan J F, Hao B L, Hsieh C H and Lee H C 2000 *Phys. Rev. Lett.* **84** 386
[18] Chikenji G, Fujisuka Y and Takada S 2006 *Proc. Natl. Acad. Sci.* USA **103** 3141
[19] Hoque T, Chetty M and Satter A 2009 *J. Comp. Biol.* **16** 85
[20] Li H, Helling R, Tang C and Wingreen N 1996 *Science* **273** 666
[21] England J L and Shakhnovich E I 2003 *Phys. Rev. Lett.* **90** 218101
[22] Levy Y, Wolynes P G and Onuchic J N 2004 *Proc. Natl. Acad. Sci.* USA **101** 13786