THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

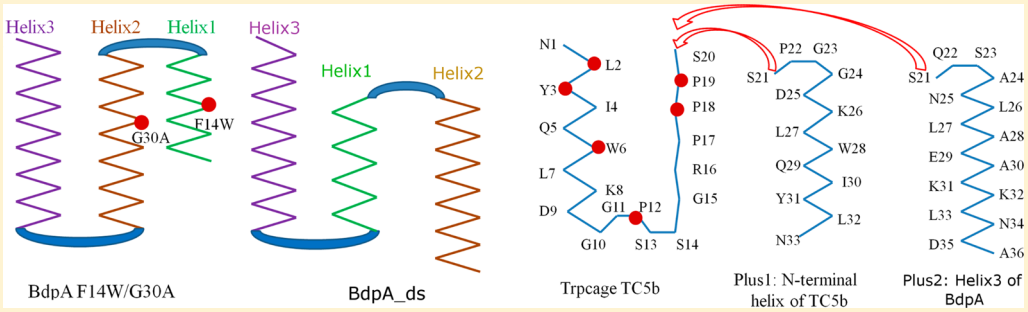
pubs.acs.org/JPCB

Robustness in Protein Folding Revealed by Thermodynamics Calculations

Qiang Shao,[†,‡] Weiliang Zhu,[‡] and Yi Qin Gao*[,†]

[†]Institute of Theoretical and Computational Chemistry, College of Chemistry and Molecular Engineering, Beijing National Laboratory of Molecular Sciences, Peking University, Beijing 100871, China

[‡]Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai, 201203, China



ABSTRACT: Long-range intraprotein interactions play important roles in protein folding. In the present study, we use two variants of the B domain of protein A (BdpA F14W/G30A and BdpA_ds) and two variants of the Trp cage (TC5b_P1 and TC5b_P2) as models to investigate how long-range hydrophobic interactions affect protein tertiary and secondary structures. The mutation of the selected residues (BdpA F14W/G30A) or the change in the sequence order of Helix1 and Helix2 (BdpA_ds) changes detailed hydrophobic interactions. However, this change does not alter the global three-helix-bundle structure of BdpA and the overall shape of the folding free-energy landscape. It does affect the formation and stability of individual secondary structures. Similarly, the addition of an extra segment to the C-terminus of Trp-cage increases the number of long-range hydrophobic interactions without making any significant change of the native structure of Trp-cage. These results show the robustness of the overall protein folding, where rather large sequence changes exert significant influences on secondary but not tertiary structures.

INTRODUCTION

The folding of a protein/polypeptide, including its formation of secondary and tertiary structures, is essentially characterized by inter-residue interactions. For instance, the local electrostatic interactions including hydrogen bonding and ion-pairing and van der Waals interactions largely dictate the secondary structures, whereas the long-range side chain−side chain interactions (mainly hydrophobic interactions) drive the polypeptide chain to folded tertiary structure.[1−3] The amino acid residues interact with each other cooperatively to construct the stable native structure. The knowledge of inter-residue interactions is thus indispensable to the understanding of folding mechanism. Nevertheless, considering the varied physicochemical properties of amino acids, the inter-residue interactions and their effects on protein structure are complex and inevitably lead to the difficulty in protein structure prediction based on only the knowledge of sequence.[4,5] For instance, the accuracy of current methods for secondary structure prediction can reach 80% but seem to be close to the optimal limit.[4,6] One of the main reasons behind this difficulty could be the lack of the consideration of long-range interactions.[7]

The B domain of protein A from *Staphylococcus aureus* (BdpA) is a good model to study long-range intraprotein interactions affecting protein structure and fold free-energy landscape.[8−12] As determined by NMR spectroscopy,[13] the native structure of BdpA in aqueous solution consists of three helices: Helix1 (Gln10-His19), Helix2 (Glu25-Asp37), and Helix3 (Ser42-Ala55, as shown in Figure 1a). Of these three helices, Helix2 and Helix3 contact intimately with each other, whereas Helix1 is tilted ∼30° with respect to the other two helices. A stopped-flow circular dichroism experiment by Bai et al.[14] showed that the isolated Helix3 fragment folds into helical structure in aqueous solution with the helix population of ∼30%, whereas the other two fragments of Helix1 and Helix2 have no detectable helical content. Therefore, one might speculate that the long-range intraprotein interactions within BdpA help form and stabilize the three helices in the native structure. Our recent molecular dynamics (MD) simulation[15] also indicated the important role of long-range intraprotein

Received: August 2, 2012
Revised: October 5, 2012
Published: November 6, 2012

dx.doi.org/10.1021/jp307684h | J. Phys. Chem. B 2012, 116, 13848−13856

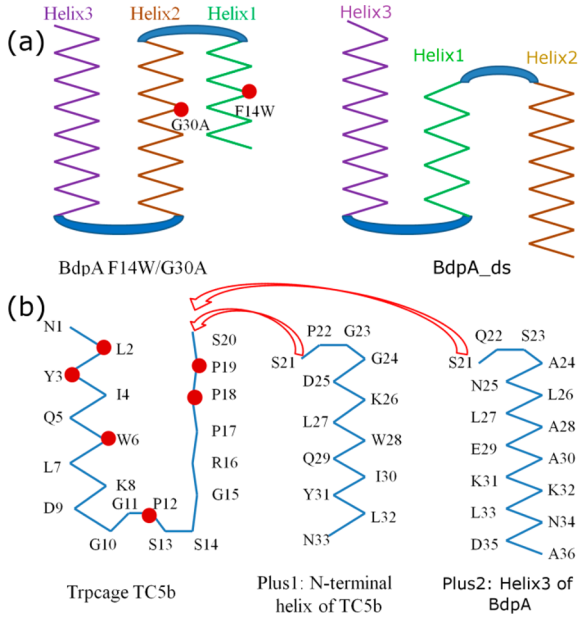


Figure 1. Schematic representations of (a) the structures of BdpA F14W/G30A and BdpA_ds polypeptides and (b) the structure of Trp cage (TC5b) and the generation of elongated polypeptides of TC5b_P1 and TC5b_P2 by connecting the sequences of N-terminal helix of TC5b and Helix3 of BdpA, respectively.

interactions in the folding of BdpA: along the lowest free-energy folding pathway, the most hydrophobic domain Helix3 forms easily; the packing of Helix3 with the other two domains with lower hydrophobicity (Helix1 and Helix2) introduces long-range hydrophobic interactions and facilitates the structuring of Helix1 and Helix2. The latter two are too hydrophilic to form stable secondary structures by themselves.

In the present study, we selected two variants of BdpA as models and ran integrated-tempering-sampling (ITS) molecular dynamics simulations[16,17] to study their folding mechanisms. The first one is the F14W/G30A double mutant of BdpA (BdpA F14W/G30A), which has an experimentally determined structure similar to wild-type BdpA. The other is a domain swapped variant (BdpA_ds) of which the sequence order of Helix1 and Helix2 is exchanged, as shown in Figure 1a. The former variant, as indicated in previous nanosecond laser-induced T-jump experiment by Dimitriadis et al.,[18] has an increased folding rate compared to wild-type BdpA. The increase of the folding rate was attributed by the authors to the increased helical propensity of Helix2 and decreased entropy of the denatured state. But how the stability of Helix2 is enhanced in the double mutant is not yet clear. On the other hand, the other variant (BdpA_ds) is a new sequence and has no known structure. Compared to the point mutations on selected residues in BdpA F14W/G30A, the designed BdpA_ds polypeptide has a sequence change in a larger scale, which makes the prediction of its structure more challenging.

In addition to changing the sequence within BdpA by either mutating residues (BdpA F14W/G30A) or changing the sequence order (BdpA_ds), we also designed two other protein models by adding an extra amino acid sequence to the C-terminus of Trp-cage (TC5b). The Trp cage is a 20-residue polypeptide that folds autonomously and cooperatively into a stable tertiary structure in aqueous solution.[19,20] In its native structure, the Trp cage contains a short α-helix in the sequence range of Asn1-Lys8, a 3_{10}-helix (residues of Gly11-Ser14), and a C-terminal polyproline II region packing against Leu2, Tyr3, and Trp6[19,21] (Figure 1b). As shown in Figure 1b, the amino acid sequence of the N-terminal helix of Trp-cage was added to the C-terminus of the Trp cage to generate the elongated polypeptide TC5b_P1, and the sequence of Helix3 of BdpA was added to generate TC5b_P2 polypeptide. These two sequence fragments are both α-helix structured but have different hydrophobicities. With the extra two sequence fragments added to the C-terminus of Trp-cage, we introduce extra (and different) long-range hydrophobic interactions within the protein, respectively, which are designed to exert an influence on the protein structuring and/or structure stability. Through the enhanced sampling simulation study on the folding of the two designed model proteins (TC5b_P1 and TC5b_P2), we investigate how the addition of supplementary intraprotein interactions to a protein affects its structure.

THEORETICAL METHODS

All MD simulations were performed using the AMBER 9.0 package.[22] In the folding simulations of all polypeptides under study, the GB^{OBC} implicit solvent model[23,24] was used. Polypeptides were modeled with the AMBER FF03 all-atom force field.[25] The salt concentration is set to 0.2 M, and the default surface tension is 0.005 kcal/mol/Å^2. The SHAKE algorithm[26] with a relative geometric tolerance of 10^{-5} is used to constrain all chemical bonds. No nonbonded cutoff was used.

For each polypeptide, we ran 6 independent trajectories, each with a length of several hundred nanoseconds. In each trajectory, the fully extended structure of a polypeptide was first subjected to 2500 steps of minimization. Then the temperature of the system was set by velocity rearrangement from a Maxwell−Boltzmann distribution to be 300 K. The system was then maintained at 300 K using the weak-coupling algorithm with a coupling constant of 0.5 ps^{-1}. Finally the long-time production run of each trajectory was performed with a step size of 0.002 ps and at the constant temperature of 300 K, using the ITS method to encourage thorough sampling over a large energy range. The calculation data were collected every 2.0 ps, and all data from the long-time production run were used for data analysis.

The details of ITS method have been described previously.[16,17] Briefly a modified potential energy is obtained from an integration function over temperature: $V' = -(1/\beta)\ \ln \sum_k f(\beta_k) e^{-\beta_k V}$, where $\beta_k = (1/k_B T_k)$ (k_B is the Boltzmann constant, and T is the temperature). The summation is over a series of discrete temperature values. Compared to the standard MD simulation, the sampled energy range in the ITS simulation is largely expanded. In the present study, 150 temperatures (β_k), evenly distributed in the range of 240−380 K, were used in the ITS method to ensure the efficient sampling of the desired energy range. A preliminary iteration process was employed to obtain converged values of discrete $f(\beta_k)$s, which were then used in the long-time production run to evenly sample the entire energy range. After the data were collected in production run using the biased potential with the set of $f(\beta_k)$s, the thermodynamic properties of the simulated system at a desired temperature β were then calculated by reweighting of each term by timing a factor $e^{\beta(V'(r)-V(r))} = e^{-\beta V(r)}/p(V') = e^{-\beta V(r)}/\sum_{k=1}^{N} f(\beta_k) e^{-\beta_k V}$. For the free-energy profiles calculated at room temperature, the thermodynamics was reweighted using $\beta_0 = 1/k_B T_0$, where $T_0 = 300$ K. Meanwhile, the

dx.doi.org/10.1021/jp307684h | J. Phys. Chem. B 2012, 116, 13848−13856

thermodynamics can also be calculated at other desired temperatures.

RESULTS

Force Field and Solvent Model Selection for Molecular Dynamics Simulation. It is worth noting here that there are no NMR structures available for three of the protein models under study (BdpA_ds, TC5b_P1, and TC5b_P2), and the present study thus makes predictions of polypeptide structures at room temperature and needs experimental tests. In principle, the accuracy of predictions largely depends on the force field and simulation methodology used. In the present study, an AMBER FF03 force field[25] was applied to model the polypeptides, and a GB^{OBC} model[23] was used to model the solvent effects. AMBER FF03 force field has been used in previous MD simulations of the Trp cage,[27−29] and its combination with the GB^{OBC} model generated the folding/unfolding thermodynamics data of the Trp cage in better agreement with experiments when compared to other force fields such as FF96.[30] Moreover, the same force field has also been used successfully in many MD simulations to fold wild-type BdpA.[15,31,32]

On the other hand, in comparison with the explicit solvent model which explicitly and thus more reasonably mimics the physical properties of water, the GB^{OBC} implicit solvent model treats solvent molecules as structureless continuums. Thus the latter model might generate folding thermodynamics parameters deviated from experimental data and/or those from explicit solvent simulation. For instance, our previous simulation on the folding of wild-type BdpA using the FF03 force field and the GB^{OBC} model[15] indicated that the absolute stabilities of individual helices within BdpA are inconsistent with the experimental ones. Nevertheless, the calculated helix stability order is consistent with that determined by experiments[14] and explicit solvent simulation.[15]

As mentioned earlier, ITS[16,17] was used in the present folding simulations to enhance the sampling capability on the potential energy surface. The high sampling efficiency of ITS has been shown in several previous studies.[15,33−36] Figure 2 shows a typical trajectory of the folding simulation using ITS for each polypeptide at room temperature, represented by the time series of potential energy sampled. It is clearly seen that both high and low potential energies are sampled frequently in each trajectory. As a result, the folding simulations allow thorough sampling of configurations including unfolded, folded, and transition states for each polypeptide and thus provide quantitative description of the protein folding free-energy landscape.

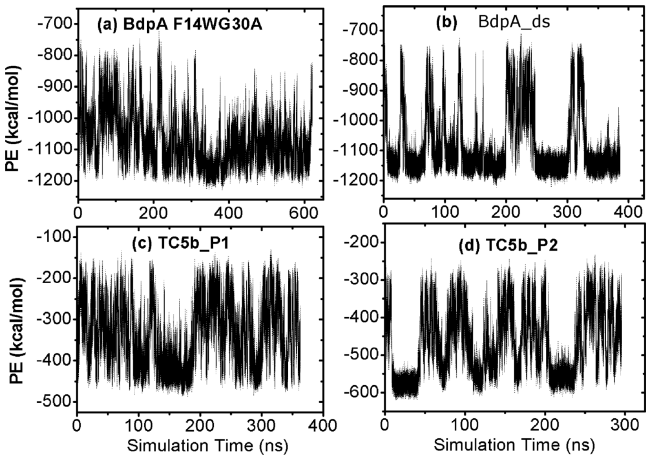


Figure 2. Time series of potential energy sampled in the typical trajectories from ITS-MD simulations on (a) BdpA F14W/G30A, (b) BdpA_ds, (c) TC5b_P1, and (d) TC5b_P2. All simulation trajectories were performed with a step size of 0.002 ps and at 300 K, using the ITS method to encourage thorough sampling over a large energy range.

Sequence Dependence of Folded Structure and Free-Energy Landscape for BdpA Variants. For each polypeptide under study, the hierarchiral clustering analysis on ∼40000 snapshots evenly selected from the trajectories was performed to indicate the most populated structure. The clustering analysis was run using the MMTSB Toolset.[37] Clustering was based on the Cα_RMSD between structures, and the cluster radius was set as 8 Å. A total of 29 clusters for BdpA F14W/G30A and 51 clusters for BdpA_ds were identified. Not surprisingly, most of these clusters were too poorly populated to be of interest and were excluded from further analysis. As shown in Figure 3, the most populated structure of BdpA F14W/G30A (population 38.97%) is the nativelike folded structure which has the similar configuration as the native structure of wild-type BdpA, with the backbone RMSD of 2.8 Å (top panel in Figure 4). Interestingly, the most populated structure of BdpA_ds (population 32.10%), which has the exchanged order of sequences between Helix1 and Helix2 segments, also closely resembles that of the wild-type native structure, although the positions of three-dimensional helical structures of Helix1 and Helix2 are switched.

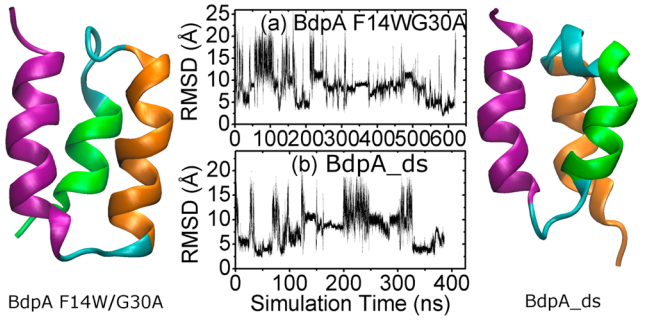


Figure 3. Time series of backbone heavy-atom RMSD (root-mean-square deviation) corresponding to the most populated structures of (a) BdpA F14W/G30A and (b) BdpA_ds polypeptides. The most populated structures of the two polypeptides are present at the left and right edges of which Helix1, Helix2, and Helix3 are green, orange, and purple colored.

The one-dimensional free-energy profile as a function of backbone root-mean-square deviation (RMSD) was then calculated at room temperature. As shown in parts b and c of Figure 4, the nativelike structure with small RMSD value has the lowest free energy for either BdpA F14W/G30A or BdpA_ds, indicating that the native structure is the most populated for each polypeptide. This result is consistent with the observation in the clustering analysis. By use of the most populated structure as the reference state, the RMSD of each polypeptide (BdpA F14W/G30A or BdpA_ds) was calculated as a function of simulation time. One can see from Figure 3 that multiple folding (RMSD <3.5 Å) and unfolding events can be observed in every trajectory for each polypeptide. For a protein such as BdpA and its variants which have complex tertiary

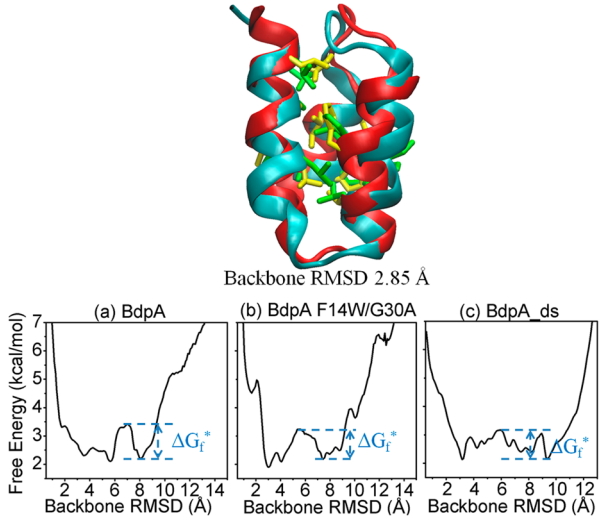


Figure 4. Top panel: the most populated structure of BdpA F14W/G30A (blue) compared to NMR structure (red). All hydrophobic side chains involved in native hydrophobic core clusters are shown with the licorice representation (the hydrophobic side chains in the most populated structure are yellow colored, and those in NMR structure are green colored). Bottom panel: the one-dimensional free-energy profile as a function of backbone RMSD calculated for (a) BdpA, (b) BdpA F14W/G30A, and (c) BdpA_ds, respectively (all free-energy profiles in the present article are calculated at 300 K).

interactions, the structures very close to NMR structure are actually not easy to be captured using the current simulation methodology and force field. For instance, a set of high-temperature (400 K) implicit solvent MD simulations by Jang et al. could transiently obtain the nativelike structure of BdpA with the lowest Cα_RMSD value of 2.9 Å.[38] The electrostatically driven Monte Carlo (EDMC) simulation by Vila et al. sampled the structure of BdpA with the lowest Cα_RMSD of 2.85 Å.[39] Our previous MD simulation sampled the structure of BdpA to the backbone heavy-atom RMSD of 1.70 Å and the overall heavy-atom (including side-chains) RMSD of 2.6 Å.[15] Development of a new and robust force field and implicit solvent modes is still pretty much in need for accurate prediction of protein structures.

Several reaction coordinates, which were used in the two-dimensional free-energy landscape analysis of wild-type BdpA in our previous study,[15] were utilized to present the data: (1) The total number of native contacts (side chain−side chain packing) among the three helices (C_{Nat}, only the native contacts between residues from different helices are included) used to estimate the interhelical side chain interactions. (2) The number of interhelical hydrophobic contacts (side chain−side chain packing of hydrophobic residues from different helices, N_{HC}) used to estimate the interhelical hydrophobic interactions. (3) The total number of backbone hydrogen bonds formed, N_{HB}, used to measure the formation of helices.

Figure 5 shows the contact maps of the two variants of BdpA, which were calculated using the contact map analysis (CMA) server in SPACE tools.[40] For all native contacts in parts a and b of Figure 5, only those between residues from different helices (C_{Nat}), which demonstrates the long-range side chain contacts from different subdomains, are evaluated in the present study. Similarly, N_{HC} is evaluated by the hydrophobic contacts only from different helices, and those within individual helices are excluded. The native interhelical hydrophobic contacts for both BdpA F14W/G30A and BdpA_ds are organized in Table 1.

Table 1. Native Interhelical Hydrophobic Contacts for BdpA F14W/G30A and BdpA_ds

BdpA F14W/G30A		BdpA_ds	
A13-L46	L20- F31	I17-Y33	L36-L45
F14-I32	L20-A49	L20-Y33	L36-L46
F14-L35	L20-L52	L20-L26	L36-A49
I17-F31	F31-L45	L23-L46	
I17-I32	F31-A49	L26-A49	
I17-L35	F31-L52	F32-A49	
I17-L46	I32-L45	F32-L45	
I17-A49	L35-L45	F32−52	
L18- I32		I35-L45	

While C_{Nat} or N_{HC} is accounted, the side chains of two residues are considered as in contact only if the minimal distance between heavy atoms from two side chains is within 5 Å. [It is worth noting that the free-energy landscapes of wild-type BdpA were also drawn in parts a and b of Figure 6 to make a direct comparison to its variants, of which the C_{Nat} and N_{HC} were recalculated using the cutoff distance of 5 Å instead of 6 Å in ref 15. The free-energy profiles of wild-type BdpA in Figure 6 have the same shapes as those in Figure 3 in ref 15 except that the detailed values of C_{Nat} and N_{HC} are different because of the

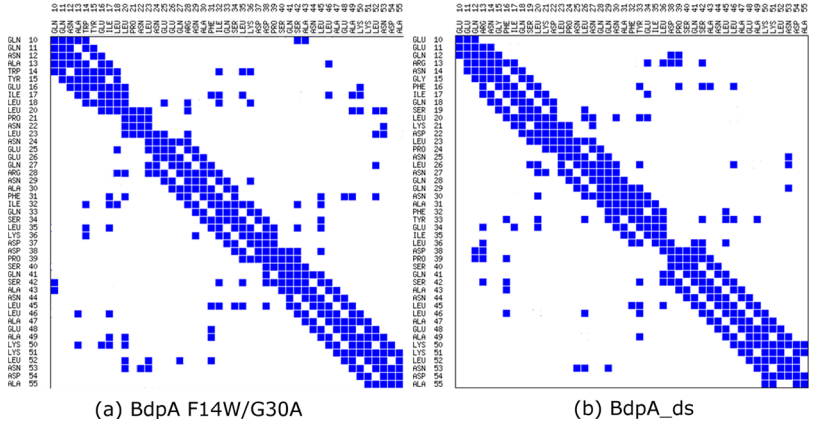


Figure 5. The contact maps of (a) BdpA F14W/G30A and (b) BdpA_ds. The contact map is calculated by the contact map analysis (CMA) server in SPACE tools.[40].

dx.doi.org/10.1021/jp307684h | J. Phys. Chem. B 2012, 116, 13848−13856

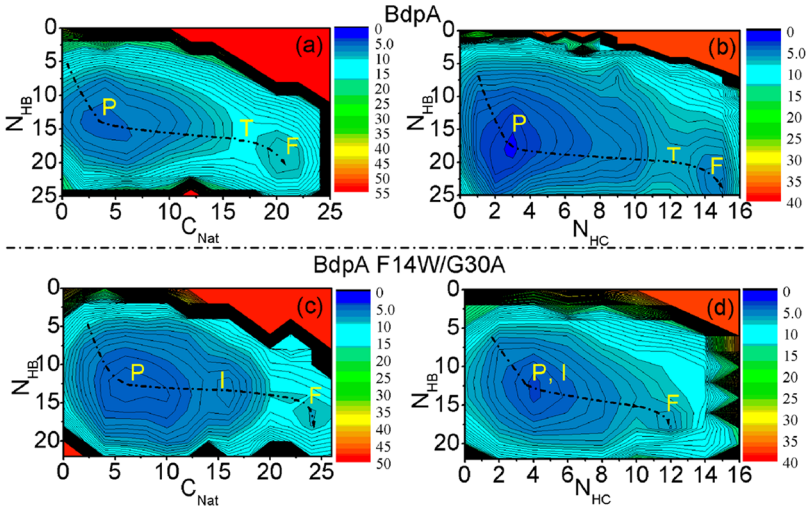


Figure 6. The free-energy landscapes as the function of C_{Nat} and N_{HB} (a) and N_{HC} and N_{HB} (b) for BdpA and C_{Nat} and N_{HB} (c) and N_{HC} and N_{HB} (d) for BdpA F14W/G30A. The contours are spaced at intervals of 0.5 k_BT.

usage of different cutoff distances.] Backbone hydrogen bonds are defined as HBs 1−25, numbered from N- to C- terminus direction (HBs 1−5 in Helix1 segment, HBs 6−15 in Helix2, and HBs 16−25 in Helix3). A backbone hydrogen bond is considered as formed only if the distance between two heavy atoms of the hydrogen bond donor and acceptor is less than 3.2 Å and the N−H−O angle is greater than 135°.

The two-dimensional free-energy landscape is calculated with the normalized probability, $P(x) = Z^{-1}\, e^{-\beta W(x)}$ from a histogram analysis, where x is any set of reaction coordinates. $W(x_2) - W(x_1) = -(1/\beta)\ln(P(x_2)/P(x_1))$ is the relative free energy or potential of mean force.[41] We first show the free-energy landscape as a function of C_{Nat} and N_{HB} for BdpA F14W/G30A in Figure 6c. Three states are observed, consisting of the physiologically unfolded (P) state ($10 \le N_{HB} \le 15$ and $5 \le C_{Nat} \le 10$, population 30.84%), the folded (F) state ($16 \le N_{HB} \le 18$ and $C_{Nat} \ge 23$, population 38.97%), and an intermediate (I) state ($12 \le N_{HB} \le 15$ and $15 \le C_{Nat} \le 17$, population 10.25%) between P and F states. The physiologically unfolded state is so-called because it has the similar structural characteristics as the physiologically unfolded state defined for wild-type BdpA:[15] it has a high population compared to the folded state and thus should not be considered as an intermediate state; meanwhile, it contains a large proportion of secondary structures but only a small proportion of native contacts.

The lowest free-energy folding pathway indentified by a connection with the lowest barrier between the initial and final states can be drawn to connect these distinct states (black dash lines in Figure 6c). This free-energy landscape has very similar features to that of wild-type BdpA, which possesses the same three distinct states (Figure 6a). The only difference between the two free-energy profiles of BdpA F14W/G30A and wild-type BdpA is that an intermediate state with relative low free-energy presents in the former profile, which is located at similar position as the transition (T) state in the latter profile. Therefore the I state of BdpA F14W/G30A should correspond to the T state of BdpA, and the double mutation of F14W and G30A lowers the free energy of the transition state and thus stabilizes it.

To gain structural insights into the folding mechanism of BdpA F14W/G30A, we performed clustering analyses on snapshots selected from the simulation trajectories which satisfy the criteria for both P and I states (e.g., $10 \le N_{HB} \le 15$ and $5 \le C_{Nat} \le 10$ for the P state, $12 \le N_{HB} \le 15$ and $15 \le C_{Nat} \le 17$ for the I state), respectively. The most populated conformations of the P state are shown in Figure 7 with their percentages

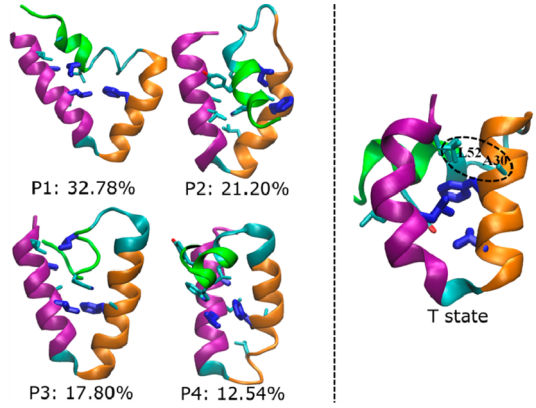


Figure 7. Left: Representative structures of the populated conformational ensembles of the physiologically unfolded (P) state of BdpA F14W/G30A (four populated conformations are obtained from the clustering analysis (P1−P4) and showed with their percentages in the ensembles). Right: Representative structures of the predominant conformation of intermediate (I) state obtained from clustering analysis. Hydrophobic side chains participating in the side chain contacts in these structures are shown with the licorice representation, and those in the native contacts are blue-colored.

in the structure ensembles indicated. Similar to the P state of wild-type BdpA, in all conformations of the P state of BdpA F14W/G30A, Helix3 is the best folded, and Helix2 is partially folded, whereas the helical content of Helix1 segment in all conformations is the least. On the other hand, all conformations of the P state possess compact structures which are stabilized mainly by non-native interhelical hydrophobic contacts (see the larger amount of non-native hydrophobic contacts than that of native ones in Table 2). This structure character of P state is again similar to that of wild-type BdpA.

dx.doi.org/10.1021/jp307684h | J. Phys. Chem. B 2012, 116, 13848−13856

Table 2. Interhelical Hydrophobic Contacts Formed in the Populated Conformations of the P and I States of BdpA F14WG30A obtained from Clustering Analysis[a]

interhelical hydrophobic contacts in BdpA F14W/G30A				
P1	P2	P3	P4	T
I17-A49	**I17-F31**	**I17-A49**	**F31-L45**	**F31-L45**
F31-L45	A13-F31	**F31-L45**	**F31-A49**	**F31-A49**
A13-L52	W14-F31	A13-F31	W14-L45	**L35-L45**
I17-L45	W14-L35	W14-F31	W14-L52	Y15-F31
I17-L52	W14-L46	W14-L45	Y15-L52	Y15-L35
L18-F31	Y15-L45	W14-A49	L18-F31	Y15-L46
L18-L45	Y15-A49	I17-L52	L18-A49	L18-F31
L18-L46	L18-L46	A30-L45	L18-L52	L18-L52
L18-A49	L35-L46	F31-L46	I32-L46	A30-L52
			L35-L46	F31-L52

[a]The native hydrophobic contacts are in bold.

In comparison to the multiple conformations of the P state, the I state (Figure 7) possesses more native hydrophobic contacts. But this difference in native hydrophobic contacts between P and I states is not large enough to separate these two states in the free-energy landscape using the native interhelical hydrophobic contacts as the reaction coordinate. Therefore the P and I states merge in Figure 6d. Among all hydrophobic contacts formed in the I state, the noteworthy one is the contact between Leu52 and the mutated residue G30A. To see how such hydrophobic contact between Leu52 and Ala30 affects the stability of individual backbone hydrogen bonds, we calculated the average formation probability of individual hydrogen bonds of BdpA F14W/G30A at different temperatures. As shown in Figure 8, the backbone hydrogen bonds close to the contact of Leu52-Ala30 (HB8, HB9, HB11) have apparently higher stabilities than the same hydrogen bonds in wild-type BdpA absent of such hydrophobic contact (Figure 8). Therefore the mutation of G30A leads to the formation of an additional hydrophobic contact between Leu52 and Ala30 in the I state, which generally increases the hydrogen bond stability of Helix2 (and thus could accelerate the folding of BdpA). Meanwhile, the mutated residue of F14W in Helix1 has no contacts with other hydrophobic residues in the I state. As a result, the hydrogen bonds in Helix1 and Helix3 are not influenced by the mutation.

In addition, in our previous study on wild-type BdpA, we observed that the stability of individual backbone hydrogen bonds in the three-helix bundle is strongly affected by their surrounding environment.[15] Generally speaking, the hydrogen bonds buried in the hydrophobic core cluster have a higher stability than those remote to the hydrophobic core cluster (HB7, HB10, HB11, HB13, and HB14 in Helix2 and HB19, HB22, and HB25 in Helix3, see Figure 8). As the temperature increases, the hydrophobic core cluster becomes loosened, and the stability difference among all hydrogen bonds becomes smaller. The same tendency can be also seen for BdpA F14W/G30A (Figure 8): the hydrogen bonds remote to the hydrophobic core cluster (HB7, HB10, HB11, HB13, and HB14 in Helix2 and HB19, HB22, and HB25 in Helix3) are apparently less stable than other hydrogen bonds. The difference of hydrogen bond stability becomes more apparent as the temperature decreases.

The free-energy landscapes as the function of C_{Nat}, N_{HB}, and N_{HC} for the BdpA_ds polypeptide were also calculated, and the results are shown in Figure 9. Two local minima corresponding to P ($5 \le N_{HB} \le 11$ and $4 \le C_{Nat} \le 8$ in Figure 9a, $5 \le N_{HB} \le 11$ and $2 \le N_{HC} \le 4$ in Figure 9b) and F states ($10 \le N_{HB} \le 13$ and $22 \le C_{Nat} \le 25$ in Figure 9a, $10 \le N_{HB} \le 13$ and $7 \le N_{HC} \le 10$ in Figure 9b) are present in the two profiles. Compared to wild-type BdpA and BdpA F14W/G30A, the folded state of BdpA_ds possesses less helical content, which correlates with the fewer interhelical hydrophobic contacts formed in the folded state. Although the total number of native contacts in the folded structures of BdpA F14W/G30A and BdpA_ds is similar, the native hydrophobic contacts in the former polypeptide are more than those in the latter. Figure 10 demonstrates the most populated conformations of P state of BdpA_ds obtained from the clustering analysis. Among these

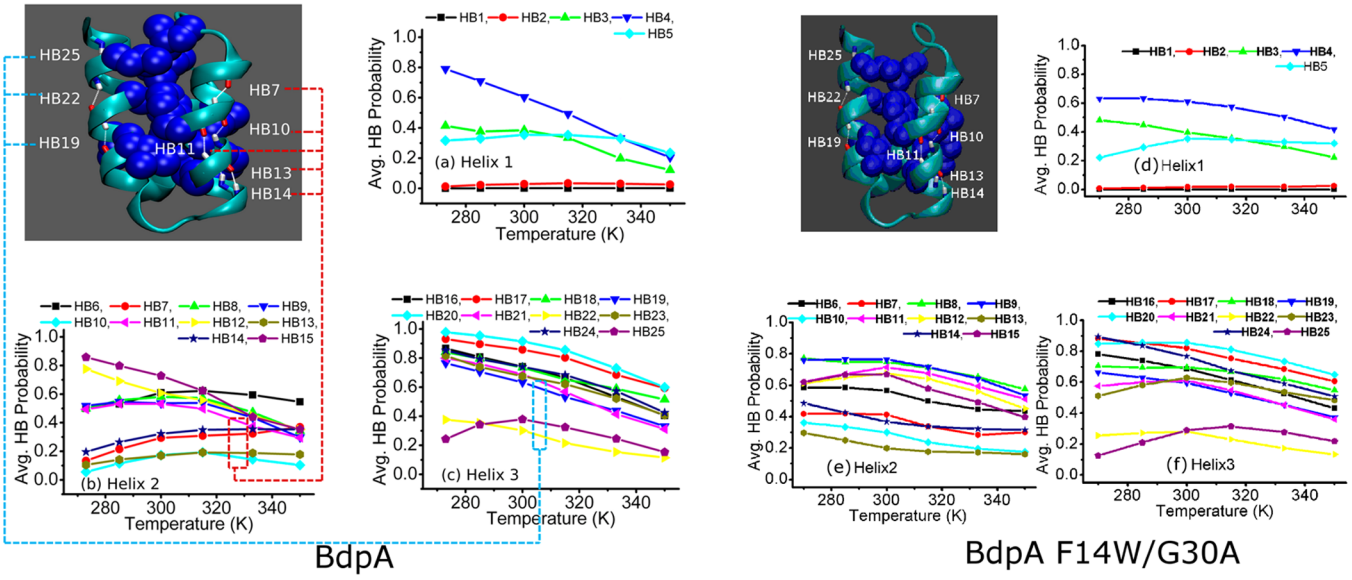


Figure 8. The average formation probability of individual backbone hydrogen bonds belonging to different helices (Helix1, Helix2, and Helix3) in BdpA and BdpA F14W/G30A at different temperatures. (Top panel: the hydrogen bonds remote to the hydrophobic core in the native structure of the polypeptide are represented by dashed lines, and the interhelical hydrophobic core is represented by VDM model.)

dx.doi.org/10.1021/jp307684h | J. Phys. Chem. B 2012, 116, 13848−13856

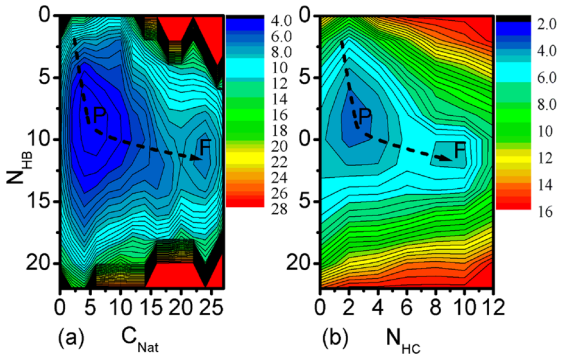


Figure 9. The free-energy landscapes as the function of C_{Nat} and N_{HB} (a) and N_{HC} and NHB (b) for the BdpA_ds polypeptide. The contours are spaced at intervals of 0.5 k_BT.

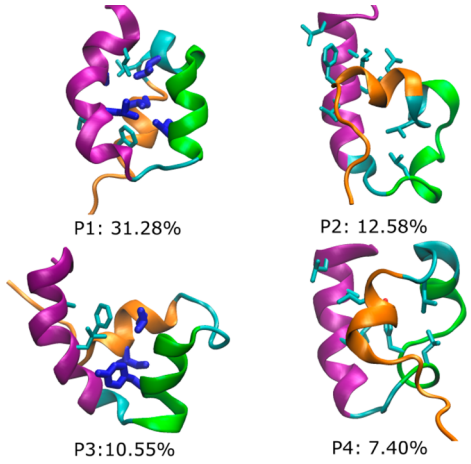


Figure 10. Representative structures of the populated conformational ensembles of the physiologically unfolded (P) state of BdpA_ds. Hydrophobic side chains participating in the side-chain contacts in these structures are shown with the licorice representation, and those in the native contacts are blue-colored.

conformations, Helix3 is again largely folded. Helix1, which is now in the middle, is also largely folded except in P2. In contrast, the helical content of Helix2, which now moves to the protein surface, is apparently less stable than the other two.

The overall shapes of the free-energy landscapes for wild-type BdpA and the variants remain largely unchanged. For instance, the distinct states such as P and F states are present at similar positions in the free-energy landscapes, respectively, and as a result, the lowest free-energy folding pathway connecting these states goes through similar positions and has a similar shape. On the other hand, parts a–d of Figure 6 do show that the mutation of F14W/G30A changes the details of the free-energy landscape and causes the appearance of new intermediate state. In addition, the change of the sequence order of Helix1 and Helix2 in BdpA_ds decreases the total number of interhelical native contacts within the protein. As a result, the distinct states in the free-energy landscape of BdpA_ds (Figure 9) have less numbers of NHC and N_{HB} compared to the analogous states in the free-energy landscapes of wild-type BdpA and BdpA F14W/G30A, respectively.

Considering P and F as the main minima (the T state has an apparently higher free energy and its population is much lower than that of P or F state), the folding of wild-type BdpA and BdpA_ds can be regarded as a two-state process. In contrast, the folding of BdpA F14W/G30A is a three-state transition. The two-state transition of wild-type BdpA folding observed here is consistent with the experimental observation of Bai et al.:[14] no stable intermediate state was observed during the denaturation of the I17W mutant of BdpA in the diluted GdmCl solution, which may suggest that BdpA folds rapidly by a two-state mechanism without the formation of stable partially folded intermediates. The consistency between the experiment and the present study indicated the reliability of the present simulation results. Moreover, the present study suggested that the mutation on Helix2 (G30A) but not Helix1 (F14W) can generate the new intermediate, which is worth testing by experiments.

The results from free-energy landscape analysis and clustering analysis together show that the point mutation on specific residues (BdpA F14W/G30A) or changing the sequence order of secondary structure components (BdpA_ds) does not change the global topology of the protein and has little influence on its minimum free-energy folding pathway. Along the minimum free-energy folding pathways of BdpA and its variants, the extended structure collapses into a compact and stable unfolded state (P state) driven by the barrierless structure collapse. In the P state, the most hydrophobic helix, Helix3, and the helix in the middle position of protein (either Helix2 in wild-type BdpA and BdpA F14W/G30A or Helix1 in BdpA_ds) are largely structured, whereas the less hydrophobic helix at the protein surface is only partially formed. Moreover, the P state is stabilized by not only native but also non-native side-chain interactions (see the interhelical hydrophobic contacts formed in various states in Table 2 and Table 3).

Table 3. Native Interhelical Hydrophobic Contacts Formed in the Populated Conformations of the P State of BdpA_ds Obtained from Clustering Analysis[a]

interhelical hydrophobic contacts in BdpA_ds			
P1	P2	P3	P4
L20-I26	F16-L45	**I17-Y33**	L20-L45
I26-A49	F16-L52	**L20-Y33**	L20-L52
I35-L45	I17-A49	F16-L45	L23-L36
F16-L46	L20-A49	F16-A49	Y33-L45
L23-L46	L20-L46	L20-L45	Y33-L46
L20- L46	L23-L36	Y33-L45	Y33-A49
I26-L46			
I35-L52			

[a]The native hydrophobic contacts are marked in bold.

To further fold into the F state, essential local structural rearrangement is needed (e.g., the breaking of non-native side-chain contacts and the formation of native side chain contacts) to form the correct tertiary interactions.

The effects of these sequence changes seem to be limited to the formation and stability of secondary structures within the protein, and the effects are essentially induced by the change of long-range hydrophobic contacts. The mutation of G30A creates an additional hydrophobic contact between Leu52 and Ala30 in the I state along the folding pathway of BdpA F14W/G30A, which leads to the increase of the stability of Helix2. Moreover, changing the order between Helix1 and Helix2 decreases the total number of hydrophobic contacts in the native structure and therefore reduces the helical content. On the other hand, the one-dimensional free-energy profiles as the function of RMSD (Figure 4) show that the folding free-energy

dx.doi.org/10.1021/jp307684h | J. Phys. Chem. B 2012, 116, 13848–13856

barrier (ΔG_f^*) is 1.23 kcal/mol for wild-type BdpA, 1.05 kcal/mol for BdpA F14W/G30A, and 1.02 kcal/mol for BdpA_ds. Therefore BdpA F14W/G30A has a lower folding free-energy barrier than wild-type BdpA, corresponding to the increased folding rate of the former polypeptide than the latter.

The Effects of Supplementary Intraprotein Interactions on the Native Structure and Folding Mechanism of the Trp Cage. As further examples for protein structure determined by sequence, we studied next the folding of two elongated polypeptides based on the Trp cage (TC5b_P1 and TC5b_P2) to investigate the influence of additionally added intraprotein interactions on the structure of the Trp cage.

Figure 11 demonstrates the most populated structures obtained from the clustering analysis for TC5b, TC5b_P1, and TC5b_P2. The most populated structure of TC5b is the nativelike structure, which has a Cα_RMSD value of 1.41 Å corresponding to the NMR structure. In the most populated structures of TC5b_P1 and TC5b_P2, the segment of TC5b also adopts the nativelike structure with very small Cα_RMSDs (0.99 Å for TC5b_P1 and 1.67 Å for TC5b_P2), indicating that the addition of an extra sequence to the Trp cage does not alter the folded structure of the Trp cage. The additional amino acids (the sequence of N-terminal helix of TC5b added in TC5b_P1 and the sequence of Helix3 of BdpA added in TC5b_P2) fold into α-helix structures and have direct side chain−side chain contacts with the TC5b segment. These extra intraprotein interactions, although do not change the overall structure of the original segment of Trp-cage, decrease the stability of native structure, revealed by the lower populations of the most populated structures of TC5b_P1 and TC5b_P2 than that of TC5b as shown in Figure 11.

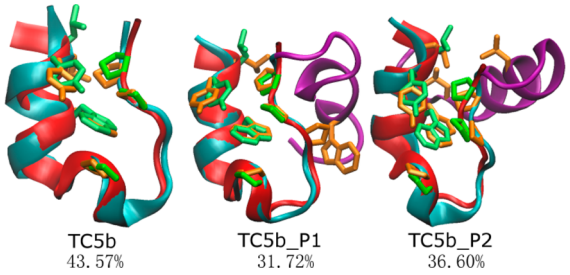


Figure 11. The most populated structures of TC5b, TC5b_P1, and TC5b_P2 obtained from the clustering analyses (blue) and the comparison to the NMR structure of TC5b (red). The artificially added segments in TC5b_P1 and TC5b_P2 are purple. All hydrophobic side chains involved in the native hydrophobic core cluster (L2, Y3, W6, P12, P18, and P19) are shown with the licorice representation (the hydrophobic side chains in the most populated structure are in orange, and those in NMR structure are in green).

To see the influence of the added intraprotein interactions on the folding mechanism of TC5b, we calculated its free-energy landscape as a function of Cα_RMSD and the radius of gyration (R_g) of the TC5b segment for the Trp cage and its variants. As shown in Figure 12, the free-energy landscapes of TC5b and its two variants share a similar shape, of which two distinct states are present, corresponding to the unfolded and folded states, respectively. All the folded structures in the three polypeptides have the similar TC5b segment configuration with small values of Cα_RMSD, whereas the unfolded states have a large variation in Cα_RMSD.

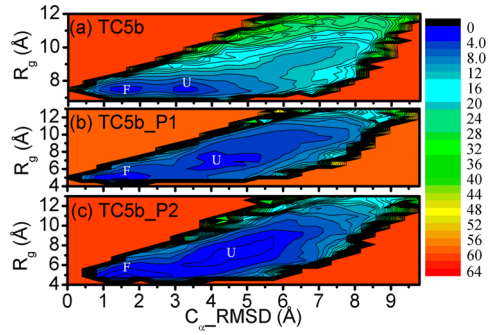


Figure 12. The free-energy landscapes as the function of RMSD and R_g for (a) TC5b, (b) TC5b_P1, and (c) TC5b_P2. The contours are spaced at intervals of 0.5 k_BT.

DISCUSSION AND CONCLUSIONS

Changing and/or optimizing the folding kinetics and thermodynamics properties by making changes in protein sequence is a widely used strategy in experimental studies of protein folding.[18,42,43] For instance, the double mutations of F14W and G30A of the B domain of protein A (BdpA) were found to increase the folding rate from 120 000 s^{-1} to 249 000 s^{-1}.[18] However, the molecular mechanism through which the sequence change affects the protein folding pathway and induces the change in folding kinetics and thermodynamics remains largely unclear.

One possible consequence of a sequence change is the change of long-range side chain−side chain interactions (mainly hydrophobic interactions) within protein, which play an important role in protein structure formation. In the present study, we used two variants of BdpA polypeptide, BdpA F14W/G30A, and BdpA_ds as models to investigate the change of intraprotein hydrophobic interactions induced by sequence mutation and their effects on protein folding.

It was observed that neither the tried mutation on BdpA sequence (BdpA F14W/G30A) nor the change of sequence order between secondary structure components of Helix1 and Helix2 (BdpA_ds) changes the overall topology of the tertiary structure of BdpA. Both BdpA variants fold into stable three-helical bundle structures similar to that of wild-type BdpA, indicating the intrinsic robustness of sequence determining protein structure. On the other hand, the stabilities of individual helices **are** affected by the long-range intraprotein hydrophobic interactions induced by the sequence change. For instance, in BdpA_ds, fewer interhelical hydrophobic contacts can be formed, and as a result a lower portion of helical contents of Helix1 and Helix2 form in its folded structure when compared to wild-type BdpA. Moreover, in BdpA F14W/G30A, the mutation of G30A in BdpA stabilizes its transition state in the folding pathway (Figure 6). In this mutant Ala30 and Leu52 form additional hydrophobic contacts and thus stabilize the backbone hydrogen bonds of Helix2, leading to its faster folding compared to wild-type BdpA. The other residue Trp14 is not involved in any hydrophobic contacts in the transition state and is shown not to contribute to its stabilization. On the basis of this observation we expect the mutation of 14th and/or 30th residues to other (more) hydrophobic residues should have little effect on the folding free-energy landscape.

In the lowest free-energy folding pathway, Helix3 (which has the highest α-helical forming propensity) remains as the first to fold. The formation of the other two helices which are less

dx.doi.org/10.1021/jp307684h | J. Phys. Chem. B 2012, 116, 13848−13856

hydrophobic needs the assistance from the long-range hydrophobic interactions among helices. The folding of the three polypeptides is generally initiated by the structural collapse to a stable compact unfolded (P) state. In this state, Helix3 is fully folded, and the helix in the middle position of protein is also largely folded, whereas the remaining helix at the edge of protein is only at most partially formed. The P state is stabilized mainly by non-native hydrophobic contacts and only by a few native ones. The further folding of the polypeptides to their folded states requires the structure rearrangement to break non-native hydrophobic contacts and to form native ones. Accompanying the hydrophobic collapse, more helical contents of Helix1 and Helix2 become formed.

The relative weak effects of the local changes of sequence on protein structure and its folding mechanism were also seen from studies on another popular protein model: After the addition of extra amino acid residues to its C-terminus, the original sequence of Trp-cage still folds into its native structure. It would be interesting to test experimentally this prediction and also to test how general such a phenomenon is.

AUTHOR INFORMATION

Corresponding Author

*E-mail: gaoyq@pku.edu.cn.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the financial support from the National Key Basic Research Foundation of China (2012CB917304 to Y.Q.G.) and the National Natural Science Foundation of China (Grant Nos. 21125311 for Y.Q.G, 21003003 for Q.S., and 2012AA01A305 for W.Z.). The simulations were run at the Shanghai Supercomputer Center and TianHe 1 supercomputer in Tianjian.

REFERENCES

(1) Gromiha, M. M.; Selvaraj, S. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235−277.
(2) Baldwin, R. L. *Science* **2002**, *295*, 1657−1658.
(3) Dougan, L.; Li, J. Y.; Badilla, C. L.; Berne, B. J.; Fernandez, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 12605−12610.
(4) Faure, G.; Bornot, A.; de Brevern, A. G. *Biochimie* **2008**, *90*, 626−639.
(5) Zhou, Y. Q.; Duan, Y.; Yang, Y. D.; Faraggi, E.; Lei, H. X. *Theor. Chem. Acc.* **2011**, *128*, 3−16.
(6) Dor, O.; Zhou, Y. Q. *Proteins* **2007**, *66*, 838−845.
(7) Fiser, A.; Dosztanyi, Z.; Simon, I. *Comput. Appl. Biosci.* **1997**, *13*, 297−301.
(8) Zhou, Y. Q.; Karplus, M. *Nature* **1999**, *401*, 400−403.
(9) Zhou, Y. Q.; Karplus, M. *J. Mol. Biol.* **1999**, *293*, 917−951.
(10) Alonso, D. O. V.; Daggett, V. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 133−138.
(11) Garcia, A. E.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13898−13903.
(12) Yang, J. S.; Wallin, S.; Shakhnovich, E. I. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 895−900.
(13) Gouda, H.; Torigoe, H.; Saito, A.; Sato, M.; Arata, Y.; Shimada, I. *Biochemistry* **1992**, *31*, 9665−9672.
(14) Bai, Y. W.; Karimi, A.; Dyson, H. J.; Wright, P. E. *Protein Sci.* **1997**, *6*, 1449−1457.
(15) Shao, Q.; Gao, Y. Q. *J. Chem. Phys.* **2011**, *135*, 135102/1−135102/13.
(16) Gao, Y. Q. *J. Chem. Phys.* **2008**, *128*, 064105/1−064105/6.
(17) Gao, Y. Q.; Yang, L. J.; Fan, Y. B.; Shao, Q. *Int. Rev. Phys. Chem.* **2008**, *27*, 201−227.
(18) Dimitriadis, G.; Drysdale, A.; Myers, J. K.; Arora, P.; Radford, S. E.; Oas, T. G.; Smith, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3809−3814.
(19) Neidigh, J. W.; Fesinmeyer, R. M.; Andersen, N. H. *Nat. Struct. Biol.* **2002**, *9*, 425−430.
(20) Culik, R. M.; Serrano, A. L.; Bunagan, M. R.; Gai, F. *Angew. Chem., Int. Ed.* **2011**, *50*, 10884−10887.
(21) Qiu, L. L.; Pabit, S. A.; Roitberg, A. E.; Hagen, S. J. *J. Am. Chem. Soc.* **2002**, *124*, 12952−12953.
(22) Case, D. A.; Darden, T. A.; Cheatham, T. E. III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; et al. *AMBER*, 9th version; University of California: San Francisco, 2006.
(23) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383−394.
(24) Shao, Q.; Yang, L. J.; Gao, Y. Q. *J. Chem. Phys.* **2009**, *130*, 195104/1−195104/6.
(25) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; et al. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(26) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.
(27) Kannan, S.; Zacharias, M. *Int. J. Mol. Sci.* **2009**, *10*, 1121−1137.
(28) Marinelli, F.; Pietrucci, F.; Laio, A.; Piana, S. *Plos. Comput. Biol.* **2009**, 5.
(29) Chowdhury, S.; Lee, M. C.; Duan, Y. *J. Phys. Chem. B* **2004**, *108*, 13855−13865.
(30) Duan, L. L.; Mei, Y.; Li, Y. L.; Zhang, Q. G.; Zhang, D. W.; Zhang, J. Z. *Sci. Chin. Chem.* **2010**, *53*, 196−201.
(31) Duan, Y.; Lei, H. X.; Wu, C.; Wang, Z. X.; Zhou, Y. Q. *J. Chem. Phys.* **2008**, 128.
(32) Chowdhury, S.; Lei, H. X.; Duan, Y. *J. Phys. Chem. B* **2005**, *109*, 9073−9081.
(33) Shao, Q.; Gao, Y. Q. *J. Chem. Theory Comput.* **2010**, *6*, 3750−3760.
(34) Shao, Q.; Wei, H. Y.; Gao, Y. Q. *J. Mol. Biol.* **2010**, *402*, 595−609.
(35) Shao, Q.; Yang, L. J.; Gao, Y. Q. *J. Chem. Phys.* **2011**, *135*, 235104/1−235104/10.
(36) Chen, L. X.; Shao, Q.; Gao, Y. Q.; Russell, D. H. *J. Phys. Chem. A* **2011**, *115*, 4427−4435.
(37) Feig, M.; Karanicolas, J.; Brooks, C. L. I. *MMTSB Tool Set*; MMTSB NIH Research Resource, The Scripps Research Institute, 2001.
(38) Jang, S. M.; Kim, E.; Shin, S.; Pak, Y. *J. Am. Chem. Soc.* **2003**, *125*, 14841−14846.
(39) Vila, J. A.; Ripoll, D. R.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14812−14816.
(40) Sobolev, V.; Eyal, E.; Gerzon, S.; Potapov, V.; Babor, M.; Prilusky, J.; Edelman, M. *Nucleic Acids Res.* **2005**, *33*, W39−W43.
(41) Garcia, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345−354.
(42) Bunagan, M. R.; Yang, X.; Saven, J. G.; Gai, F. *J. Phys. Chem. B* **2006**, *110*, 3759−3763.
(43) Bunagan, M. R.; Gao, J. M.; Kelly, J. W.; Gai, F. *J. Am. Chem. Soc.* **2009**, *131*, 7470−7476.