Biochemical and Biophysical Research Communications 392 (2010) 95–99

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

# Water dynamics clue to key residues in protein folding

Meng Gao [a], Huaiqiu Zhu [a,*], Xin-Qiu Yao [a,b], Zhen-Su She [a,*]

[a] State Key Laboratory for Turbulence and Complex Systems, and Department of Biomedical Engineering, and Center for Theoretical Biology, and Center for Protein Science, Peking University, Beijing 100871, China
[b] Department of Biophysics, Kyoto University, Sakyo Kyoto 606-8502, Japan

ARTICLE INFO

*Article history:*
Received 16 December 2009
Available online 7 January 2010

*Keywords:*
Key residues
Water dynamics
Protein folding
Trp-cage

ABSTRACT

A computational method independent of experimental protein structure information is proposed to recognize key residues in protein folding, from the study of hydration water dynamics. Based on all-atom molecular dynamics simulation, two key residues are recognized with distinct water dynamical behavior in a folding process of the Trp-cage protein. The identified key residues are shown to play an essential role in both 3D structure and hydrophobic-induced collapse. With observations on hydration water dynamics around key residues, a dynamical pathway of folding can be interpreted.



## Introduction

With a series of experiments by residue mutation [1,2], it is now recognized that a few key residues in a protein sequence play key roles in the function, stability and folding of proteins. Meanwhile, many theoretical approaches have been proposed to investigate and even to predict key residues which are important in a structure of protein. Generally speaking, these theoretical studies can be summarized in two categories [3]. The first is based on sequential or structural alignments, under the argument that structurally or functionally important residues are highly conserved, which has been shown to depend on the known homolog information among proteins [3]. The second strategy aims at identifying key residues using contact or interaction energy evaluated by *ab initio* quantum chemical calculation [3], or studying on residues vibrational fluctuation pattern based on Gaussian network model or molecular dynamics (MD) simulation [3–5]. The latter mainly considers the key residues in folded protein structure stability analysis or in specific transition states of folding. However, there were few attempts to investigate key residues during a folding process, which are important to the understanding how proteins produce the correct 3D structure. A recent study by our group identified four key residues in the folding of the Trp-cage miniprotein (20 amino acid residues) using all-atom molecular dynamic simulation, which led to a proposal of key residue-dominated reconfiguration mechanism, in addition to spontaneous reconfiguration mechanism in protein folding [5]. However in that study, the experimental information, i.e. NMR structure, is needed as a reference to calculate a criteria function for key residues [5].

In most previous studies of key residues, a problem deserving more attention is the water–protein interaction. Until recently, the role of hydration water as an active component in the structure and the folding process of proteins has become accepted [6]. In fact, hydrophobic interactions have been shown to be the main driven force in protein folding [6]. More specifically, local hydration dynamics around key residues have cooperative effects with protein dynamics, and thus can provide information for protein structure and dynamics [7]. In a recent study by our group on wild type and mutant α-lytic protease differing by only one amino acid, we demonstrated that there are obvious distinctions in dynamic behaviors of hydration water [8]. Therefore, in such a hydrophobic-induced folding process, hydration dynamic behaviors is key to the understanding of water–protein interplay, and can be used to identify key residues.

In this paper, we present a computational method based on hydration water dynamics to recognize key residues in protein folding; the new method has the advantage of being independent of experimental protein structure information. Using data from all-atom MD simulations, two key residues are determined for the Trp-cage protein from their distinct water dynamical behaviors in the folding process. The key residues are shown to play an essential role in both structure and folding. With observations on hydration water dynamics around key residues, a dynamical pathway of folding can be interpreted.

## Materials and methods

*All-atom simulations.* All analyses here are based on data of 4-μs folding trajectories from our previous work [5] for the Trp-cage system (PDB entry 1l2y), which was simulated using the GROMACS

* Corresponding authors. Fax: +86 10 6276 7261.
*E-mail addresses:* hqzhu@pku.edu.cn (H. Zhu), she@pku.edu.cn (Z.-S. She).

0006-291X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2010.01.003

package [9] (Version 3.3) with the OPLS/AA force field and SPC water model. Starting from a partly unfolded configuration with randomly selected initial velocities, totally 40 simulations of the folding processes are performed in parallel at 282 K for 100 ns. The trajectories are saved every 10 ps, and then about 4,00,000 conformations are collected for analysis in this paper. There are 7 among 40 trajectories reaching the folded state.

*Coarse-grained Gō-model simulations*. To study the role of key residues in the folding kinetics of Trp-cage, coarse-grained simulations are performed with scaling of the interactions within a given residue pair or a contact group by a factor α. A similar strategy has been used in simulations of the coupled folding-binding process of intrinsically disordered proteins [10]. Herein, the Gō-model with coarse-grained $C_\alpha$ chain representation is used, and the potential model applied here is similar to that of the "without-solvation" model in [11]. Other parameters used here are set the same as in [11]. Langevin dynamics is used in dynamic simulations.

## Results and discussion

### *Determining key residues without NMR information*

To identify key residue, parameters such as RMSD and number of native contact have been used in previous studies. However, due to the need of a reference structure, the methods deploying those parameters are seriously dependent on the experimental information (e.g., NMR structure data). In the present study, we use the parameter *Rg*, radius of gyration (for all heavy atoms), to describe protein folding state, which can provide the fundamental information of protein collapse process without need for pre-known protein 3D structure. First of all, we classify all configurations collected in protein folding trajectory into 20 bins according to its *Rg* values. Then, taking *Rg* as a reaction coordinate, free energy may be calculated by

$$\Delta G_i = -k_B T \ln(Z_i) \quad (1)$$

where $i$ is the index of 20 bins, $Z_i$ is the probability of the system staying at the bin $i$, calculated by the percent of configurations in which *Rg* stays, and $k_B$ is Boltzmann constant, $T$ is Kelvin temperature, here taken to be 282 K.

Among the 20 bins characterized by *Rg*, the one where protein stayed with highest probability and hence with lowest free energy is defined as the reference set $C^*$. Herein the state $C^*$ means a stable state and protein in the state $C^*$ has collapsed structures. Then, the configurations with greater *Rg* are recognized as less collapsed, and are classified into four subsets by their *Rg* according to their free energy level. The subset classification schema is illustrated in Fig. 1A, and named from the most extended state to the most collapsed one with $C_1$, $C_2$, $C_3$, $C_4$, to $C^*$.

It has been argued that the hydration state of a protein shows the hydrophobic-driven protein collapse process [7]. Herein we define a parameter $N_{hw}$, the number of water molecules around a given residue side chain within a hydration shell (see below), to describe the hydration state of protein in folding process. By counting water molecules within the hydration shell around a given side chain along folding trajectory, then changes of $N_{hw}$ can be measured for all residues in all trajectories. For Trp-cage in the current study, the hydration shell is defined to be 0.55-nm-thick with an irregular shape which covers all heavy atoms (non-hydrogen atoms) surface side chain, so that there will be no more than two layer of water molecules in the hydration shell. If oxygen atom of a water molecule is within 0.55 nm from any heavy atom of a given side chain, it will be regarded as one water molecule around this residue's side chain. During a protein collapse, some other residues take the places of water molecules which initially surround the center residues, and $N_{hw}$ around the center residue then decreases. So, $N_{hw}$ should be a sensitive measure to the residue which is located at the core center during a collapse process. In other words, the large fluctuation of $N_{hw}$ will be an indicator of key residues in protein folding process, especially in hydrophobic-driven protein.

To measure the fluctuation of $N_{hw}$ in a protein collapse process, we define the sensitivity, $Sn_i(k)$, of given residue $k$ by

$$Sn_i(k) = \frac{\langle N_{hw}(k)\rangle_{C_i}}{\langle N_{hw}(k)\rangle_{C^*}} \quad (2)$$

here $C_i$ represents the set $C_1$, $C_2$, $C_3$ or $C_4$. Apparently, $Sn_i(k)$ shows the relative value of $N_{hw}$ around residue $k$ when protein stays at a non-native state $C_i$ relative to the near-native state $C^*$. Thus, when protein collapse from $C_1$ to $C_4$, some residues are packed more closely, and their *Sn* will decrease from a greater va-

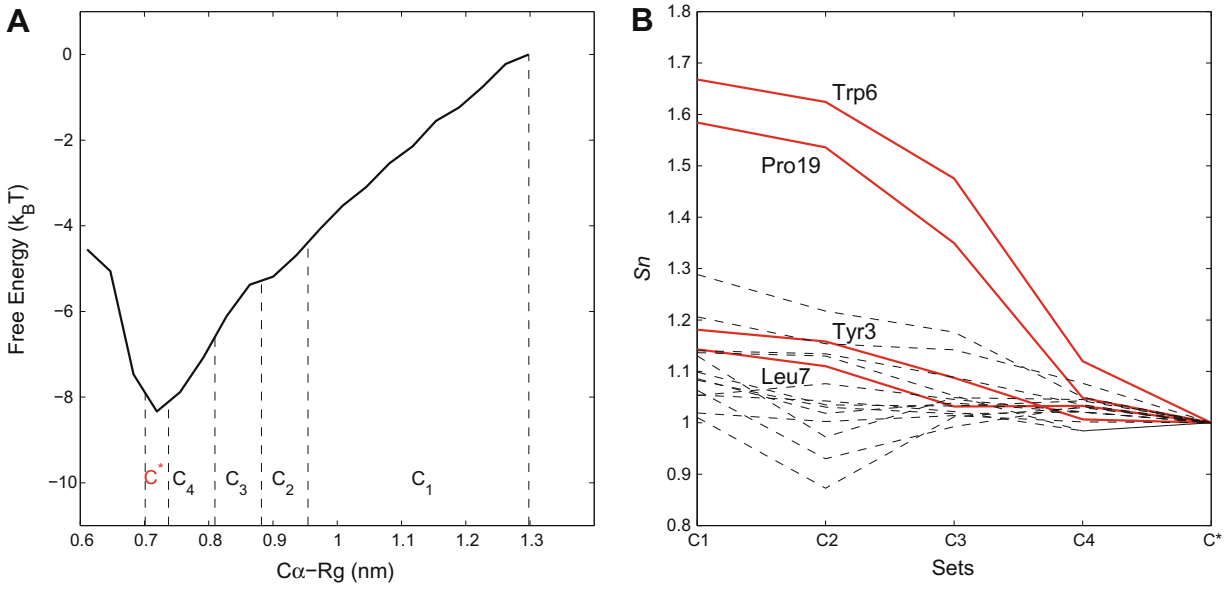


**Fig. 1.** (A) Free energy with Cα-*Rg* as a reaction coordinate. Here the maximum free energy is set to be 0 $k_BT$. $C_1$, $C_2$, $C_3$, $C_4$ is four defined subsets and $C^*$ is the reference state, which is the bin with minimum free energy (see text). (B) Sensitivity scores (*Sn*) of 17 residues (Gly 10, 11, 15 are not included for being lack of side chains) in Trp-cage. The four key residues (Trp6, Pro19, Tyr3 and Leu7) identified in Ref. [5] are labeled and plotted as solid lines; *Sn* for other residues are plotted as dash-dot lines.

lue to 1; whereas other residues show little change in their packing and their *Sn* will keep around 1. Residues packed closely in the folded state while exposed in the unfolded state are usually recognized as key residues in the folding process. Here, these key residues can be recognized easily by their *Sn*.

With the change of *Sn* along $C_1$, $C_2$, $C_3$ and $C_4$ in the current case, we can see that two residues, Trp6 and Pro19, among the 17 residues of Trp-cage, show distinct sensitivity to change of global protein collapse (as shown in Fig. 1B). More specifically, *Sn* of both Trp6 and Pro19 are significantly greater when protein is extended, and decrease quickly to 1 when protein collapses. That is to say, Trp6 and Pro19 are highly exposed with about 1.6–1.7 times water molecules than at stable state when Trp-cage is relatively loosely packed, while they are packed more closely and far from water after the protein collapse to a similar hydration state. On the contrary, *Sn* of other residues are about 1, showing few changes with protein collapse. So from Fig. 1B, one may draw a conclusion that change of water molecules around Trp6 and Pro19 shows a much closer correlation to Trp-cage folding process with the hydrophobic-induced collapse, thus these two residues should be identified as key residues in protein collapse process.

*Role of key residues in structure and folding dynamics*

The above results show that our computational method can identify key residues without any experimental structure information. We now investigate the role of identified key residues in both structure and folding of Trp-cage. It should be noticed that besides Trp6 and Pro19, two other residues, Tyr3 and Leu7, were also recognized as key residues in previously published work based on the analysis of side chain relaxation pattern [5], but do not show any remarkable difference from others in the present study (Fig. 1B). This suggests that residues Trp6 and Pro19 are involved in both the protein–water interaction process and side chain relaxation process, whereas Tyr3 and Leu7 only in the latter. Although all four residues are constituents of the hydrophobic core of the protein, they act differently. In fact, NMR experimental structure of Trp-cage has shown that Trp6 happens to be the center of the hydrophobic core, while Pro19 is one of the C-terminal $(Pro)_3$-unit coupling with Pro12 and Tyr3 to complete hydrophobic cluster [12]. Such a structure responsible for the stability of Trp-cage has been already noticed in previous work [5,12,13]; the present work provides a confirmation through computational analysis.

Furthermore, by examining the initial trajectories start from the partially unfolded structure, where Pro19 instead of Trp6 stays at the center, we found that, to achieve the correct configuration, C-terminal should bring Pro19 out of the center position, so as to loosen the wrong hydrophobic cluster and to put Trp6 back to the center position as a hydrophobic core. Then, Pro19 would pack on the upper face of the indole-ring of Trp6. During this process, repacking of Pro19 and Trp6 is a very important step for modifying the wrong hydrophobic cluster, and thus the two residues are the most sensitive ones to the global protein collapse. In fact, in all hydrophobically driven proteins, encapsulating of certain key residues are the main correlated events with the global protein dynamics; hence, the sensitivity measure related to the water dynamics may be of general interest. Moreover, the above scenario of exchanging of Pro19 and Trp6 during the folding process of Trp-cage further supports the argument that water–protein interplay is essential in leading to a native state in a hydrophobic-induced folding process.

Note that in the above definition of the sensitivity, we have used all trajectories to define the free energy other than part of them folded to native state, due to the premise that the method needs not a reference native structure. To show this treatment does not depend on folded trajectories and thus is robust, we use a subset of trajectories classified as not reaching a folded state by Yao and She. [5] to calculate *Sn* for each residue. The same strategy as above is used here to classify the sets $C_1$, $C_2$, $C_3$, $C_4$, and $C^*$, except that configurations used here are all produced from unfolded trajectories. The results present almost exactly the same trend as Fig. 1B, with Trp6 and Pro19 significantly more sensitive than others (data not shown). Therefore, even when protein does not reach the folded state, our method still works well on recognizing key residues in folding dynamics. Moreover, it indicates a direction in which some residues are gradually packed in protein collapse process rather than only in the final folded state, so these residues are important in the whole protein collapse process, while are identified as key residues here.

To study the role of identified key residues during the folding, herein we employ the Gō-model by mutating contact energy between residues. For a given residue, we use two mutation strategies by (i) adjusting contact energies of all pairs formed by the residue with others, and (ii) adjusting contact energy of single pair formed by the residue with another one. The former strategy aims to show how given residue acts in the whole protein structure and folding rate, while the latter compare effect of different contact pairs separately. As illustrated in method, here $\alpha$ is used to quantify the change of the contact energies: $\alpha = 0.5$ is a decrease by half and $\alpha = 2$ is a twofold increase. At each $\alpha$, 400 trajectories are created to get an average folding time $t_s$ (for $\alpha = 0.5$) or $t_q$ (for $\alpha = 2$). Then, a ratio $t_s/t_q$ is used to measure the effect of mutation on folding dynamics; a significant change of $t_s/t_q$ implies a big effect of a particular residue or residue-pair during the protein folding. The computational results show that the mutation on contact energies with Trp6 indeed leads to the biggest change of folding time. As shown in Fig. 2A, increasing on all contacts energies with Trp6 have a remarkable effect on folding time ratio ($t_s/t_q > 3$). Mutation on contact groups with Pro19 also shows a similar increase ($t_s/t_q > 1.6$), although not as significant as Trp6. This may be partially due to the fact that Pro19 has fewer contacts than Trp6. On the other hand, mutations on contact group with other residues, such as Leu2, Tyr3, Asp9, Gly11, Arg16, Pro17, Pro18, also have some effect on folding time. Our explanation is that Trp6 or Pro19 forms a native contact network with a certain number of residues, so an adjusting of the contact energy with Trp6 or Pro19 alone may bring remarkable change on folding time (Fig. 2B). The present calculation of the folding time ratio implies an effective way to reveal such a network, and the heterogeneousness of the peptide in which 20 amino acid residues do not play equal role in its folding. It is clear that Trp6 and Pro19 play more important roles in the folding dynamics of the Trp-cage protein.

*Describing folding process by key residue Trp6*

We then take Trp6 as an example to show how water dynamics around key residues can form an effective description for the dynamical steps of Trp-cage folding. Two key-residue-based order parameters, the number of water around Trp6 denoted by $N_{hw\text{-}6}$ and the side chain dihedral angle of Trp6 by Trp6-$\chi_2$, are used here as reaction coordinates to calculate free energy landscape. In the region $70^\circ <$ Trp6-$\chi_2 < 90^\circ$ on free energy landscape (Fig. 3), which has been recognized by our previous work as native states for Trp6-$\chi_2$ [5], we found two stable states, I and II, which correspond to a fewer and a greater number of water around Trp6, and which illustrate the packed state and exposed state, respectively (Fig. 3). In fact, as observed on simulation data, Pro19 closely packs with Trp6 in the state I, whereas Pro19 leaves Trp6 so that Trp6 is exposed to water in the state II. To judge the extent to which $N_{hw\text{-}6}$ and Trp6-$\chi_2$ describe folding events, herein we employ the folding sets determined in previous work by using native structure as reference [5]. Interestingly, configurations corresponding to

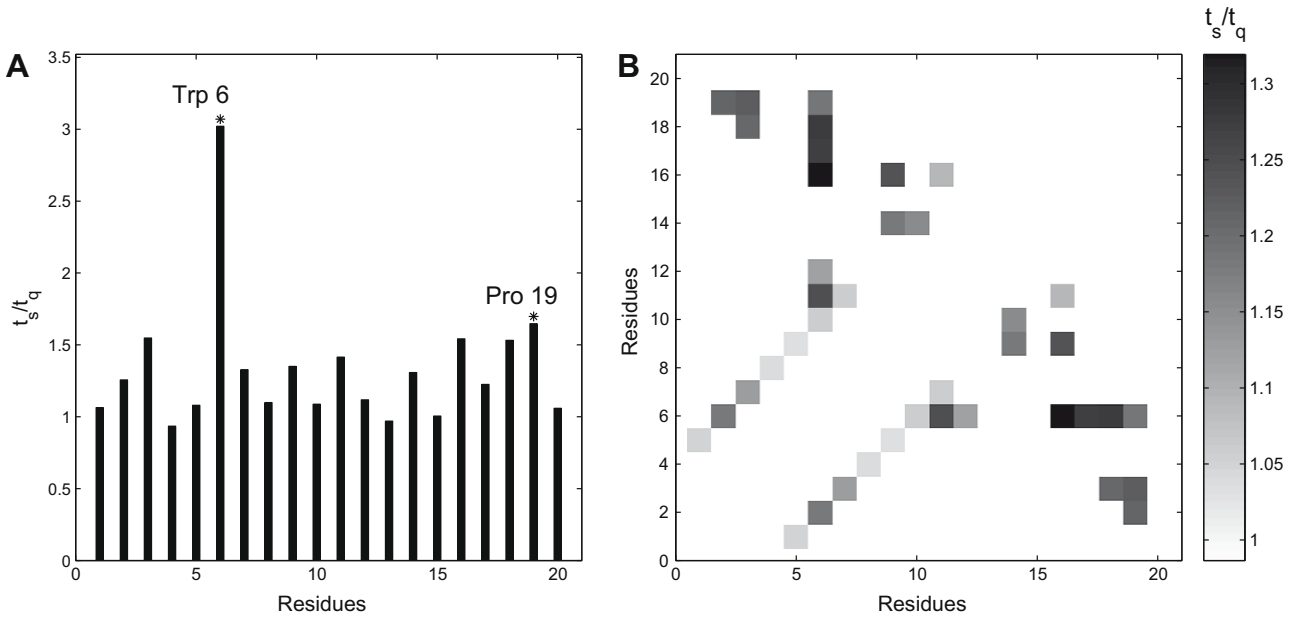


**Fig. 2.** (A) Ratio of folding time ($t_s/t_q$) after adjusting all pairwise contact energy of a particular residue. (B) Ratio of folding time ($t_s/t_q$) after adjusting contact energy between one pair of residues. Darker color represents greater $t_s/t_q$.

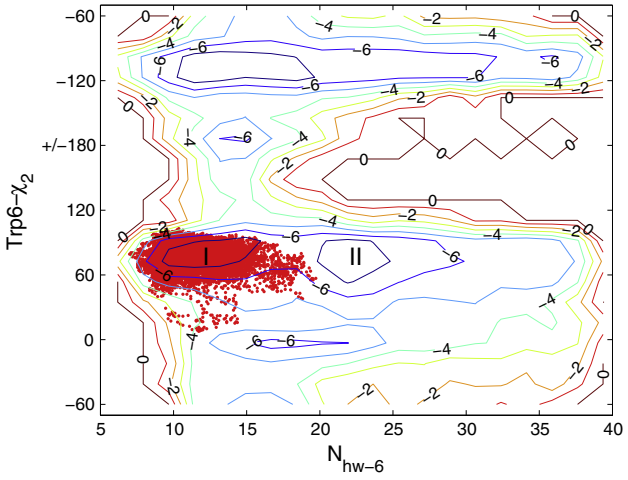


**Fig. 3.** Free energy landscape with Trp6-$\chi_2$ and $N_{hw-6}$ as reaction coordinates. Trp6-$\chi_2$ and $N_{hw-6}$ are averaged from 40 trajectories. When Trp6-$\chi_2$ is in (−50°, 100°), there are two stable states (i.e., I and II). Red points indicate snapshots of folding states of 40 trajectories. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

folding states spread at the region of state I (Fig. 3), showing that Trp6-$\chi_2$ and $N_{hw-6}$ act as good reaction coordinates.

Using a typical folding trajectory, a dynamical pathway of folding can be interpreted through free energy landscape (Fig. 4A). At first, the protein stays at a state around -180° < Trp6-$\chi_2$<−150° and 10 < $N_{hw-6}$ < 15; then a rapid conformational change happens with Trp6-$\chi_2$ turning to (50°, 100°) (close to state II), resulting in Trp6's exposure to water. During this process, Trp6 and Pro19 are separated for dozens of picoseconds, leaving a possibility for water molecules to move towards Trp6. Immediately after, there is a decrease of $N_{hw-6}$ to (8, 12), and a slight change of Trp6-$\chi_2$. So, the protein reaches the state I from the initial state in two steps: a first step of great conformational change of Trp6-$\chi_2$ and a second step of a remarkable change of $N_{hw-6}$. It is obvious that the interplay between water molecules and side chains of key residues accomplishes the repacking of hydrophobic cluster; this explains why Trp6 and Pro19 are identified as key residues in both two different strategies based on side chain relaxation [5] and water dynamics in the present study. After the state I, the packing between key residues Trp6 and Pro19 opens again and Trp-cage reaches the state II and stays there for about 20 ns, then Pro19 repacks with Trp6 and the protein returns to the state I (Fig. 4B). The free energy barrier of the transition between the state I and II is very low (~1·$k_BT$), thus, the initial state does not lead to the state I directly, but via an intermediate state around the state II. This is because at the initial state, limited by Trp6-$\chi_2$, it is difficult to destroy the wrong packing between Trp6 and Pro19. When Trp6-$\chi_2$ reaches (50°, 100°), protein oscillates between the state I and II at a low free energy barrier, which shows increasing flexibility. A limited space of Trp6 side chain provides the possibility that protein moves between hydrophobic collapse and solvation configurations at low free energy cost. So the correct packing between key residues Trp6 and Pro19 becomes possible. The observation on the transition between the two states provides support for the argument that the native state of a protein is dynamic rather than static [14]. In summary, using properties of key residues as reaction coordination, we are able to show important dynamical steps on free energy landscape and demonstrate the effect of key residues and hydration water on global protein dynamics.

## Conclusion

It has been known that a small number of key residues are sufficient for understanding the structure, stability and folding of a protein. However, the challenge is to develop an effective method for determining key residues, in view of the state-of-the-art approaches mainly reckoning on known homolog or structural information. In this paper, we present a computational method to recognize key residues in protein folding by studying hydration water dynamics. The method has an advantage in being independent of experimental protein structure information. The results demonstrate how capable the hydration water dynamics describes protein folding process, and reveal the heterogeneousness of a given peptide in which amino acid residues do not play equal role in its folding. Taken the Trp-cage protein as an illustration, two predicted key residues are an essential part of the hydrophobic

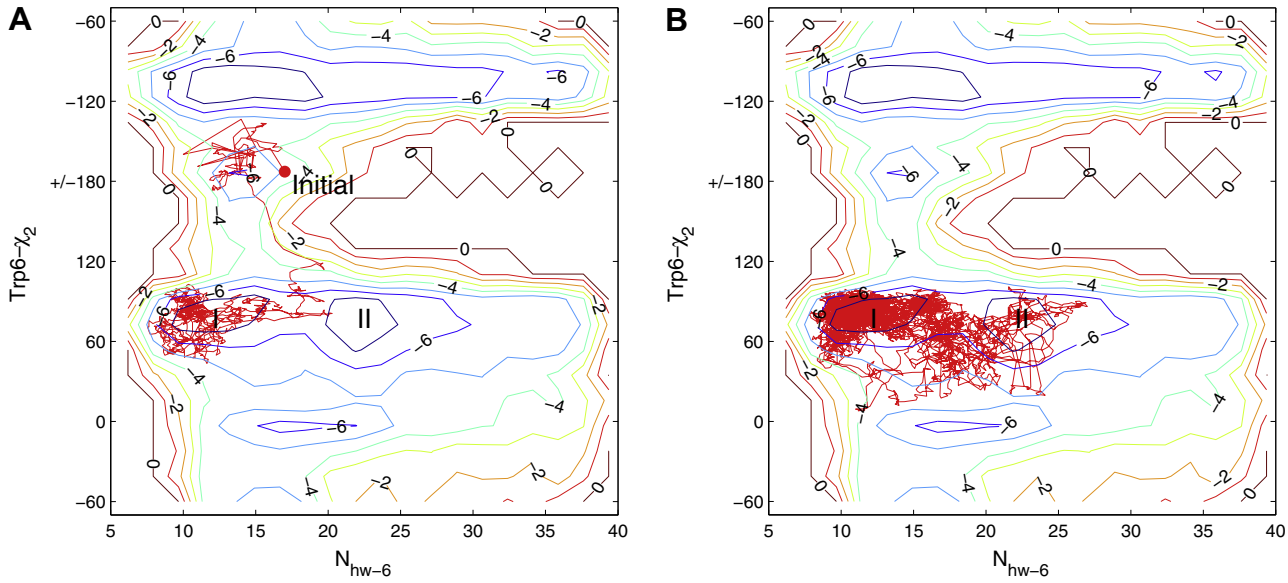


**Fig. 4.** (A) Configurations of the beginning 16 ns of a typical trajectory (the 7th) are projected on free energy landscape (red line). Trp-cage folds into its native state (I) from an initial partial unfolded configuration (labeled with a red point). (B) After reaching folded state (16–100 ns), Trp-cage partly unfolded again, and evolved between the two states, I and II. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

core and form nontrivial contacts with other residues, and are shown to have quite distinct water dynamics in the folding process by hydrophobic-induced collapse. Moreover, these two key residues are important hydrophobic residues that induce protein collapse. We further show important transition steps via free energy landscape, which demonstrates the interplay between key residues and hydration water strongly correlated with global protein dynamics.

Determining how a protein folds is a central problem in structural biology. However, it will be a formidable task in MD simulation to trace every residue in protein folding process even for a miniprotein like Trp-cage with 20 amino acids. The heterogeneousness of dynamics along the peptide implies that the residues playing important role on folding should be paid more attention to in sufficient details, whereas those less influence may be simplified in a way. That is to say, key residues acting in such important roles in protein folding may facilitate understanding protein folding mechanism by offering a reduced key-residue-based phase space analysis [15]. Therefore it is very important to determine key residues in folding for a protein, while the present study suggests a possible prospective solution to it. Above all, by investigating hydration dynamics, key residues can be identified without any pre-known experimental structure information. The MD simulation may be performed starting from a compact or stretched denature state of protein. With an ensemble average over all trajectories in simulation, key residues have distinct pattern of hydration water dynamics around their side chains. It is worth noting that even when those trajectories do not lead to a native state, our method still works well on recognizing key residues in folding dynamics. That means the simulation does not have to be performed for a long working time to the native state for protein. Finally, with predicted key residues, our work suggests a possible new route to design a solvent scheme: the explicit solvent model is used for key residues while the implicit solvent model for other residues. That seems to be very helpful to design a smart and faster MD simulation scheme.

## Acknowledgments

We thank Prof. Zhirong Liu and Yongqi Huang of Peking University for providing the data of Gō-model simulations, also thank Changsheng Zhang, Gang-Qing Hu, and Xiaobin Zheng of Peking University for beneficial discussions and helps to the work. The work received partial support by the 973 project grant 2009CB724100 and was also supported by the National Natural Science Foundation (30970667, 30770499 and 10721403) and MOST Project 2009ZX09501-002 of China.

## References

[1] M. Vendruscolo, E. Paci, C.M. Dobson, M. Karplus, Three key residues form a critical contact network in a protein folding transition state, Nature 409 (2001) 641–645.
[2] A.R. Fersht, Nucleation mechanisms in protein folding, Curr. Opin. Struct. Biol. 7 (1997) 3–9.
[3] L. Bendová-Biedermannová, P. Hobza, J. Vondrášek, Identifying stabilizing key residues in proteins using interresidue interaction energy matrix, Proteins 72 (2008) 402–413.
[4] C.J. Chen, L. Li, Y. Xiao, Identification of key residues in proteins by using their physical characters, Phys. Rev. E 73 (2006) 041926.
[5] X.Q. Yao, Z.S. She, Key residue-dominated protein folding dynamics, Biochem. Biophys. Res. Commun. 373 (2008) 64–68.
[6] K.A. Dill, Dominant forces in protein folding, Biochemistry 29 (1990) 7133–7155.
[7] L. Zhang, Y. Yang, Y.T. Kao, L. Wang, D. Zhong, Protein hydration dynamics and molecular mechanism of coupled water–protein fluctuations, J. Am. Chem. Soc. 131 (2009) 10677–10691.
[8] H. Kang, X.Q. Yao, Z.S. She, H. Zhu, Water–protein interplay reveals the specificity of alpha-lytic protease, Biochem. Biophys. Res. Commun. 385 (2009) 165–169.
[9] D.V.D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A.E. Mark, H.J.C. Berendsen, GROMACS: fast, flexible and free, J. Comp. Chem. 26 (2005) 1701–1718.
[10] Y.Q. Huang, Z.R. Liu, Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: a critical assessment of the "fly-casting" mechanism, J. Mol. Biol. 393 (2009) 1143–1159.
[11] Z. Liu, H.S. Chan, Solvation and desolvation effects in protein folding: native flexibility, kinetic cooperativity and enthalpic barriers under isostability conditions, Phys. Biol. 2 (2005) S75–S85.
[12] J.W. Neidigh, R.M. Fesinmeyer, N.H. Andersen, Designing a 20-residue protein, Nat. Struct. Biol. 9 (2002) 425–430.
[13] K.H. Mok, L.T. Kuhn, M. Goez, I.J. Day, J.C. Lin, N.H. Andersen, P.J. Hore, A pre-existing hydrophobic collapse in the unfolded state of an ultrafast folding protein, Nature 447 (2007) 106–109.
[14] R.G. Smock, L.M. Gierasch, Sending signals dynamically, Science 324 (2009) 198–203.
[15] A.B. Law, E.J. Fuentes, A.L. Lee, Conservation of side-chain dynamics within a protein family, J. Am. Chem. Soc. 131 (2009) 6322–6323.