Eur. Phys. J. E **32**, 103–107 (2010)
DOI 10.1140/epje/i2010-10591-5

THE EUROPEAN
PHYSICAL JOURNAL E

Regular Article

The role of secondary structure in protein structure selection

Yong-Yun Ji[1,a] and You-Quan Li[2]

[1] Department of Physics, Wenzhou University, Wenzhou 325035, P.R. China
[2] Department of Physics, Zhejiang University, Hangzhou 310027, P.R. China

Received 16 November 2009 and Received in final form 10 March 2010
Published online: 25 May 2010 – © EDP Sciences / Società Italiana di Fisica / Springer-Verlag 2010

Abstract. The presence of highly regular secondary structure motifs in protein structure is a fascinating area of study. The secondary structures play important roles in protein structure and protein folding. We investigate the folding properties of protein by introducing the effect of secondary structure elements. We observed the emergence of several structures with both large average energy gap and high designability. The dynamic study indicates that these structures are more foldable than those without the effect of secondary structures.

1 Introduction and motivation

The native conformation of proteins generally displays extremely regular motifs called secondary structure. Roughly 50% of the structure of all proteins is in some form of secondary structure [1]. There are two main kinds of regular, recurring secondary structure elements: α-helix and β-sheet. The secondary structure is an important level in the hierarchical classification of protein structure and it is used to identify protein features for fold recognition. Predicting the secondary structure of protein is an essential intermediate step on the way to predicting the three-dimensional structure. Researchers have tried to explain the emergence of this regular motif. Stimulating explanations, ranging from detailed chemistry to geometrical principles, have been invoked to rationalise the existence of secondary structures.

Pauling *et al.* first realised that proteins have secondary structures through the use of early crystallographic studies of proteins and made the brilliant deduction of the ability of an α helix promoting its stability with the correct accommodation of hydrogen bonds [2]. Since then, there have been many studies which attempt to understand the origins of secondary structures. Many studies attempted to explain the emergence of secondary structures only from geometrical principles rather than from detailed chemistry. Some works indicated that the compactness of protein is sufficient to create secondary structures [3,4]. Chan and Dill found that, as the compactness of the chains increased, so did the percentage of secondary structures present [5–7]. They also found that the maximal compact chains had roughly the same amount of secondary structures as real proteins and the proportion of helices to sheets was also approximately the same. This result was tested by Socci *et al.* with a simple non-lattice protein model [8]. The emergence of a helical order in compact structures was obtained by Maritan *et al.* They formulated a dynamical variational principle for selection in conformation space, based on the requirement that the backbone of the native state of biologically viable polymers could be rapidly accessible from the denatured state [9]. Helices are important and ubiquitous in biology because of the theorem that regularly assembled identical objects form a helix. Cahill *et al.* explained this theorem with a proof which is simple, direct, self-contained and has reachable implications for biology [10]. It has also been proposed that the α-helix may be an energetically favourable structure for main-chain atoms [3,11]. Despite the concerted effort of several groups, the simple general explanation remains elusive.

Some other works have been carried out to explore the role of secondary structure in protein structure and protein folding. Silva *et al.* [12] studied the protein folding rate with various contents of secondary-type proteins, and they found that the folding rate constants are largely influenced by topological details of the native structure. Kuwajima *et al.* suggested that the rapid formation of a secondary structure framework in protein folding is a common property observed in a variety of globular proteins [13]. The experiments of Myers *et al.* indicated that preorganisation of one or more elements of secondary structure in the unfolded protein is an important determinant of fast protein folding [14]. Their results are consistent with the results of a model study which investigated fast-folding proteins with a diffusion-collision mechanism.

There are many researchers who try to explore the origin of secondary structures or emphasise the important role that secondary structures play in protein folding. The

[a] e-mail: yyji@wzu.edu.cn

secondary structure is very important to protein in the evolution, size and geometry selection of the secondary structure motifs [10], but what role will it play in the selection of protein conformations? Banavar *et al.* [15] emphasised the parallel tendency of the structure elements in protein structure with lattice models and obtained many protein-like properties. More recently, the tube model has been widely used to study the protein folding. The results show that (a) the observed protein folds are determined by general considerations of symmetry and geometry, and (b) sequences and functionalities evolve within the fixed backdrop of these immutable folds [16–20]. These show us that it is significant to study this problem with a simplified model by introducing the effect of secondary structure elements. In this article, we try to obtain the answer by exhaustive emulation of the lattice model. By investigating the folding properties of proteins under the effect of secondary structures, we observed that there are some structures which emerge with both large average gaps and high designability. The dynamic study indicates that these structures are more foldable than those without the effect of secondary structures.

2 Models

For the analytical and computational simplicity, the lattice model and its extended ones have been widely used in studying essential properties of protein folding and evolution to date [15,21,22]. Li *et al.* [23] have introduced the designability concept to interpret the natural selection of protein structures with the HP lattice model firstly introduced by Dill *et al.* [21]. We have investigated the medium effects on the selection of sequences folding into stable proteins with the simple HP model and obtained some meaningful results [24]. We have also obtained the prion-like folding behavior of protein by using simple HP lattice model [25].

In order to investigate the folding behaviour of proteins with the effect of secondary structure elements, we reconstructed the original HP model by accounting for a much lower pair contact energy of amino acids when this contact is involved in a secondary structure element. The protein is figured as a cube formed by a chain of 27 beads occupying the discrete sites of a lattice in a self-avoiding way, with two types of beads: hydrophobic and polar (fig. 1(a)). The energy of the model protein is given by

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

where i, j, respectively, denote the labels of residues in a sequence, r_i denotes the position (of the i-th residue) on the lattice sites, and σ_i the H (P), corresponding to hydrophobic (polar) residue, respectively. Here the delta notation is adopted, for example, $\delta_{a,b} = 1$ if $a = b$ and $\delta_{a,b} = 0$ if $a \neq b$. The hydrophobic force drives the protein to fold into a compact shape with as much hydrophobic residues inside as possible [26], The H-H contacts are more favourable in this model, which can be characterised by choosing $E_{PP} = 0$, $E_{HP} = -1$, and $E_{HH} = -2.3$, as adopted in ref. [23]. γ indicates the enhancement of the pair contact energy while this contact is involved in a secondary structure element: α-helix or β-sheet. In our calculations, it ranges from 0.1 to 1. The δ_s is adopted as: $\delta_s = 1$ if a contact is involved in a secondary structure element and $\delta_s = 0$ for other contacts. It means that the formation of a secondary structure possesses much lower energy, which is consistent with the viewpoint that secondary structures are energetically favourable structures [3,11].

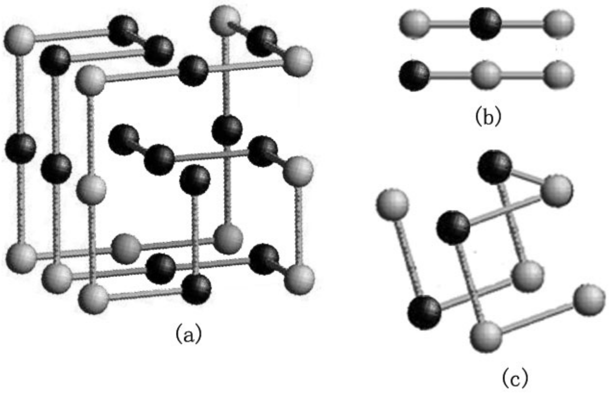

Fig. 1. (a) An example structure of HP model with hydrophobic residues (grey) and polar residues (black) and the secondary-structure-like elements: (b) sheet-like, (c) helix-like.

3 Results and discussion

We calculated and analysed the energy of all 2^{27} sequences among the maximally compact structures unrelated by symmetries. The maximally compact structures are analysed firstly to find out and denote the secondary-structure-like elements. There are two main kinds of secondary-structure-like elements, sheet-like (fig. 1(b)) and helix-like (fig. 1(c)), which are considered in our calculation. There are more sheet-like elements than helix-like elements in this simplified model. There are 38210 structures which include at least one sheet-like element, while only 4983 with helix-like elements.

By the exhaustive enumeration, there are 51940504 sequences which have a unique lowest energy state (these sequences, respectively, take only one structure as the lowest energy state) with the effect of secondary structure elements and $\gamma = 1$. This equals nearly 40% of all sequences which is much more than without the effect of secondary structure elements (4.6%). We analysed the Z-scores of some sequences which take some particular structures as the unique native state. The Z-score of a sequence, firstly introduced by Bowie *et al.* [27], is defined as the energy separation between the native fold and the average of an ensemble of misfolds in units of the standard deviation of the ensemble. It is a good parameter to use to measure the stability of the native structure for a protein sequence. The Z-score is a widely used and efficient optimisation method for sequence design or assessing energy

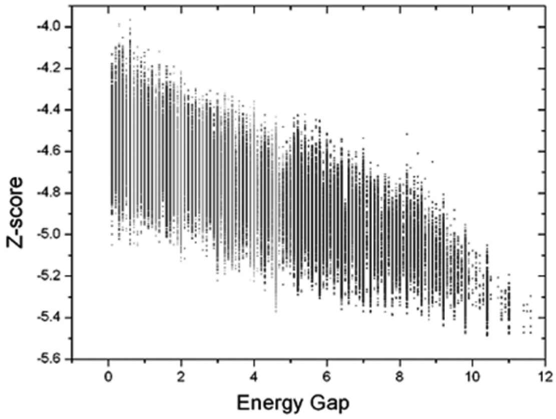

Fig. 2. The Z-score *versus* energy gap for sequences belonging to the structure with the largest average energy gap ($\gamma = 1$).

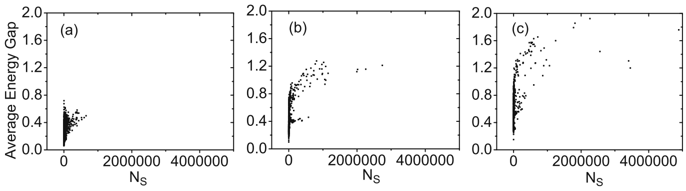

Fig. 3. The average energy gap *versus* N_s for different γ: (a) $\gamma = 0.1$, (b) $\gamma = 0.5$, (c) $\gamma = 1$.

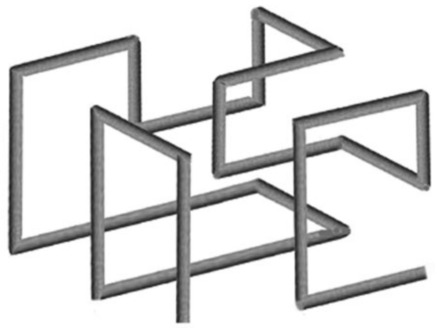

Fig. 4. The schematic structure with the highest designability ($\gamma = 1$).

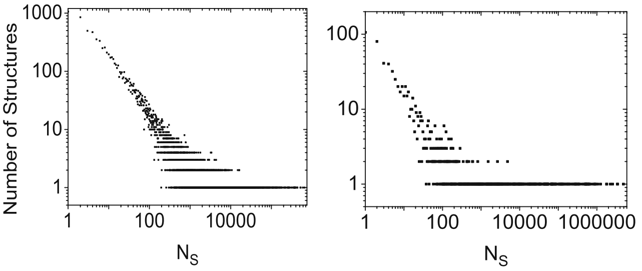

Fig. 5. The number of structures *versus* N_s for different γ's: $\gamma = 0.1$ (left), $\gamma = 1$ (right).

parameters in protein studies [28–30]. The Z-score is defined as

$$Z_{\text{nat}} = (E_{\text{nat}} - \langle E \rangle)/\sigma_E. \qquad (2)$$

E_{nat} is the energy of the native state of the sequence, $\langle E \rangle$ is the average energy of the sequence in all compact conformations and the σ_E is the standard deviation in energy for the entire compact ensemble. The native state is much more stable than the average member of a non-native ensemble (it is said, when the Z-score for the native state has a large negative value, the sequence is assumed to fold into the native state). This means that the better sequence holds smaller Z-scores. We calculate the Z-scores of the sequences belonging to a particular structure which possessed the largest average energy gap when $\gamma = 1$. The result shows that the sequences with a large energy gap hold small values of Z-scores (fig. 2). The smaller the Z-score is, the better the sequence will be. Thus a good sequence measured by an energy gap is also a good one measured by a Z-score. As show in fig. 2, on the whole, the Z-scores decrease with the increase of an energy gap.

Considering the effect of secondary structure elements, the average energy gap and designability of structures were analysed. The designability of a structure (indicated as N_s) is defined as the number of sequences which take the structure as the lowest energy state and the average energy gap is defined as the average of the energy gaps of these sequences. As show in fig. 3, when γ is large enough, some structures emerge with both large average energy gap and high designability, such as there are nine structures with average energy gaps larger than 1.0 and N_s larger than 1000000. There is a structure with an average gap of 1.92412 and N_s being 2260106, which is much larger than that in the case without the effect of secondary structures (where the highest designability is 3794). There are a large number of secondary-structure-like elements in these structures. There is a structure, with the highest designability ($N_s = 4891104$) and a large average energy gap ($\bar{\delta}_s = 1.7591$) in the case of $\gamma = 1$, which emerges with both the helix-like and the sheet-like elements (fig. 4). While γ is small, there is no structure with both high designability and large average energy gap. Two structures, one with the highest designable and the other with the largest average energy gap when $\gamma = 0$, are not designable again when $\gamma \neq 0$. Considering the effect of secondary structure elements, the number of structures, which have at least one sequence taking as unique lowest energy state, is much less than in the case without the effect of secondary structures. There are only 1426 structures while $\gamma = 1$, but there are about 47 thousands structures in the case of $\gamma = 0$. This number also decreases with γ increasing. The correlated number of structures is 15185 for $\gamma = 0.1$, 2056 for $\gamma = 0.5$ and 1426 for $\gamma = 1$. It implies that the effect of secondary structures promotes the selection of a small number of specific structures for protein from the large conformation space. The distribution of structure number with N_s for different γ's has also been analysed. The number of structures with a given N_s monotonically decreases with the increase of N_s, and it is similar for all parameters of γ, as shown in fig. 5. There is a large number of structures for small N_s, but as to large N_s, the number of structures is small, usually only one structure.

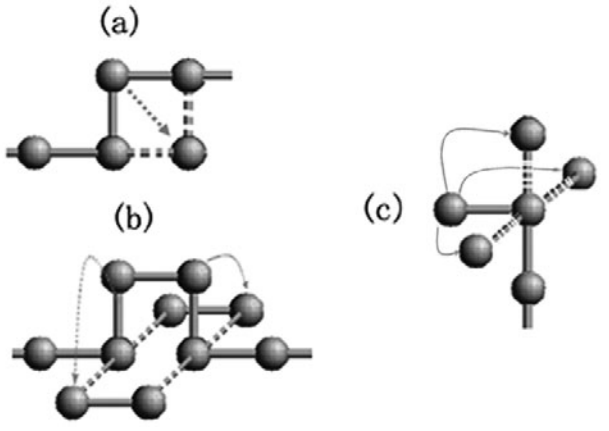

Fig. 6. The move sets: (a) corner move, (b) crankshaft move, (c) end move.

The dynamic properties of two structures, with the largest average energy gap in $\gamma = 1$ and $\gamma = 0$, are studied. The dynamic properties are calculated with the simple Go potential [31] with a simple lattice model. The Monte Carlo algorithm with a variety of move sets is also used. The move sets are corner moves, end moves and crankshaft moves, as Socci *et al.* selected [32] (fig. 6). The σ factor is calculated to describe the foldability of the model proteins [33]. σ is defined as $\sigma = |T_\theta - T_f|/T_\theta$, where T_θ is the collapse transition temperature, at which the chain changes its shape from a coil state to a compact form, and T_f is the folding transition temperature, at which the chain undergoes a first-order transition to the folded state. T_θ is obtained with the entropy sampling [34]. T_f is the peak temperature of the fluctuation $\delta\chi$ of a structure overlap function χ as defined in ref. [33]. χ is given as $\chi = 1 - \sum_{i \neq j, j\pm 1} \delta(r_{ij} - r_{ij}^N)/(N^2 - 3N + 2)$, where r_{ij}^N are the coordinates of the native state and δ is the Kronecker delta function. The $\delta\chi$ is defined as $\delta\chi = \langle\chi^2\rangle - \langle\chi\rangle^2$, where the angular brackets indicate thermodynamic averages. Figure 7 shows $\delta\chi$ *versus* the temperature and the entropy $S(E)$ *versus* the energy E of the two structures, with the largest average gap: one under the effect of secondary structures ($\gamma = 1$) (fig. 7(a), (c)) and one without the effect($\gamma = 0$) (fig. 7(b), (d)). By simulation, the σ factors of these two structures are obtained. We find that the σ of the structure without the effect of secondary structure is about 0.63 and it is larger than the one with the effect of secondary structures, which is 0.58. It implies that, according to the foldability, the structure selected with the effect of secondary structures is better than the one selected without the effect.

Our study indicates that the enhancement of the secondary structure effect can promote the selection of favourable structures for proteins. These structures possess more secondary-structure-like elements. The accessible formation of regular secondary structures sharply decreases the number of available conformations for protein, the peptide segments forming secondary structure hold fixed the ψ angle and the ϕ angle. The dynamic study shows that these structures selected with the effect of the secondary structures are more foldable than those without the effect. The secondary structure and the 3D structure selection are tightly related. Much better understanding of the secondary structure may be greatly helpful for our understanding and prediction of the three-dimensional structures of proteins. In fact, researchers have tried to deeply understand the secondary structure and made it as an intermediate step to the final destination: predicting the 3D structures of proteins. In our model, there are more sheet-like elements than helix-like, so taking a proper parameter for sheet and helix or importing more realistic protein-like models, such as the tube model, may give more instructive results.

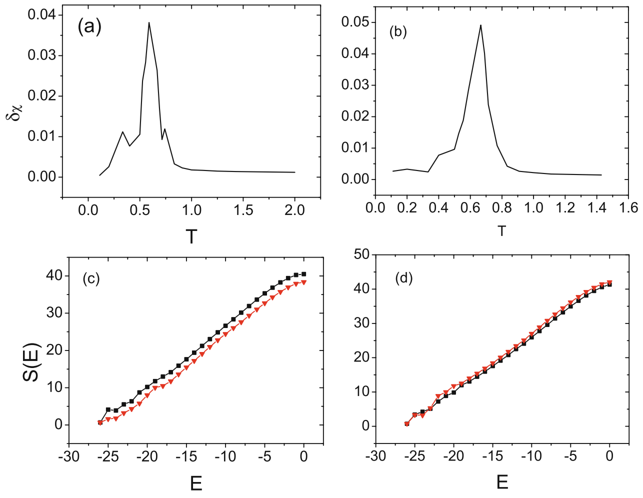

Fig. 7. $\delta\chi$ *versus* temperature of two specific structures: (a) the largest average gap one with $\gamma = 1$ and (b) the one with $\gamma = 0$. $S(E)$ *versus* E of these two structures: (c) the one with $\gamma = 1$ and (d) the one with $\gamma = 0$.

References

1. W. Kabsch, C. Sander, Biopolymer **22**, 2577 (1983).
2. L. Pauling, R.B. Corey, H.R. Branson, Proc. Natl. Acad. Sci. U.S.A. **37**, 205 (1951).
3. N.G. Hunt, L.M. Gregoret, F.E. Cohen, J. Mol. Biol. **241**, 214 (1994).
4. D.P. Yee, H.S. Chan, T.F. Havel, K.A. Dill, J. Mol. Biol. **241**, 557 (1994).
5. H.S. Chan, K.A. Dill, Macromolecules **22**, 4559 (1989).
6. H.S. Chan, K.A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 6388 (1990).
7. H.S. Chan, K.A. Dill, Annu. Rev. Biophys. Biochem. **20**, 447 (1991).
8. N.D. Socci, W.S. Bialek, J.N. Onuchic, Phys. Rev. E **49**, 3440 (1994).
9. A. Maritan, C. Micheletti, J.R. Banavar, Phys. Rev. Lett. **84**, 3009 (2000).
10. K. Cahill, Phys. Rev. E **72**, 062901 (2005).
11. R. Aurora, T.P. Creamer, R. Srinivasan, G.D. Rose, J. Mol. Biol. **272**, 1413 (1997).

12. I.R. Silva, L.M.D. Reis, A. Caliri, J. Chem. Phys. **123**, 154906 (2005).
13. K. Kuwajima, H. Yamaya, S. Miwa, S. Sugai, T. Nagamura, FEBS Lett. **221**, 115 (1987).
14. J.K. Myers, T.G. Oas, Nature Struct. Biol. **8**, 552 (2001).
15. J.R. Banavar, M. Cieplak, A. Maritan, Phys. Rev. Lett. **93**, 238101 (2004).
16. J.R. Banavar, A. Maritan, Annu. Rev. Biophys. Biomol. Struct. **36**, 261 (2007).
17. J.R. Banavar, A. Flammini, D. Marenduzzo, A. Maritan, A. Trovato, Complexus **1**, 4 (2003).
18. T.X. Hoang, L. Marsella, A. Trovato, P. Seno, J.R. Banavar, A. Maritan, Proc. Natl. Acad. Sci. USA **103**, 6883 (2006).
19. A. Maritan, C. Micheletti, A. Trovato, J.R. Banavar, Nature **406**, 287 (2000).
20. L.M. Luheshi, D.C. Crowther, C.M. Dobson, Curr. Opin. Chem. Biol. **12**, 25 (2008).
21. K.A. Dill, Biochemistry **24**, 1501 (1985).
22. G. Salvi, P. DeLosRios, Phys. Rev. Lett. **91**, 258102 (2003).
23. H. Li, R. Helling, C. Tang, N.S. Wingreen, Science **273**, 666 (1996).
24. Y.Q. Li, Y.Y. Ji, J.W. Mao, X.W. Tang, Phys. Rev. E **72**, 021904 (2005).
25. Y.Y. Ji, Y.Q. Li, J.W. Mao, X.W. Tang, Phys. Rev. E **72**, 041912 (2005).
26. W. Kauzmann, Adv. Protein Chem. **14**, 1 (1959).
27. J.U. Bowie, R. Luthy, D. Eisenberg, Science **253**, 164 (1991).
28. S. Park, X. Yang, J.G. Saven, Curr. Opin. Struct. Biol. **14**, 487 (2004).
29. L. Zhang, J. Skolnick, Protein Sci. **7**, 1201 (1998).
30. M.J. Sippl, J. Comput. Aid. Mol. Des. **7**, 473 (1993).
31. H. Taketomi, Y. Ueda, N. Go, Int. J. Pept. Protein Res. **7**, 445 (1975).
32. N.D. Socci, J.N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
33. D.K. Klimov, D. Thirumalai, Phys. Rev. Lett. **76**, 4070 (1996).
34. M.H. Hao, H.A. Scheraga, J. Phys. Chem. **98**, 4940 (1994).