Stabilities and Dynamics of Protein Folding Nuclei by Molecular Dynamics Simulation*

Yong-Shun Song (宋永顺),[1] Xin Zhou (周昕),[1] Wei-Mou Zheng (郑伟谋),[2] and Yan-Ting Wang (王延颋)[1,2,†]

[1]School of Physical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[2]Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100190, China

(Received April 17, 2017; revised manuscript received May 5, 2017)

Abstract *To understand how the stabilities of key nuclei fragments affect protein folding dynamics, we simulate by molecular dynamics (MD) simulation in aqueous solution four fragments cut out of a protein G, including one α-helix (seqB: KVFKQYAN), two β-turns (seqA: LNGKTLKG and seqC: YDDATKTF), and one β-strand (seqD: DGEWTYDD). The Markov State Model clustering method combined with the coarse-grained conformation letters method are employed to analyze the data sampled from 2-μs equilibrium MD simulation trajectories. We find that seqA and seqB have more stable structures than their native structures which become metastable when cut out of the protein structure. As expected, seqD alone is flexible and does not have a stable structure. Throughout our simulations, the native structure of seqC is stable but cannot be reached if starting from a structure other than the native one, implying a funnel-shape free energy landscape of seqC in aqueous solution. All the above results suggest that different nuclei have different formation dynamics during protein folding, which may have a major contribution to the hierarchy of protein folding dynamics.*

PACS numbers: 87.15.Cc, 87.15.A-, 87.15.ap, 87.15.Bg **DOI:** 10.1088/0253-6102/68/1/137

Key words: protein folding, molecular dynamics simulation, structure prediction

1 Introduction

Proteins form major components in cell, whose structures are essential for their biological functions. Although it is well-known that the three-dimensional (3D) structures of most globular proteins are encoded in their 1D sequences,[1] there still lacks a systematic way to determine the structure of a protein according to its sequence. Therefore, the number of known sequences is far larger than the number of known protein 3D structures.[2] The reliable prediction of the protein structure from sequence thus has a significant importance.

One way for predicting the protein structure from sequence is utilizing the known native structures of subsequences, which is known as "template-based modeling" methods. By assuming that similar sequences lead to similar structures, sequence search and alignment methods, such as Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST),[3] were developed. Thanks to the expanded Protein Data Bank (PDB) and better algorithms, the prediction accuracy of protein structures by template-based modeling has been improved substantially in the recent 20 years.[4–5] However, the improvements are mainly for small proteins no more than 100 residues, and the prediction accuracy is still less than 30% for larger and more complex proteins.[6–7] Further improvements based on template-based modeling methods are limited, especially for those proteins whose sub-sequences cannot be found in the PDB databases. Another strategy for protein structure prediction is the so-called "free modeling", which predicts the folded structure from scratch. The most successful implementation of free modeling is probably the "fragment assembly approach".[8–11] This method makes up the drawback of the template-based modeling by a "divide and conquer" strategy which divides the target protein sequence into several small fragments containing 8-12 residues[8,12] and searches for strong sequence signals in these fragments.

The central step of the fragment assembly approach is searching strong sequence signals for local geometries. Many sequence-structure libraries have been established in previous two decades, among which the most famous one is the bioinformatics-based I-sites library.[13] Usually the strong sequence signals of local geometries are secondary structural fragments, such as α-helix and β-turn. As the preferred structure is roughly estimated, the corresponding 3D structure of a strong sequence occupies a particular volume of the conformational space. Thus, it is very useful to have a coarse-grained description of the 3D structure containing secondary structure information. There are mainly three ways for coarse graining the 3D coordinates of the fragments: based on phase partition, based on representative points, or based on distribution in the phase space, with the third the most precise one. Clustering based on the distribution of the three angles in a tetrapeptide (one torsional and two bending angles) provides an alphabet of 17 conformation letters.[14–15] With the help of these conformation letters, we can identify the

*Supported by the National Basic Research Program of China under Grant No. 2013CB932804, the National Natural Science Foundation of China under Grant No. 11421063, and the CAS Biophysics Interdisciplinary Innovation Team Project

†E-mail: E-mail: wangyt@itp.ac.cn

conservative motifs in a protein by searching in the protein structure database.

Besides the bioinformatics-based methods mentioned above, physics-based models may also contribute to the sequence-structure library based on the assumption that peptide fragments of proteins have intrinsic propensities to form their native conformations, which has been tested by both experiments and simulations.[16–20] The most notable simulation test is the full range test done by Ken A. Dill *et al.*,[19–20] whose main conclusion is that, physics-based modeling (molecular dynamics (MD) simulation in their case) can be utilized to find the folding initiation sites.

Using a protein G as an example, Dill *et al.*[19–20] found eight fragments that show a preference to form their native structures and several strong sequence signals of local geometries serving as the folding initiation sites. However, they did not analyze in detail the stableness of those strong sequence signals, which may better help us identify them. Furthermore, we are interested in the following questions: why some sequences are more structured than others? are there any competing substates besides the native structures? what are those substates? what are the features of the free energy landscapes of the fragments in aqueous solution?

To tackle those questions, we carried out long-time equilibrium MD simulations on several fragments that have particular meanings as folding initiation sites. In the α-helix region, we select seqB (KVFKQYAN) because it has been identified by both the I-sites prediction and the conformation letters prediction but claimed unstable by Dill *et al.*[19] In the N-terminal β-hairpin region, we select seqA (LNGKTLKG) since it has been identified by Dill *et al.*[19] as a stable nucleus but not by both the bioinformatics-based methods. In the C-terminal β-hairpin region, we select seqC (YDDATKTF) as it has been identified by all the three methods and it is interesting to explore the stableness of seqC and find the difference between seqA and seqC. In addition, a native β-strand fragment (seqD: DGEWTYDD), which is expected to be amorphous when standing alone, has also been chosen for the purpose of comparison. The data analyses on our MD simulation results of the four fragments indicate that both seqA and seqB have a more stable structure than the native structure; seqD is flexible and does not have a stable structure; seqC can be stabilized on its native structure but difficult to reach the native structure if starting from an initial configuration different from the native structure. Those results suggest that fragments in key locations of protein structure have different stableness and play different roles in protein folding dynamics which may contribute differently to the hierarchy of protein folding dynamics.

2 Methods

2.1 *Simulation Methods*

The four 8-mer fragments mentioned above were cut out from the PDB ID 2GB1 protein structure[21] and capped with Ace and Nme. Their locations in the full protein G are shown in Fig. 1. For each fragment, the data were collected from totally 2-μs equilibrium simulations.

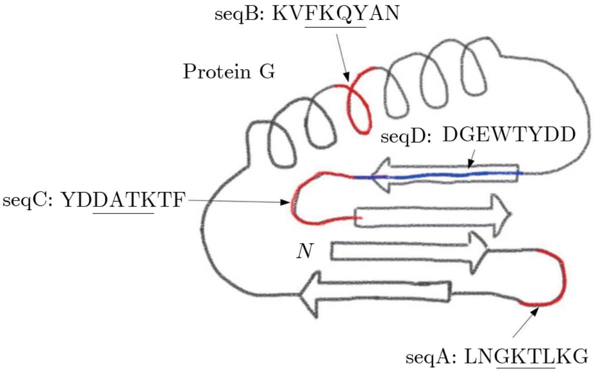

Fig. 1 (Color online) The four simulated fragments cut out from the protein G. For seqA, seqB, and seqC (colored by red), the underlined four-letters central sequences have specific structures generally supposed to serve as the folding nuclei during protein folding. The centers of seqA and seqC are β-turns and the center of seqB is a helical ring. A single β-strand fragment seqD (colored by blue) is also cut out and simulated for comparison.

MD simulations were performed with the GROMACS 4.6.2 simulation package.[22] The peptides were modeled by the OPLS-AA/L force field[23] because it had been widely used in MD simulations of β-hairpin folding,[24–25] and water molecules were modeled by the TIP4P explicit water model.[26] The particle mesh Ewald (PME) method[27] was adopted to handle the long-range electrostatic interactions. The NPT ensemble was realized by the Nosé-Hoover temperature coupling scheme[28] and the Parrinello-Rahman pressure coupling scheme.[29] For each peptide, 4000 solvent water molecules were added in the cubic simulation box. The simulation timestep was 2 fs and the instantaneous configurations were sampled every 100 ps. A 520-ns NPT simulation was carried out for each instance with the first 20-ns equilibration period discarded. For each peptide, four independent 500-ns instances were simulated, with two starting from the extended structure and the other two from the native structure, to check the initial structure dependency and examine the folding pathway.[30]

2.2 *Clustering of the Trajectories*

To analyze the conformational space of the fragments accessed by the MD simulations, we designed a clustering technique as follows. For each fragment, the total 2-μs trajectories were first divided into 100 coarse-grained microstates. These 100 microstates were further lumped into several metastable states. These metastable states were then analyzed to have a comprehensive understanding of the fragments.

A similarity score between different conformations should be defined to group the configurations in a trajectory into different microstates. The most commonly used one is the root-mean-square deviation (RMSD) between different samples. However, RMSD contains only structural information and cannot eliminate the possibility that two configurations have a small RMSD whereas a big energy barrier between them. To avoid this problem, one can first split the conformational space into several coarse-grained regions using the *a priori* knowledge of the conformational space distributions, and then group the configurations into those coarse-grained regions. This method greatly reduces the possibility that two energetically unaccessible configurations are grouped together. Another advantage of this method is that it is independent of the sample size. We chose the coarse-grained 17 conformation letters,[14] whose convenience also includes that it has a substitution matrix CLESUM that can be used to calculate the similarity score between different conformations letters. Each 8-mer in our simulations can be translated into a 5-letters string, since each conformational letter represents a structure formed by 4 residues. The similarity score between two conformations is defined as

$$S_{mn} = \sum_{i=1}^{5} w_i C(X_{ni}, X_{mi}) = \boldsymbol{wC}\,, \qquad (1)$$

where $C(X_{ni}, X_{mi})$ is the matrix element of (X_{ni}, X_{mi}) in CLESUM. The weight vector $\boldsymbol{w} = (1/10, 1/5, 2/5, 1/5, 1/10)$ for the 5-letters string allows the more centered ones contribute more to S_{mn}.

This similarity score is then incorporated into the Daura's clustering method,[31] which has been extensively used in protein structure analyses,[32–33] to group all the sampled structures into 100 microstates. Two strings are considered neighbors if their similarity score is less than 150. The clustering procedure takes the following steps. (a) The number of neighbors for each configuration is calculated according to the neighboring criterion. (b) The configuration that has the most neighbors is grouped along with its neighbors into one cluster. (c) The configurations in this cluster are deleted from the configurations pool. (d) The above steps are repeated for 100 times, which generates 100 microstates and associated 100 representative configurations of those microstates.

The Markov State Model (MSM) clustering method[34–35] is then employed to further adaptively lump those 100 microstates into several metastable substates. The MSM clustering method is on the basis of the hypothesis that the random transitions between adequately large metastable substates are less frequent than the transitions inside each metastable state. One of the routinely used software EMMA[36] implementing the MSM method was adopted to perform the clustering and related data analyses.

3 Results and Discussion

3.1 *Stable Regions Identified with PDB*

As mentioned above, conformation letters can be used to identify the stable regions of a protein according to its amino acid sequences. To find the stable regions in the protein G, we used PDB 3D structure library PDB_Select_25[37] to assess the structural conservativeness of a specific sequence. This library has a good balance between reducing the sequence redundancy and preserving the sequence diversity.

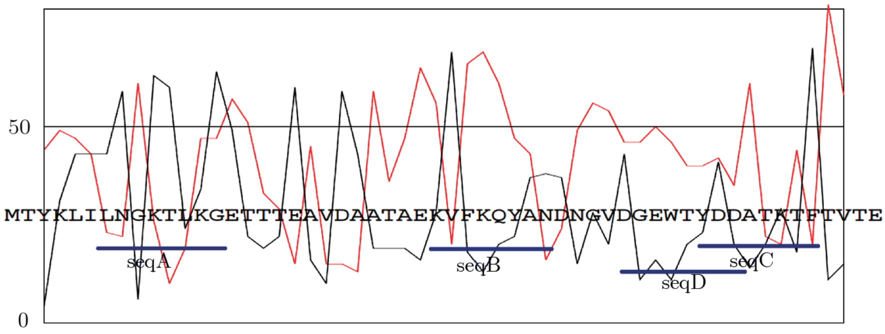

Fig. 2 (Color online) The values of N_1 (red curve) and N_2 (black curve) along the amino acid sequence of the protein G. N_1 denotes the number of sequences similar to A_1 for each 5-mer. N_2 denotes the number of sequences similar to A_k (the first sequence not similar to A_1) for each 5-mer. A larger N_1 and a smaller N_2 indicate that the corresponding sequence is more conservative. The locations of the four selected fragments are marked below the amino acid sequence.

Every 5-mer in protein G is set as a unit, corresponding to a coarse-grained 2-letters string of the conformation letters description. With a sliding window of 1, we can obtain all the 5-mers in the 56-residues sequence of the protein G. For each 5-mer, it is then compared with PDB_Select_25. The 100 most similar sequences in the library were kept and denoted as A_1, A_2, …, A_{100} from the most similar one to the least similar one. The corresponding 2-letters strings $\sigma(A_i)$ were also recorded.

All the 100 2-letters strings were compared with the first string $\sigma(A_1)$, and those with a structural similarity score higher than 0 were considered as similar. The number of sequences similar to A_1 was denoted as N_1. The first apparently different sequence was denoted as A_k. By the same rule, we calculated the number of sequences similar to A_k, denoted as N_2. The values of N_1 and N_2 along the residue sequence are shown in Fig. 2. Using the conformation letters and the PDB_Select_25 database, we then identified several conservative sequences in the protein G: FKQY is the most conservative component followed by DATK. These two regions are the central parts of seqB and seqC, consistent with our MD simulation results below.

3.2 *Stabilities of Fragments Studied by MD Simulations*

(i) ***α-helix Fragment seqB***

The simulation data for seqB were first analyzed by calculating the radius of gyration R_g ($\equiv \sqrt{(1/N)\sum_{k=1}^{N}(\boldsymbol{r}_k - \boldsymbol{r}_{\rm mean})^2}$, where N is the total number of atoms in the peptide, $\boldsymbol{r}_{\rm mean}$ is the mean position of the peptide) and the RMSD ($\equiv \sqrt{(1/N)\sum_{k=1}^{N}\delta_k^2}$, where δ_k is the distance between C_α atom k and its reference) with respect to its native conformation in the protein G. The probability map in the RMSD-R_g space for seqB is shown in Fig. 3. Although the configurations are clustered naturally in this map, it is still too rough to manually identify the substates. Therefore, we further applied the MSM method to the probability map to automatically identify the substates, which groups the configurations into four substates: the native-like substate BF1 (A/B/C/D stands for seqA/seqB/seqC/seqD, F/E stands for folded/extended), the half-folded substate BF2, another isolated stable substate BF3, and the extended substate BE, as shown in Figs. 3 and 4. The lifetime and portion of each substate is also shown in Fig. 3.

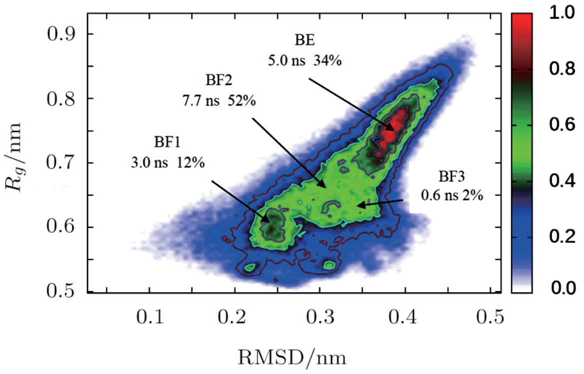

Fig. 3 (Color online) The RMSD-R_g probability map for seqB, marked with the names, lifetimes, and appearance probabilities of the four substates.

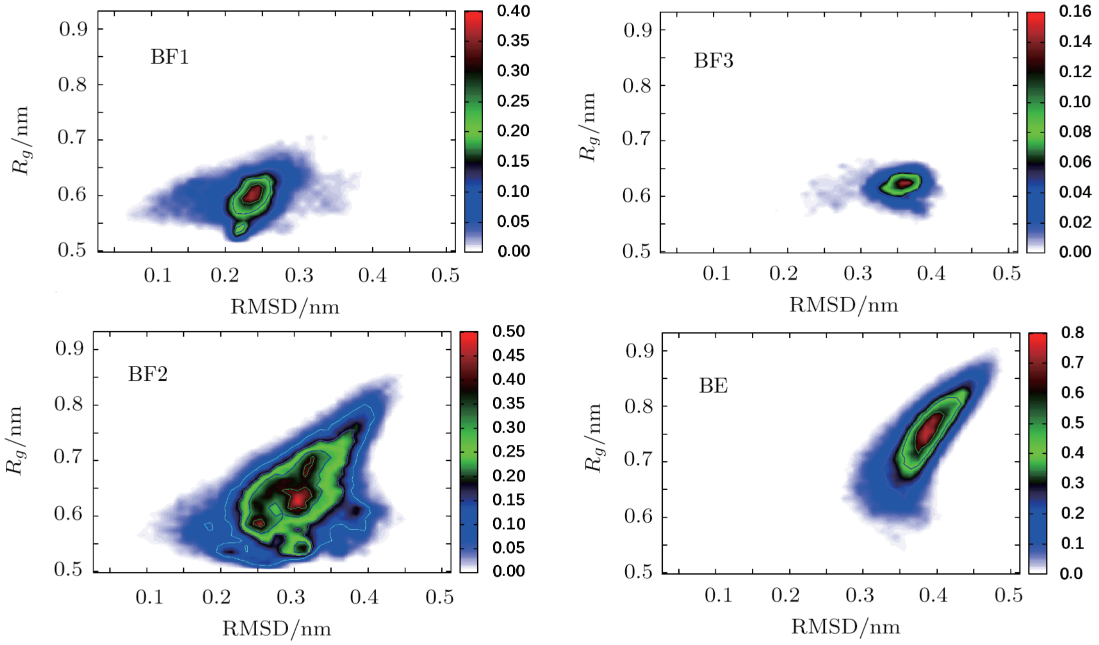

Fig. 4 (Color online) The four substates of seqB obtained by the MSM clustering method.

Figure 5 shows the native structure of seqB and the representative structure of substate BF1. We can see that the substate BF1 is very close to its native state. In the native structure, four typical α-type hydrogen bonds (H-bonds) ($i \rightarrow i+4$) are formed and two aromatic side chains are stretched out towards the same direction. The representative structure of BF1 is stabilized by three H-bonds: two (Val29:O-Gln32:NH and Phe30:O-Tyr33:NH) are 3_{10}-type H-bonds ($i \rightarrow i+3$) and one (Phe30:O-Ala34:NH) is a typical α-type H-bond. The contents of Val29:O-Gln32NH, Phe30:O-Tyr33:NH, and Phe30:O-Ala34:NH are 47.5%, 72.5% and 36.7%, respectively. The representative structure is the one with the largest appearance probability in BF1, while the native structure also appears with a smaller probability. In Fig. 6, we show a transition process between the representative structure of BF1 and the native structure observed during the MD simulation characterized by the formation of a particular H-bond between Val29:O and Tyr33:NH. The break of the H-bond between Val29:O and Gln32:NH is the precondition for the structural adjustment to form the H-bond Val29:O-Tyr33:NH. Thereafter, the break of the H-bond Phe30:O-Tyr33:NH indicates the completion of the transition from the 3_{10}-helix to the α-helix. The high content of the H-bond Phe30:O-Tyr33:NH indicates that there is an energy barrier for this structural transition.

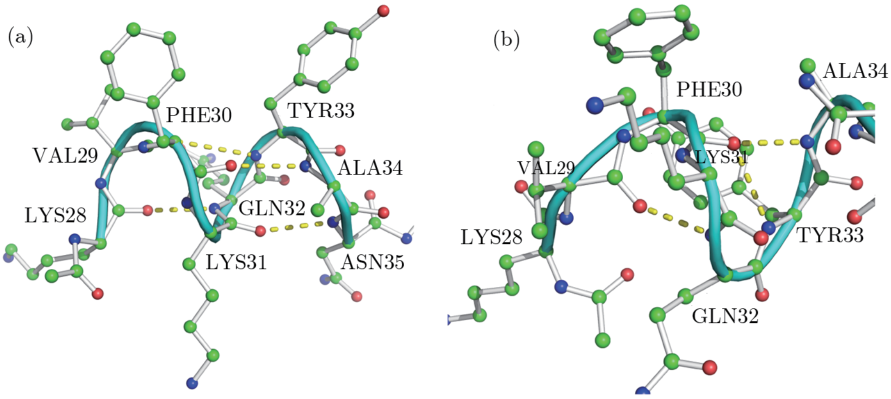

Fig. 5 (Color online) The native structure of seqB (a) and the representative structure of substate BF1 (b). H-bonds are depicted by dashed yellow lines.

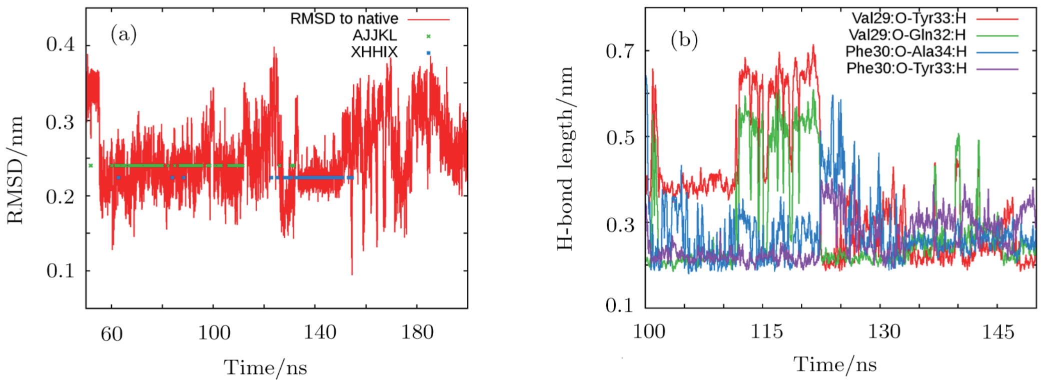

Fig. 6 (Color online) A transition process obtained in the MD simulation between the representative structure of substate BF1 and the native structure of seqB. (a) RMSD with respect to the native structure vs. time. Samples that represent the BF1 representative structure (5-letters string AJJKL) and native structure (5-letters string XHHIX, where X means an arbitrary conformations letter) are explicitly shown with green points and cyan points. The RMSD values of these points are set as the average value of all the corresponding sample points. The transition process (100 ns to 150 ns) are illustrated with more detail in (b), shown with the length of the four H-bonds that form or break during this process. Typically, the length of O-H is less than 0.2 nm indicates that an H-bond forms.[38−39]

Two other folded substates BF2 and BF3 besides BF1 have also been identified, whose representative structures are shown in Fig. 7. BF3 is stabilized by the H-bonds Phe30:O-Tyr33:NH and Lys31:O-Ala34:NH. Since it has a small probability (2%) and a large RMSD with respect to the native structure, it can be regarded as an unstable substate affecting little to the folding process of seqB. By contrast, BF2 has a large probability (52%) and a relatively small difference of RMSD comparing to BF1. Figure 7 shows that BF2 has a much looser structure than BF1 and thus includes several adjacent microstates. It is stabilized by two independent 3_{10}-type H-bonds: Lys28:O-Lys31:NH and Lys31:O-Ala34:NH with the contents of only 32.9% and 14.6%, respectively. Therefore, BF2 can be regarded as a collection of several intermediate metastable (or half-folded) substates towards the native structure of seqB.

Figure 3 shows that the lifetime of these metastable substates ranges from 0.6 ns to 7.7 ns, indicating that both the forming and breaking processes of this α-helix are very fast, consistent with the previous observation by Noé and Fischer.[35]

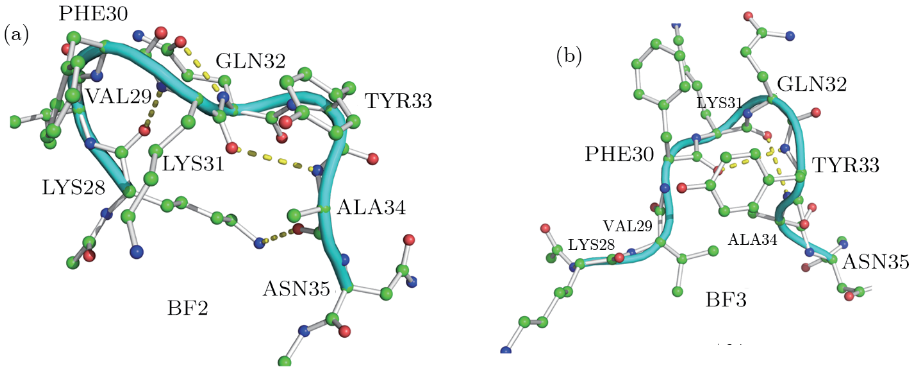

Fig. 7 (Color online) Representative structures of BF2 and BF3 for seqB. H-bonds are depicted by dashed yellow lines.

(ii) ***β-turn Fragment seqA***

The probability map of the RMSD-R_g space for seqA is shown in Fig. 8 and the substates clustered with the MSM method is shown in Fig. 9. All the configurations sampled from the MD simulations are grouped into three substates: the extended substate AE, the native-like substate AF1, and an alternative folded substate AF2. Representative structures of substates AF1 and AF2 are shown in Fig. 10.

As shown in Fig. 10(a), the representative structure of AF1 is close to its native structure. It is mainly stabilized by four H-bonds: the upper two (with Gly9 and Leu12 involved) and the lower two (with Leu7 and Gly14 involved). The forming probabilities of these hydrogen bonds were calculated to be 28.4% and 25.8% for the upper two (Gly9:NH-Leu12:O and Gly9:O-Leu12:NH), and 6.7% and 4.2% for the lower two (Leu7:O-Gly14:NH and Leu7:NH-Gly14:O), respectively. These numbers indicate that, in the native structure, the upper two H-bonds are more stable than the lower two.

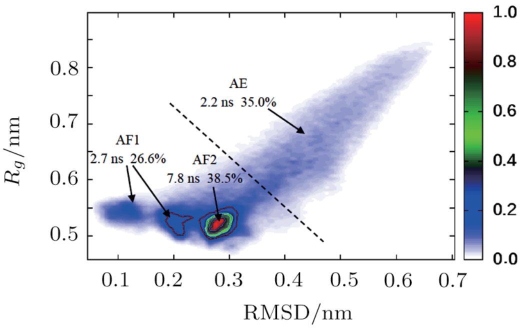

Fig. 8 (Color online) RMSD-R_g probability map for seqA, marked with the names, lifetimes, and appearance probabilities of the three substates.

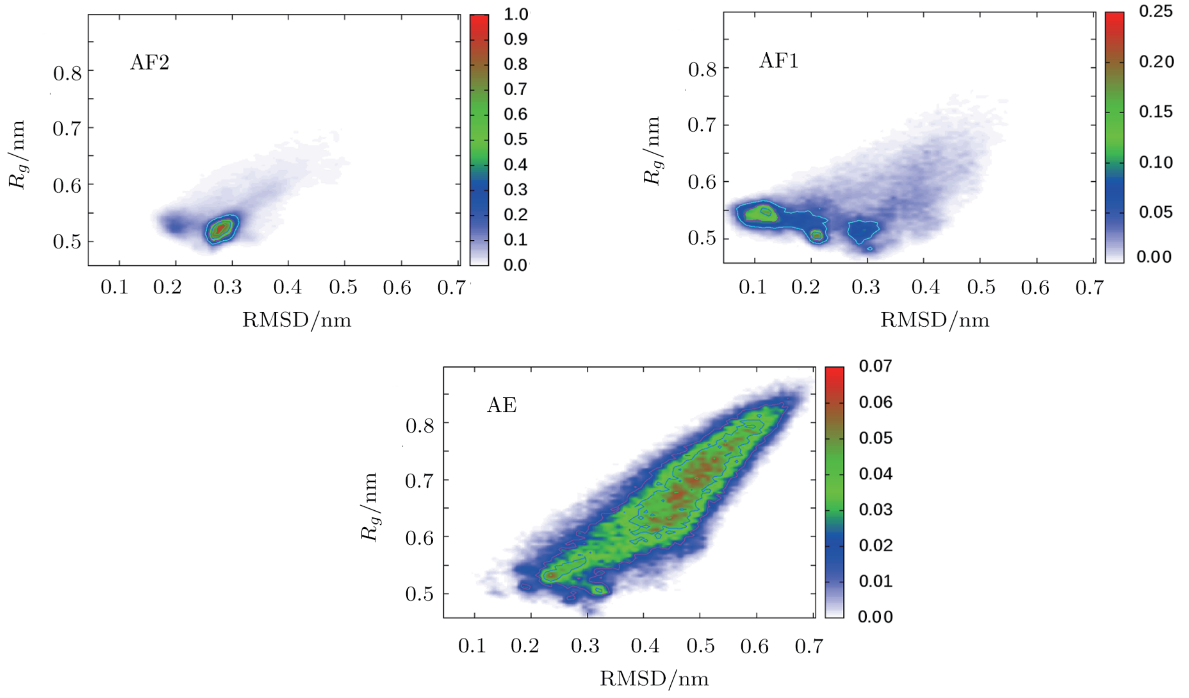

Fig. 9 (Color online) Substates of seqA determined by the MSM clustering method.

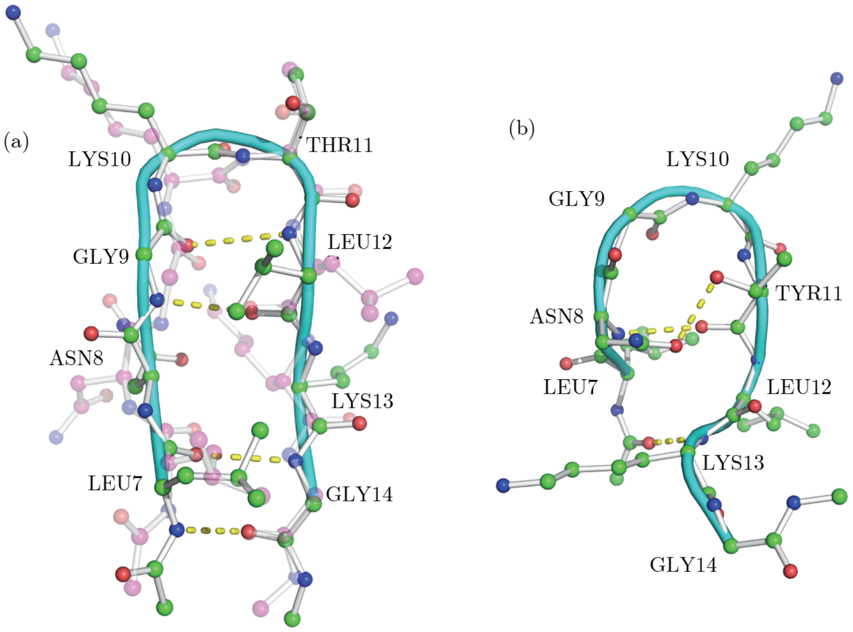

Fig. 10 (Color online) Representative structures of substates AF1 (a) and AF2 (b) for seqA. In (a), the native structure is shown as pink background and has the C_α atoms aligned with the representative structure of substate AF1. H-bonds are depicted by dashed yelow lines.

Because the content of AF2 is 38.5% and has a narrow RMSD deviation, it can be regarded as a dominant substate during the simulation and has a well-characterized structure. This can be further confirmed by the simulation result drawn in Fig. 11, which shows that seqA stays in AF2 for more than 100 ns in one of our four 500-ns simulations. The representative structure of AF2, shown in Fig. 10(b), is also a β-turn structure forming three H-bonds—Ace6:O-Lys13:NH, Asn8:NH-Thr11:O, and Asn8:OD-Thr11:OH with the content of 42%, 50% and 34%, respectively.

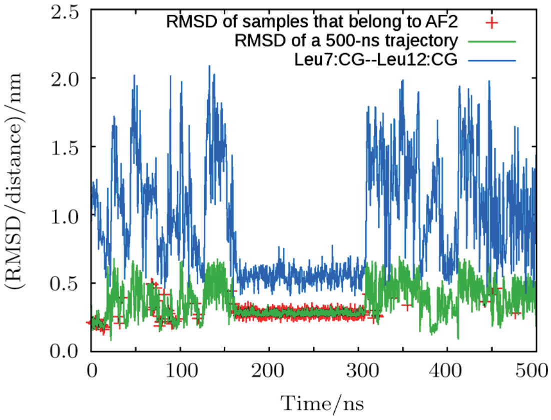

Fig. 11 (Color online) Time evolvement of the RMSD of all configurations in a 500-ns trajectory (green line) with respect to the native structure of seqA. The configurations belonging to AF2 (red plus sign), and the distance between Leu7:CG and Leu12:CG (blue line) are also shown.

The fact that AF1 (native-like substate) is less stable than AF2 can be understood by considering the following two reasons. First, seqA, whose amino acid sequence is LNGKTLKG, contains two Gly residues. Its native structure is stabilized by four H-bonds, two of which have Gly involved (Gly9:NH-Leu12:O and Leu7:O-Gly14:NH). Since Gly has a small side chain, in the short peptide it prefers to stay in the center region of the β-turn rather than form hydrogen bonds.[40] Second, each of the two Leus in this sequence has a hydrophobic side chain that do not like to reside in an aqueous environment. In the full protein G, these two side chains protrude to the inner side of the protein (shown in Fig. 12(a)) to effectively lower their free energy cost. When the peptide is cut out and isolated in water, the two side chains are pushed to approach to each other by the hydrophobic interaction to reduce the accessible area to water. The above two factors allow AF2 to be more stable than the native structure of seqA. The structures of AF1 and AF2 are further compared in Fig. 12(b) with the difference of the distance between Leu7:CG and Leu12:CG emphasized. Note that our results for seqA are different from Dill *et al.*,[19–20] who have concluded that the native structure is the most stable one. The difference might be attributed to the different models adopted in each simulation work. Since the explicit TIP4P solvent model we have used is more accurate than the GB/SA implicit solvent model they have used, we believe our results are more convincing than theirs.

As shown in Fig. 13, the native-like substate of seqA (AF1) can be stable for 30 ns, a little shorter than the 100-ns stable time of AF2. Therefore, AF2 is a major obstacle for seqA to fold to its native structure. Nevertheless, since the native structure is still metastable, we propose that seqA still serves as an early nucleus during the protein folding process of the protein G, which might have to spend a lot of time to go through the energy barrier between AF2 and AF1.

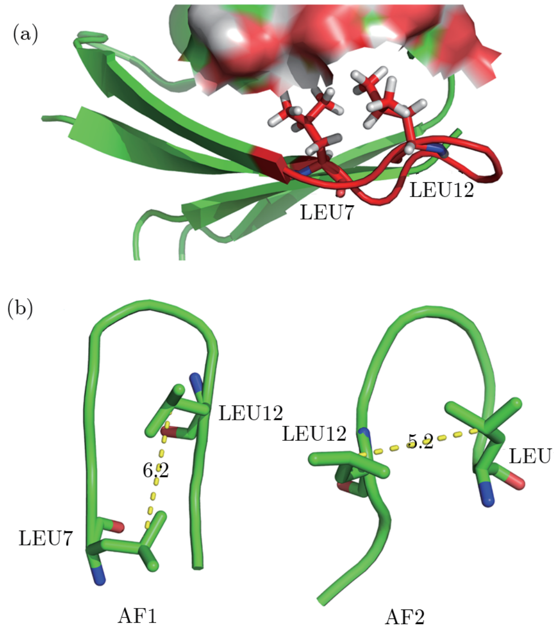

Fig. 12 (Color online) Locations of the hydrophobic side chains of Leu7 and Leu12 in seqA in the native structure of the protein G (a) and the representative configurations of AF1 and AF2 (b). The distances between Leu7:CG and Leu12:CG are marked by dashed yellow lines.

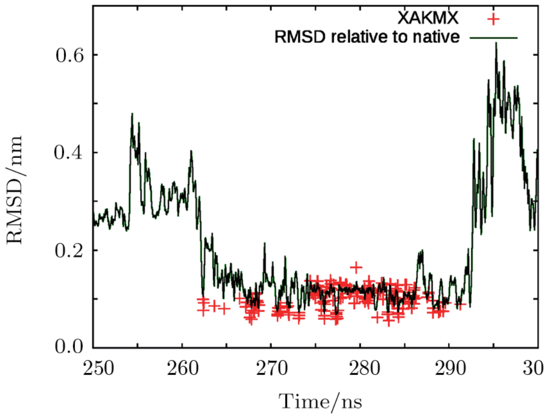

Fig. 13 (Color online) RMSD of seqA with respect to its native structure vs. time. The native structure represented by the 5-letters conformational string XAKMX (red plus signs) can last continuously for 30 ns.

(iii) *β-turn Fragment seqC*

As shown in Fig. 14, the native structure of seqC starting from the native structure is stable and does not transform into other structures during the 500-ns MD simulation. On the other hand, simulations starting from extended configurations can hardly visit the native structure. The RMSD-R_g probability maps shown in Fig. 15 indicate that there is a big gap between the conformation space explored by the simulations with the native initial structure and the extended initial structure. By applying the MSM clustering method, the RMSD-R_g space starting from the extended structure can be further split into three substates, whose appearance probabilities are 36.6% for CF1, 20% for CF2, and 43.4% for CF3 (Figs. 15(b) and 16). Their representative structures are illustrated in Fig. 17.

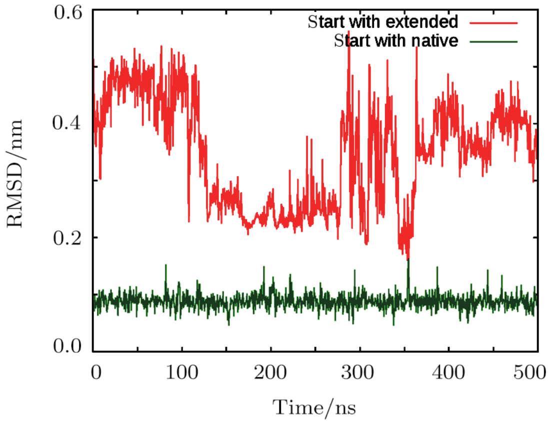

Fig. 14 (Color online) RMSDs with respect to the native structure for the MD simulations starting from the native structure and the extended structure of seqC.

After a long time simulation, the stable structure of seqC slightly deviates from its native structure, as shown in Fig. 18. Two extra H-bonds between Asp47:COO$^-$ and Lys50:NH$_3$ as well as between Tyr45:OH and Asp47:COO$^-$ are formed, and the H-bond between Asp46:O and Thr49:NH is replaced by Asp46:COO$^-$-Thr49:NH. In addition, Asp46:COO$^-$ forms H-bond with Thr49:OH, consistent with the results by Liao *et al.*[24]

Why is it difficult for seqC to reach the native structure starting from the extended conformation? We can see from Fig. 17 that, for the two most populated substates CF1 and CF3, it is not difficult to form the H-bonds in the β-turn central region Asp47-Ala48-Thr49-Lys50, as the contents for Asp47:O-Lys50:NH in CF3 and CF1 are 17.5% and 34.4%, respectively. Moreover, the RMSDs shown in Fig. 19 indicate that the central four residues Asp47-Ala48-Thr49-Lys50 form a structure similar to the native structure. We propose that the lack of hydrophobic interactions in the two ends of this sequence allows the formation of other structures, which is supported by the recent works on β-hairpin.[25,41–43] Actually, the C-terminal β-hairpin of the protein G (41-56 in the sequence), whose central region is seqC, has been extensively studied as a model β-hairpin [44–46] and found to be very stable in aqueous solution. There had been disputed on whether the hydrophobic region or the turn region forms first during the folding of this β-hairpin,[25,47] but the enhanced computational capacity has been providing more evidences to support that the turn forms first, followed by the hydrophobic region.[25,41] Our results also indirectly support the turn-centric folding mechanism for β-hairpin formation.

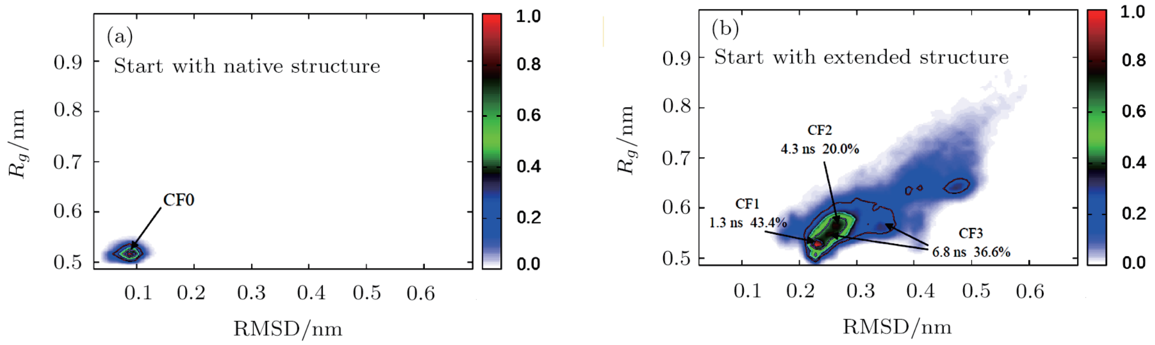

Fig. 15 (Color online) Probability map of RMSD-R_g space for seqC generated from the MD simulations starting from the native structure (a) and from the extended structure (b). The latter can be further split into three major substates: CF1, CF2 and CF3, shown with their names, lifetimes, and appearance probabilities.

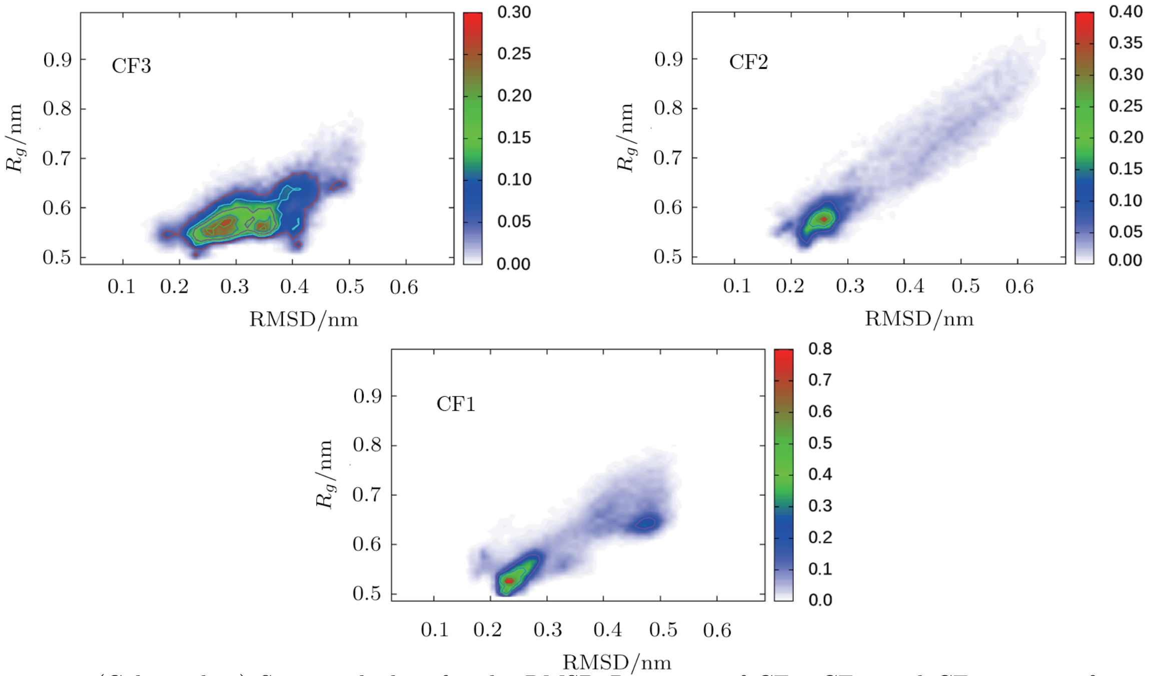

Fig. 16 (Color online) Separated plots for the RMSD-R_g spaces of CF1, CF2, and CF3 starting from the extended structure, obtained by MSM clustering method.

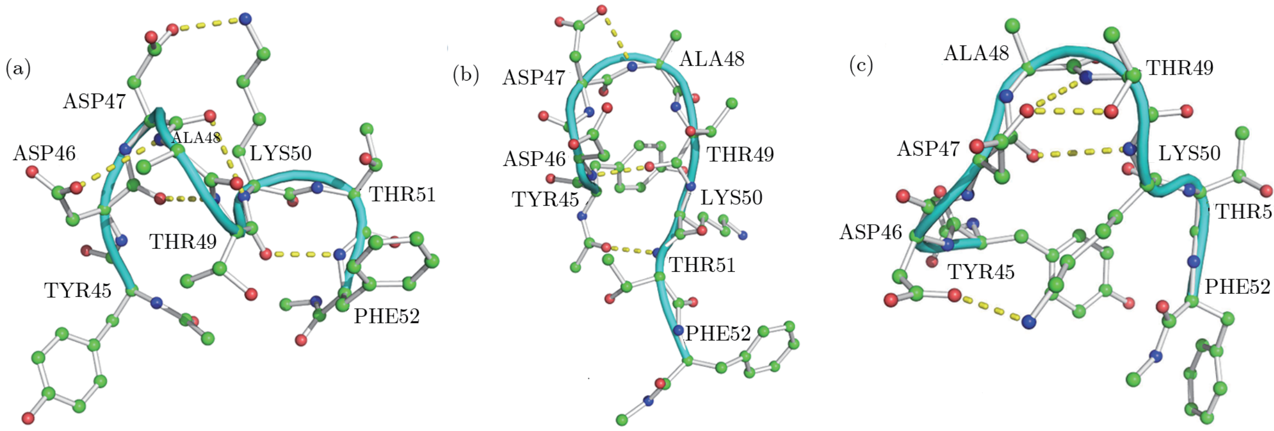

Fig. 17 (Color online) Representative structures of the three major substates starting from the extended structure of seqC. (a) CF3 is mainly stabilized by three 3_{10}-type H-bonds of Asp46:O-Thr49:NH, Asp47:O-Lys50:NH and Thr49:O-Phe52:NH, and another two H-bonds of Asp46:COO$^-$-Ala48:NH and Asp47:COO$^-$-Lys50:NH$_3$. (b) CF2 is mainly stabilized by three H-bonds of Tyr49:O-Asp46:NH, Ace44:O-Thr51:NH, and Asp47:COO$^-$-Ala48:NH. (c) CF1 is stabilized mainly by four H-bonds: Asp47:COO$^-$-Thr49:NH, Asp47:COO$^-$-Thr49:OH,Asp47:O-Lys50:NH, and Asp46:COO$^-$-Lys50:NH$_3$.

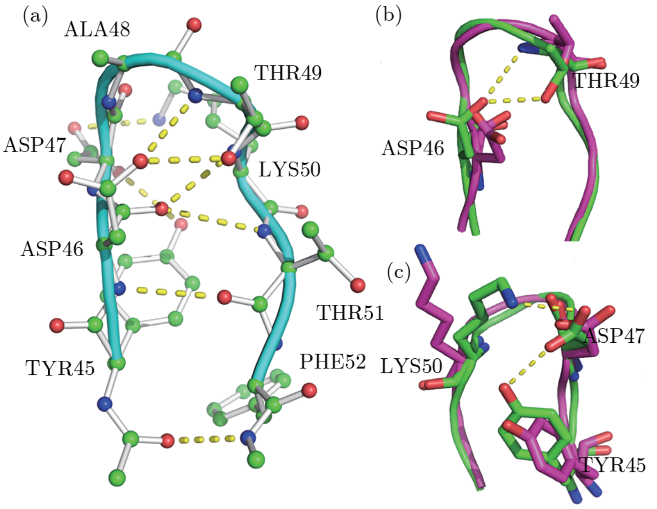

Fig. 18 (Color online) (a) H-bonds formed during the simulation of seqC starting from its native structure. The newly formed H-bonds compared with the native structure are shown in (b) and (c) from two perspectives.

The metastable substate CF3 is a helix-coil structure. Some previous studies [19,48–49] have also found this substate as a misfolded one during the folding process of the C-terminal β-hairpin of the protein G. Since Asp46:COO$^-$ does not form any H-bonds, CF2 is not as stable as the other two substates.

Both seqA and seqC are β-turns in the native structure of the protein G, but they show different stabilities when being solvated in water as individual short peptides. The native structure of seqC can survive more than 500 ns, implying a funnel-shape free energy landscape[50–51] of seqC in aqueous solution; whereas the native structure of seqA transforms into other structures.

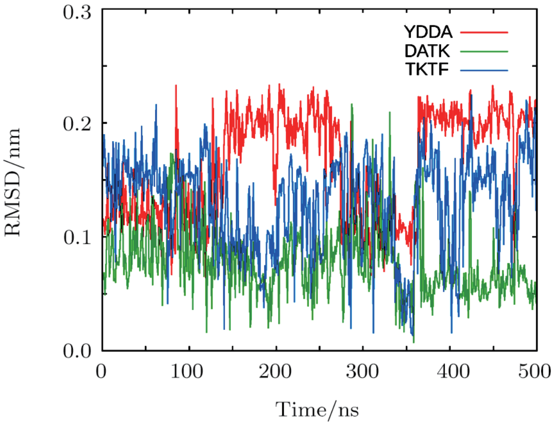

Fig. 19 (Color online) RMSDs with respect to the native structure for different segments of seqC. The center segment DATK has the least RMSD comparing to the other two segments.

(iv) ***β-strand Fragment seqD***

Since seqD is an unpaired β-strand, it is expected not to form the native structure when cutting out from the protein and standing alone in aqueous solution. Two structured substates besides the unstructured extended substate are identified by the combination of the probability map and the MSM method, whose representative structures are shown in Fig. 20. DF1 is stabilized mainly by two 3_{10} type H-bonds (Gly41:O-Thr44:NH and Glu42:O-Tyr45:NH). DF2 is stabilized by three H-bonds formed between Glu42:COO$^-$ and Trp43:NH, Thr44:NH, Thr44:OH. The appearance probabilities of DF1 and DF2 are also shown in Fig. 20. All other configurations belong to the extended substate, which has the largest of probability as 45.8% and is much larger than the previous three fragments. This result shows that seqD is very flexible and does not has a very stable structure.

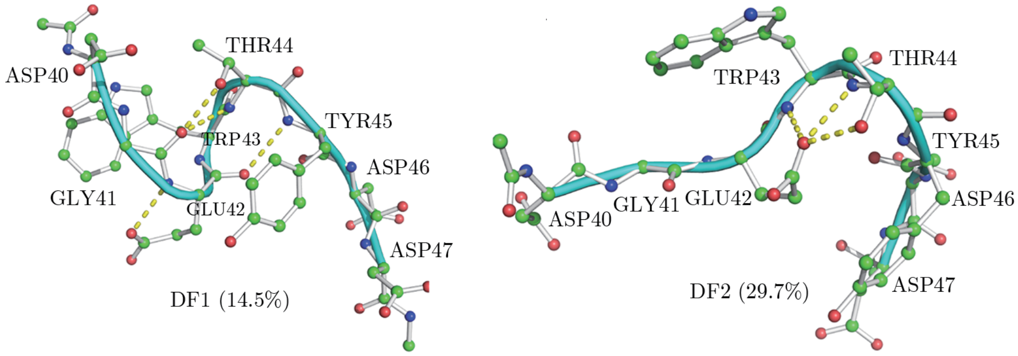

Fig. 20 (Color online) Representative structures and appearance probabilities of DF1 and DF2 of seqD.

To compare seqD with the other three short peptides, we plot the probability maps in the R_g–H-bonds space for all the four peptides in Fig. 21. It can be seen that seqA and seqC are more structured, whereas seqB and seqD are difficult to form compact structures.

(v) ***Relation between 3_{10}-helix and α-helix***

There are mainly two kinds of helical secondary structures: α-helix (with $i \rightarrow i+4$ hydrogen bonding) and 3_{10}-helix (with $i \rightarrow i+3$ hydrogen bonding). α-helix is the most common one constituting about 80% of all helical structures, and 3_{10}-helix is the second most populated one and constitutes about 20% of all helical structures,[52–53] whose portion is larger in short peptides and termini of longer α-helix proteins.[54] The general explanation is that

α-helix has more interactions than 3_{10}-helix with surrounding atomic groups and solvent molecules, so the entropy cost associated with folding is more readily compensated in a long helix or full protein. Experiments and MD simulations on 3K and MW peptides (both are 16-mers) also showed that 3_{10}-helices are present throughout the peptides, with a great contribution at the termini.[54−55] Moreover, Millhauser proposed that 3_{10}-helix is an important intermediate state along the folding/unfolding pathway of α-helix.[56]

Our results for seqB, seqC, and seqD also show that 3_{10}-helix is a very common structure that short peptides tend to form. The transition of seqB from 3_{10}-helix substate BF1 to the native structure supports the statement that 3_{10}-helix can act as an important intermediate state to form α-helix. Three out of five H-bonds in seqC that stabilize the structure of helical-coil substate CF3 are 3_{10}-type H-bonds. Two H-bonds in seqD that stabilize the structure of substate DF1 are also 3_{10}-type H-bonds. All the above results manifest the preference of 3_{10}-helix in short peptides, consistent with previous simulation works and proposed theories.[54−55]

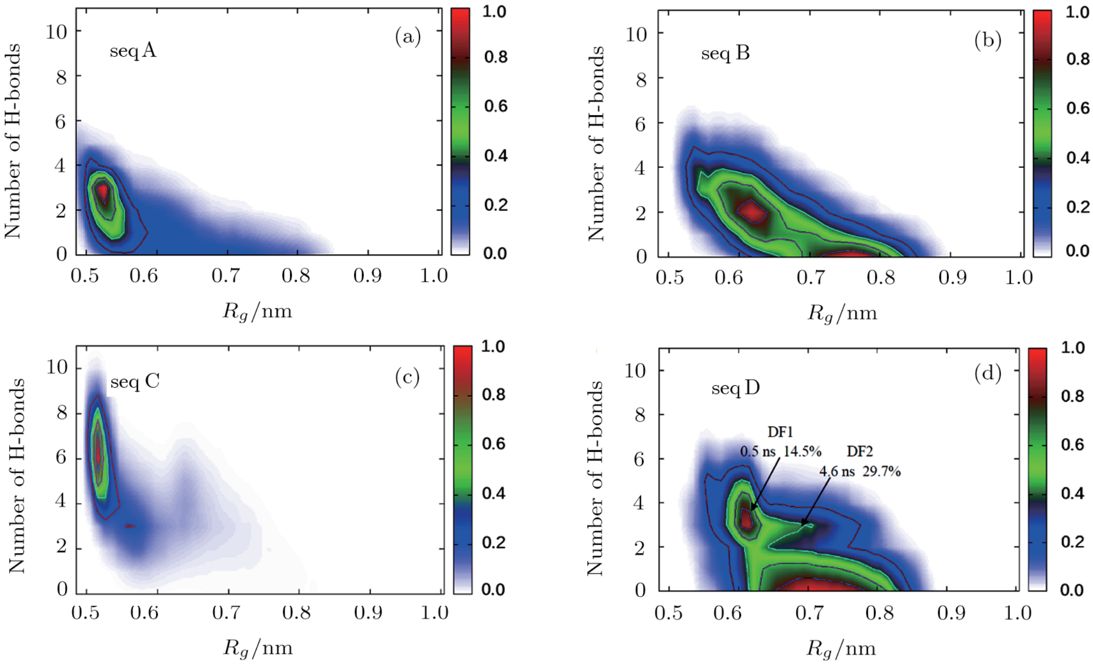

Fig. 21 (Color online) The R_g–H-bonds probability maps for all the four short peptides. (a) seqA; (b) seqB; (c) seqC; (d) seqD, marked with the names, lifetimes, and appearance probabilities of substates DF1 and DF2.

4 Conclusions

In summary, the stand-alone stabilities in aqueous solution of four key fragments, one α-helix, two β-turns, and one unpaired β-strand, essential for the folding dynamics of a protein G are simulated and analyzed in detail. The most stable substates of β-turn fragments are more compact than the α-helix and β-strand. The native substate of the α-helix fragment seqB is less stable than a 3_{10}-helix substate, in good agreement with many previous simulation results.[54−55] Since α-helix structure also appears in our simulations, there is still a great possibility that seqB acts as an early nucleus during the folding of the protein G. The β-turn fragments seqA and seqC show different folding mechanisms despite the fact that they have the same type of secondary structure. The native structure of seqA is not stable when isolated in water, and an alternative substate AF2 becomes more stable during the simulations. Whereas, the native structure of seqC is very stable, but the simulations starting from an extended initial structure can hardly find the native structure, implying a funnel-shape free energy landscape of seqC in aqueous solution. As expected, the unpaired seqD alone is not stable in aqueous solution.

These results indicate that most of the key small fragments have several possible metastable structures when standing alone and many times the native structure is not the most stable one. However, they will be stable in their native structures when they are part of the whole protein under appropriate solvation condition. Thus, we speculate that their interactions with the whole protein and solvation condition are important for them to stable at the final structures. The several possible metastable structures when standing alone show the softness of those peptide fragments, which may be important for the adjustment of the protein during the folding process. On the other hand, the conservativeness of the key peptide fragments showed in our simulation may also be important for guiding the protein folding process towards the correct folded structure. This inference needs further evidence from the simulations of whole protein folding. Overall, our simulation results demonstrate the heterogeneity of key small fragments and provide some information to the complex hierarchial folding dynamics of protein.

References

[1] C. Levinthal, *Mossbauer Spectroscopy in Biological Systems*, University of Illinois Press, Urbana (1969) p. 22.

[2] K. A. Dill and J. L. MacCallum, Science **338** (2012) 1042.

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Nucleic Acids Res. **25** (1997) 3389.

[4] J. Moult, Curr. Opin. Struct. Biol. **15** (2005) 285.

[5] Y. Zhang, Curr. Opin. Struct. Biol. **18** (2008) 342.

[6] C. Venclovas, A. Zemla, K. Fidelis, and J. Moult, Proteins **53** (2003) 585.

[7] A. Kryshtafovych, K. Fidelis, and J. Moult, Proteins **79** (2011) 196.

[8] D. T. Jones, Proteins **45** (2001) 127.

[9] D. T. Jones and L. J. McGuffin, Proteins **53** (2003) 480.

[10] K. T. Simons, R. Bonneau, I. Ruczinski, and D. Baker, Proteins **suppl 3** (1999) 171.

[11] R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. Strauss, and D. Baker, Proteins **suppl 5** (2001) 119.

[12] S. B. Ozkan, G. A. Wu, J. D. Chodera, and K. A. Dill, Proc. Natl. Acad. Sci. U. S. A. **104** (2007) 11987.

[13] C. Bystroff and D. Baker, J. Mol. Biol. **281** (1998) 565.

[14] W. M. Zheng and X. Liu, *Transactions on Computational Systems Biology II*, Springer-Verlag, Berlin (2005) p. 59.

[15] W. M. Zheng, Chin. Phys. B **23** (2014) 078705.

[16] V. Munoz and L. Serrano, Nat. Struct. Mol. Biol. **1** (1994) 399.

[17] F. J. Blanco, G. Rivas, and L. Serrano, Nat. Struct. Mol. Biol. **1** (1994) 584.

[18] K. S. Rotondi and L.M. Gierasch, Biopolymers **716** (2003) 638.

[19] B. K. Ho and K. A. Dill, PLoS Comput. Biol. **2** (2006) 228.

[20] T. Urbič, T. Urbič, F. Avbelj, and K. A. Dill, Acta Chim. Slov. **2008** (2008) 385.

[21] A. M. Gronenborn, D. R. Filpula, N. Z. Essig, A. Achari, M. Whitlow, P. T. Wingfield, and G. M. Clore, Science **253** (1991) 657.

[22] B. Hess, C. Kutzner, D. Spoel, and E. Lindahl, J. Chem. Theory Comput. **4** (2008) 435.

[23] G. A. Kaminski, R. A. Friesner, J. Tiradorives, and W. L. Jorgensen, J. Phys. Chem. B **105** (2001) 6474.

[24] C. Y. Liao and J. Zhou, Acta Chim. Sin. **71** (2013) 593.

[25] S. Enemark and R. Rajagopalan, Phys. Chem. Chem. Phys. **14** (2012) 12442.

[26] W. L. Jorgensen and J. D. Madura, Mol. Phys. **56** (1985) 1381.

[27] T. A. Darden, D. M. York, and L. G. Pedersen, J. Chem. Phys. **98** (1993) 10089.

[28] S. Nosé, J. Chem. Phys. **81** (1984) 511.

[29] M. Parrinello and A. Rahman, Phys. Rev. Lett. **45** (1980) 1196.

[30] D. Wang, B. Jaun, and W. F. Gunsteren, Chem. Bio. Chem. **10** (2009) 2032.

[31] X. Daura, W. F. Gunsteren, and A. E. Mark, Proteins **34** (1999) 269.

[32] L. Deng, P. Zhou, Y. R. Zhao, Y. T. Wang, and X. Hai, J. Phys. Chem. B **118** (2014) 12501.

[33] S. T. Xu, S. X. Zou, and L. C. Wang, *Lecture Notes in Computer Science*, Springer, Cham (2014) p. 356.

[34] F. Noé, I. Horenko, C. Schüette, and J. C. Smith, J. Chem. Phys. **126** (2007) 155102.

[35] F. Noé and S. Fischer, Curr. Opin. Struct. Biol. **18** (2008) 154.

[36] M. Senne, B. Trendelkamp-Schroer, A. Mey, C. Schutte, and F. Noé, J. Chem. Theory Comput. **8** (2012) 2223.

[37] U. Hobohm and C. Sander, Protein Sci. **3** (1994) 522.

[38] A. Luzar and D. Chandler, Phys. Rev. Lett. **76** (1996) 928.

[39] Y. Wang, N. O. Hodas, Y. Jung, and R. A. Marcus, Phys. Chem. Chem. Phys. **13** (2011) 5388.

[40] E. G. Hutchinson and J. M. Thornton, Protein Sci. **3** (1994) 2207.

[41] A. Lewandowska, S. Oldziej, A. Liwo, and H. A. Scheraga, Biophys. Chem. **151** (2010) 1.

[42] R. B. Best and J. Mittal, Proc. Natl. Acad. Sci. U. S. A. **108** (2011) 11087.

[43] L. Thukral, J. C. Smith, and I. Daidone, J. Am. Chem. Soc. **131** (2009) 18147.

[44] V. Muñoz, P. A. Thompson, J. Hofrichter, and W. A. Eaton, Nature (London) **390** (1997) 196.

[45] B. Y. Ma and R. Nussinov, J. Mol. Biol. **296** (2000) 1091.

[46] D. G. Du, Y. J. Zhu, C. Y. Huang, and F. Gai, Proc. Natl. Acad. Sci. U. S. A. **101** (2004) 15915.

[47] A. R. Dinner, T. Lazaridis, and M. Karplus, Proc. Natl. Acad. Sci. U. S. A. **96** (1999) 9068.

[48] D. L. Minor and P. S. Kim, Nature (London) **380** (1996) 730.

[49] B. Zagrovic, E. J. Sorin, and V. S. Pande, J. Mol. Biol. **313** (2001) 151.

[50] J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14** (2004) 70.

[51] P. G. Wolynes, Biochimie **119** (2015) 218.

[52] A. M. Lesk and C. Chothia, J. Mol. Biol. **136** (1980) 225.

[53] D. J. Barlow and J. M. Thornton, J. Mol. Biol. **201** (1988) 601.

[54] R. Armen, D. O. V. Alonso, and V. Daggett, Protein Sci. **12** (2003) 1145.

[55] G. L. Millhauser, C. J. Stenland, P. Hanson, K. A. Bolin, and F. J. M. Ven, J. Mol. Biol. **267** (1997) 963.

[56] G. L. Millhauser, Biochemistry **34** (1995) 3873.