# Comparative all-atomic study of unfolding pathways for proteins chymotrypsin inhibitor 2 and barnase

Yuebiao Sheng[1] and Wei Wang[1,2,*]
[1]*National Laboratory of Solid State Microstructure, Institute of Biophysics, and Department of Physics, Nanjing University, Nanjing 210093, China*
[2]*Interdisciplinary Center of Theoretical Studies, Chinese Academy of Sciences, Beijing 100080, China*
(Received 22 August 2005; revised manuscript received 11 October 2005; published 27 February 2006)

The features of transition states and intermediates are important in the study on protein folding. However, transition states and intermediates could not be obviously identified from trajectories obtained by dynamic simulations. In this work, a different method to identify and characterize the transition states and intermediates by combining the root mean square deviation of $C_\alpha$ atoms and the similarity factor $Q$ to the native state is proposed. The unfolding processes based on all-atomic simulations for proteins chymotrypsin inhibitor 2 and barnase are studied, and the related transition states and intermediates are identified by observing an unfolding factor $U=1-F$. Comparisons between the conformational cluster analysis and experimental results are also made. The various analyses on the unfolding behaviors indicate that our method can well define the transition states and intermediates, and the factor $U$ (or $F$) can be used as a reaction coordinate of the folding and unfolding process. It is also found that three-state folding proteins might experience more complicated pathways and have more rugged energy landscapes than two-state folding proteins.

DOI: 10.1103/PhysRevE.73.021915 PACS number(s): 87.15.He, 87.15.Aa, 87.10.+e

## I. INTRODUCTION

Protein folding is a fundamental, important, and unsolved problem in molecular biology. In the past two decades, much progress has been achieved. The study realm of protein folding has extended widely in recent years, and much of the field has moved on to questions such as structural dynamics, molecular recognition, and folding diseases, etc. [1–9]. Nowadays, it is generally accepted that folding resembles a diffusive process on a rugged funnel-like energy landscape [10–18]. It is also known that the time scale of protein folding events in cells is about several milliseconds to tens of seconds. For the study on protein folding, simulations based on all-atom models are very important since much detailed information of folding can be obtained. Some empirical all-atomic force field methods have been proposed and used to calculate the energy of a protein system as a function of the atomic positions for characterizing the folding. However, the all-atom simulations can only reach a time scale of several nanoseconds in one run or several microseconds by combining a number of runs. To characterize the nature of the energy landscape and the kinetics of folding, ensemble averaging over a number of simulations or long time running with full atomic representation of both protein and solvent is still quite difficult and beyond the current computational capacity. Therefore the all-atom simulation of protein folding is not a trivial work to study the folding processes in detail. Since the first molecular dynamics (MD) simulation made for the bovine pancreatic trypsin inhibitor (BPTI) about 27 years ago, many efforts have been done to increase the accuracy of the various all-atomic models, to improve the methods of the computations, and to extend the time scale of the simulations [19–21]. Recently, some achievements have been reported. Duan and Kollman characterized the folding in a 1-$\mu$s all-atomic simulation for a protein HP36 in aqueous solution in 1998 [22,23]. Pande and his collaborators implemented all-atomic folding simulations on the submillisecond time scale using world wide distributed computing in 2003 [24]. Now it seems to be possible to study some ultrafast-folding proteins which are accessible on microsecond time scale [25].

Although it is very difficult to study the folding processes directly by the all-atomic folding simulations, we can study partially the folding nature, such as transition states, intermediates and conformation changes, by the all-atomic unfolding simulations at high temperature. As we know, if we have an infinitely long run of the MD simulation at room temperature and the ergodic hypothesis holds, the trajectory may sample all of conformational space in principle. However, if the simulation time is short, the trajectory obtained from the simulation at room temperature will not cover all the conformational space, and will be restricted to some limited regions around local minima due to the effects of high energy barriers. A common solution to sample a broad conformational space is simply to raise the temperature of the simulation. In such a way the system is "shaken" and the molecule will be more likely to cross the high energy barriers [26].

In the folding and unfolding study, it is important to characterize the features of transition states (TS) and intermediates (I). Here the transition state is related to the main barrier, which affects the folding dynamics [10,11,13,18,27] on the free energy landscape. The intermediates involve the local minima around the main barrier of the free energy landscape. The study on the transition states and intermediates is essential for understanding the mechanism of protein folding, and has become a hot topic in recent years. However, the features of the transition states and intermediates are difficult to be characterized. This is due to that, on one hand, the transition

*Electronic address: wangwei@nju.edu.cn

 

states and intermediates could not be obviously identified from the trajectories obtained by dynamic simulations; on the other hand, they could not be easily obtained thermodynamically as the result of the difficulties in deriving the free energy landscape. Previously, a method based on the variation of the root mean square deviations (RMSDs) of the $C_\alpha$ atom coordinates was used to find the transition states and intermediates [28,29]. However, this method is not well feasible for identifying the transition states and intermediates because of the large fluctuation in the values of RMSD. In some cases, it is even hard to determine exactly the location of the transition states and intermediates in the RMSD profile.

In this work, we suggest a different method to identify and characterize the transition states and intermediates by combining the RMSD of $C_\alpha$ atom coordinates and the similarity factor $Q$ to the native state. In this method, we define a new factor $F$, namely the fraction of the native structure, as a reaction coordinate to identify the transition states and intermediates. We then describe the unfolding processes using a related factor $U=1-F$. We obtain a set of conformations as the transition state ensembles when the values of $U$ for these conformations show a sharp jump from a plateau. Differently, the intermediates are identified as a set of conformations when the related values of $U$ have small fluctuation around a certain average value for a rather long period, say 150 ps. Based on such a method, we make a comparative study on the folding and unfolding pathways for two typical proteins, namely protein chymotrypsin inhibitor 2 (CI2) and protein barnase. Both the transition states and intermediates of CI2 and barnase are obtained distinctly using our method. To check the validity of this method, we also make the conformational cluster analysis. We find out that the three-state folding protein barnase experiences a more complicated and rugged pathway. These agree well with those observed in experiments [30–37]. Especially, the identified transition states and intermediates are in good agreement with those found in experiments, and the correlation coefficients between our theoretically calculated $\Phi_{MD}$ values and the experimentally derived $\Phi_F$ values are larger than 0.92. Thus our study indicates that our method can identify the transition states and intermediates from the all-atom simulations more clearly and effectively.

## II. METHODS

### A. MD simulations

MD simulations are performed using the program CHARMM [38] with the all-hydrogen parameter set PARAM22. Both CI2 and barnase are solvated by the TIP3P explicit water molecules. The SHAKE algorithm is also used to fix the lengths of bonds involving hydrogen atoms. Thermal unfolding of these two proteins is performed using constant pressure and temperature dynamic technique. Six runs of simulations for each protein are carried out at 500 K. For these six runs, two are for the canonical ensemble (NVT) and the other four are for the isobaric-isothermal ensemble (NPT).

The initial structures of two proteins are the same as the native crystal structures from the Brookhaven Protein Data Bank [39] (for CI2, PDB code: 2CI2 and for barnase, PDB code: 1BNI). For each run, the initial structure is minimized for 1000 cycles to reduce the bad contacts. Then the protein is solvated in a body-centered-cubic box of explicit water molecules with the shortest distance between any protein atoms of the initial structure and the edge of the box larger than 8 Å at the beginning of the simulation to keep the protein not run out of the boundary of the box (for two of the six runs, 10 Å is used for each protein). Periodic boundary conditions are used to minimize the boundary effects. The particle mesh Ewald algorithm is used for calculating the electrostatic interactions [40]. The entire system is further minimized for another 1000 cycles. After these preparations, the system is heated to 500 K, and stays at this temperature for equilibrium for about 50 ps. Then the productive simulations are followed. During the simulations, a 10-Å effective nonbonded cutoff distance is used, and the nonbonded list is updated when necessary. The cutoff distance for image atoms is 12 Å, which is larger than the nonbonded cutoff 10 Å. This is to ensure that outside the 8-Å buffer between the protein and box edge, there will be image atoms interacting with the protein. The 8- and 10-Å buffers are found to have very similar behaviors. The detailed discussions about the setup of cutoff distances can be found in Ref. [41]. Structures of the protein molecule obtained from simulations are saved every 1 ps for analysis. The time steps of integration for the potential energy of protein CI2 and barnase are taken as 1 and 2 fs, respectively. The simulations are performed for 4 ns each at 500 K.

### B. Definition and calculation of $Q$ and $F$

The factor $Q$ is the fraction of the native contacts formed in a conformation, $Q=1$ corresponds to the native state and $Q=0$ to the fully unfolded state [42]. To obtain the value of $Q$, we can simply use the formula

$$Q_t = N_t/N_{total}, \tag{1}$$

where $N_t$ represents the number of native contacts in a certain conformation at time $t$, and $N_{total}$ represents the total number of native contacts in the native conformation. The native contact is defined when the distance between any pair of heavy atoms of two residues $i$ and $j$ ($j>i+2$) is less than 5.4 Å.

Although the reaction coordinate $Q$ can describe the similarity between structures at a certain extent, it is not on the basis of the conformational coordinates. It only relies on the number of the native contacts, or in other words, only relies on the distances between parts of the residues which form contacts in the native structure. Such a treatment may lose some information about the structures of other parts of the protein since only the native contacts are considered. Differently, the RMSD denotes the dissimilarity of two structures, which takes the whole details of protein structures. However, it is not easy to identify the transition states and intermediates by using only one of the factors $Q$ and RMSD. If $Q$ is used as the reaction coordination, the profile of $Q$ versus

time after the transition state is relatively flat, and it is hardly to find out the intermediates. While RMSD is used as the reaction coordinate, it is difficult to identify the transition states and intermediates distinctly due to the larger fluctuations in the values of RMSD.

By combining both $Q$ and RMSD, we define a factor $F$, namely the fraction of the native structure, as a reaction coordinate of the folding process,

$$F(t) = (Q - D_{\mathrm{RMS}}(t)/D_{\max\ \mathrm{RMS}} + 1)/2. \tag{2}$$

Here $D_{\mathrm{RMS}}(t)$ represents the root mean square deviation between a structure at time $t$ and the reference structure, and $D_{\max\ \mathrm{RMS}}$ represents the maximal value of $D_{\mathrm{RMS}}(t)$. Thus the unfolding process can be described clearly using a factor $U$,

$$U(t) = 1 - F(t). \tag{3}$$

Obviously, $F$ will approach 1 when the protein folds to its native state and 0 when the protein is fully unfolded. Similarly, during the unfolding process, $U$ will be close to 1 when the protein is fully unfolded. It should be noted that the value of $F$ will never equal to 1, but just approach 1 gradually. From the definition of $F$ or $U$ above, it is clear that both the important native contacts and the information of the whole structure of the protein are included. Thus the profile of $U$ versus time could present some obvious features of the protein structure and is a rational way to identify the transition states and intermediates. That is, in the profile a sharp jump would appear at the transition state region, and a relatively small variation of $U$ values will be observed for a long period of time which corresponds to the intermediate. These can be distinguished easily from other parts of the profile. Therefore the transition states and intermediates can be identified more distinctly or even exactly. The factors $F$ and $U$ are more suitable reaction coordinates to describe the folding and unfolding process.

### C. Definition and calculation of $\Phi_{MD}$

In order to compare our results with experiment, we use the simulated $\Phi$ values (termed $\Phi_{MD}$ here) to characterize the transition states and intermediates. This $\Phi_{MD}$ have been used intensively by Daggett and co-workers [30,43,44], and is a similar structural interpretation to $\Phi_F$ provided by the protein engineering experiment [44]. In experiment, the $\Phi_F$ values are derived from the free energy changes of the $X$ state (i.e., the transition state or intermediate) to the native state by

$$\Phi_X = 1 - (\Delta\Delta G_{X-F}/\Delta\Delta G_{U-F}). \tag{4}$$

In simulations, similar to Daggett *et al.*'s definition [43,44], the $\Phi_{MD}$ is defined as below:

$$\Phi_{MD} = \frac{N_{\ddagger,wt} - N_{\ddagger,mut}}{N_{N,wt} - N_{N,mut}} = \frac{\Delta N_{\ddagger}}{\Delta N_N}, \tag{5}$$

where ‡ represents *TS* or *I*, for characterizing the transition states and intermediates, respectively. The contact is defined as the same as for the native contacts, i.e., when the distance between any pair of heavy atoms of two residues $i$ and $j$ ($j > i+2$) is less than 4.5 Å. Note that the contacts in the transition states or in the intermediates can be either native or non-native. In Eq. (5), $\Delta N_N$ is the difference in the number of native state contacts, and is calculated through the total number of contacts for the residue of interest in native state of the wild type ($N_{N,wt}$) and that in native state of the mutant ($N_{N,mut}$). The mutant is constructed by replacing the wild-type residue with a mutant residue, the orientation of the side-chain of the mutant residue is retained as in the wild-type protein. The resulted structure is then minimized for ten steps to release any bad contacts that have been introduced. To the calculation of $\Delta N_{\ddagger}$, the same procedure is applied.

Each $\Phi_{MD}$ value is averaged on a set of structures (structure ensemble). The $\Phi_{MD}$ values can be larger than 1 if the mutant residue has more nonbonded contacts in the transition or intermediate state than it has in the native protein. The calculated $\Phi_{MD}$ values are in good agreement with experiment for hydrophobic deletion mutations, however, it should be noted that to interpret $\Phi_{MD}$ for other mutations is difficult, because no account is made for secondary structures, solvent effects, and other interactions related to the surface residues of the protein. In other word, such a definition of $\Phi_{MD}$ is best suitable for hydrophobic deletion mutations, especially mutations inaccessible to solvent.

For hydrophobic deletion mutations, the accuracy of the simulated $\Phi_{MD}$ values will only rely on two aspects: the validity of the definition for $\Phi_{MD}$, and the veracity of the identification for the transition state or intermediate. The validity of the definition has been discussed in detail in Ref. [43]. Thus upon the same definition of $\Phi_{MD}$, the identification of the transition states or intermediates will play a key role in determining the $\Phi_{MD}$ values. Alternatively, for a better determined transition state or intermediate, $\Phi_{MD}$ will have higher correlation coefficient with the experimentally obtained $\Phi_F$.

### D. Conformational cluster analysis

The unfolding processes can be further studied via the conformational cluster analysis which uses a set of points to fit the RMSD values between a set of protein structures. Assuming that the RMSD value between any two structures $i$ and $j$ is defined as a distance $D_{ij}$, we can use two points in a three-dimensional space to represent these two conformations by their distance. That is, the distance between the points $(x_i, y_i, z_i)$ and $(x_j, y_j, z_j)$ in the three-dimensional space is equal to $D_{ij}$. Thus all the distances between all pairs of conformations can be assigned as a set of points in a three-dimensional space. Obviously, points close in space indicate that the related conformations are similar. Thus a cluster of points represent conformations that are similar to each other [43,45]. From this method, we know that the transition state corresponds to a watershed of two clusters of conformations. Once the protein leaves the first cluster to the second one, it will never return. In addition, since the intermediate represents a set of similar conformations, the points, which correspond to the conformations of the intermediate, will gather into a relatively concentrated region.

## III. RESULTS

Now we report the detailed study on the unfolding processes of protein CI2 and protein barnase which behave typi-

cal two-state and three-state folding behaviors, respectively. The reasons for choosing these two proteins are that they are two representative proteins of two-state and three-state folding, respectively, and have been studied intensively. Our simulations can be well compared with the previous results by others. A two-state folding means that there is a main barrier between the denatured and the native states. Differently, for a three-state folding, there will be a local minimum of free energy related to the intermediate. In order to characterize the transition states and intermediates, except for the factor $U$, the radius of gyration $R_g$ and solvent accessible surface area (SASA) are also used. The compactness of a conformation can be understood quantitatively by monitoring the radius of gyration $R_g$. Thus the time evolution of the radius of gyration $R_g$ can be regarded as a good measure of the dynamics of collapse for a protein. The solvent accessible surface area of a protein is a factor to measure the hydrophobic effect. The hydrophobic force is considered to be one of the most important one among various forces that determine the tertiary structures of proteins. The hydrophobic effects cause nonpolar side chains to tend to cluster together in the protein interior. The accessible surface is part of the complex surface in direct contact with solvent. It is traced out by the probe sphere center as it rolls over the protein. An atom or a group of atoms is defined as accessible if a solvent (water) molecule of specified size can be brought into the van der Waals' contact. Therefore the accessible surface is a kind of expanded van der Waals surface. The overlapping surface of envelope by the neighboring atoms are eliminated from the area summation.

### A. Unfolding process of CI2

CI2 is of small single-domain protein with only 65 residues and behaves the two-state folding. CI2 contains an $\alpha$ helix and a three-stranded $\beta$ sheet, and the helix packs against the $\beta$ sheet to form the major hydrophobic core [28,43] [cf. Fig. 1(a)]. One of the six simulations has been extended to 6 ns for studying the unfolding. The change of $U$ as a function of time is shown in Fig. 2(a), three main stages corresponding to three large variations in structures can be clearly observed in the figure. The first plateau spans from $t=70$ to 450 ps, the second one spans from $t=950$ to 2400 ps, and the third one from $t=2800$ to 4000 ps.

The initial rapid increasing in $U$ for about 70 ps from $t=0$ is mainly due to the temperature equilibrium. The structure of the protein extends in response to the increasing in the temperature. As a result, the hydrophobic core expands quickly [cf. Fig. 2(e)], and this process lasts for about 25 ps. After that period of time, the structure has well equilibrated and then is kept to be relatively stable from $t=70$ to 450 ps. At about $t=210$ ps, the $N$ terminus of the protein becomes more mobile, and begins to move away from the hydrophobic core. That is, after the unfolding of the $N$ terminus, the hydrophobic core becomes unstable, especially in the time period from $t=470$ to 510 ps. At about $t=600$ ps, the hydrophobic core begins to expand rapidly, water molecules enter the core occasionally, and at about $t=650$ ps the hydrophobic core is disrupted. After that a new, more dynamic, hydrophobic core forms. Thus a sharp peak appears in the plot of the solvent accessible surface area. Between $t=580$ and 680 ps, there is also a sharp increasing in the factor $U$, corresponding to the opening of the hydrophobic core.

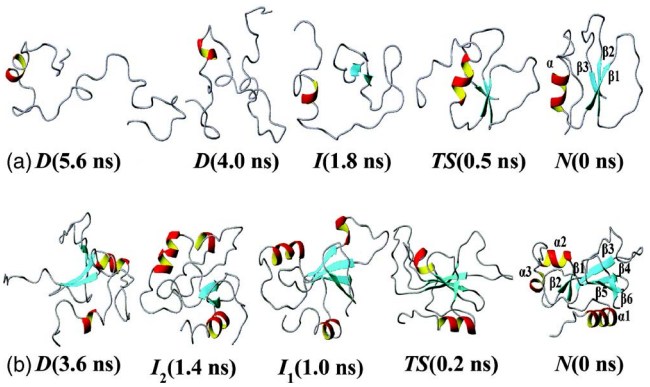


FIG. 1. (Color online) The folding pathways of CI2 and barnase. $N$, $TS$, $I$, and $D$ correspond to the native state, the transition state, the intermediate state, and the denatured state, respectively. (a) Snapshots of CI2 (omitting the water) from one of the unfolding simulations at 500 K, which are presented in reverse time. CI2 consists of an $\alpha$ helix (residues 14–24), and three $\beta$ strands: $\beta_1$ (residues 28–35), $\beta_2$ (residues 46–52), $\beta_3$ (residues 62–64). (b) Snapshots of barnase (omitting the water) from a 500-K unfolding simulation, presented in reverse time. Barnase contains three helices and six strands: $\alpha_1$ (residues 7–17), $\alpha_2$ (residues 27–33), $\alpha_3$ (residues 42–45), $\beta_1$ (residues 23–25), $\beta_2$ (residues 49–51), $\beta_3$ (residues 70–75), $\beta_4$ (residues 86–91), $\beta_5$ (residues 95–99), $\beta_6$ (residues 106–108). All the graphics presented above were constructed by the software MOLMOL [49].

From $t=1720$ to 1890 ps, an intermediate is observed. During this period of time, all of the solvent accessible surface area, the radius of gyration, and the factor $U$ are kept to be relatively stable. But due to the short lasting time, such an intermediate is hardly to be observed in experiment. After the intermediate state, the new hydrophobic core unfolds furthermore and $\beta_1$ is fully disrupted. A small number of water molecules have entered the new core, and interact with the core residues. But for a moment, $\beta_1$ forms partly again and interacts with $\beta_2$, corresponding to a transient decreasing in the solvent accessible surface area. Finally, a rapid variation of the structures occurs between $t=2400$ and 2800 ps, and the native structure is lost step by step till the protein is fully unfolded.

### B. Unfolding process of barnase

Barnase is of multidomain with 110 residues and shows three-state folding behavior. Barnase is comprised of three $\alpha$ helices and a five-stranded $\beta$ sheet [cf. Fig. 1(b)]. There are three main hydrophobic cores. The first core ($\text{core}_1$) is formed by the helix $\alpha_1$ packing against one side of the $\beta$ sheet. The second core ($\text{core}_2$) contains the hydrophobic residues from $\alpha_2$, $\alpha_3$, $\beta_1$, $\beta_2$, and the loops between them. The loops between $\beta_2$ and $\beta_3$ consist of the third core ($\text{core}_3$) [29,48].

After the protein has well equilibrated, the structure keeps relatively stable for about 100 ps (from $t=34$ to 132 ps), the

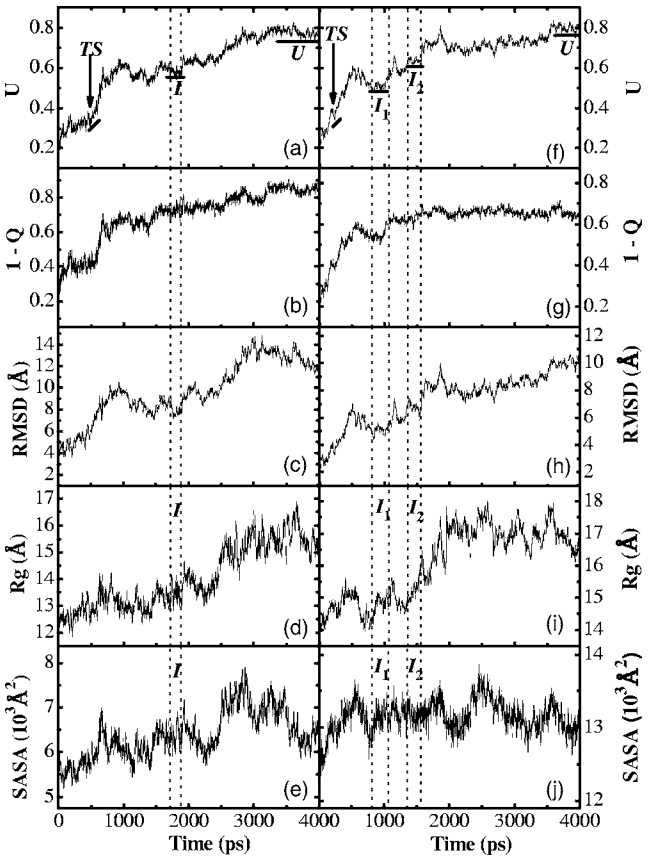


FIG. 2. Global properties of the proteins CI2, (a)–(e), and barnase, (f)–(j), as a function of time. The broken lines in (a)–(e) mark the $t=1720$ and $t=1890$ ps time points, and the region between them corresponds to the intermediate $I$ of CI2; the region between the broken lines marked at $t=810$ and $t=1070$ ps in (f)–(j) corresponds to the intermediate $I_1$ of barnase, and the region between the broken lines marked at $t=1350$ and 1560 ps corresponds to the intermediate $I_2$. The radius of the probe to calculate the solvent accessible surface area is specified to be 1.6 Å.

various conformations in this period of time are similar, with the values of $U$ around 0.27. Then the value of $U$ increases rapidly from $t=150$ to 460 ps. Three obvious halts from the plot of $U$ are observed at this phase, corresponding to the disruption of the three hydrophobic cores one by one.

About 350 ps later, the value of $U$ reaches a plateau, with an average value of about 0.5, which lasts for about 250 ps until $t=1070$ ps. During this period of time, the solvent accessible surface area also keeps relatively stable but the radius of the gyration fluctuates slightly. This part of the trajectory might correspond to the unfolding intermediate $I_1$. At this phase all the three hydrophobic cores are disrupted, and $\alpha_1$ loses the $N$-terminal turn, $\beta_1$ is fully unfolded, but the cent of the protein remains intact.

Between $t=1000$ and 1160 ps, another rapid increasing in $U$ value occurs due to the complete broken of the $\beta$ sheet. At about $t=1350$ ps, another plateau of $U$ arises and keeps relatively stable until $t=1560$ ps. Accordingly, the solvent accessible surface area fluctuates at a small extent. The average value of $U$ is about 0.63 at this phase, which indicates that most of the native structure has been lost. During this time period, the $\beta$ sheet has been completely disrupted, all the three helices are perturbed and the hydrophobic $\text{core}_3$ is lost. However, some hydrophobic contacts of $\text{core}_1$ and $\text{core}_2$ still remain. This phase corresponds to another unfolding intermediate $I_2$, but this intermediate is less structured and lasts for less time than $I_1$. About 420 ps later, the value of $U$ reaches another plateau, corresponding to the denatured state. Finally, until $t=3560$ ps, the protein reaches the unfolded state.

### C. Characterization of transition states and intermediates

The transition state ensembles are defined as a set of conformations when the related $U$ values of these conformations have risen up to a relatively stable value, and will increase rapidly but will not decrease to this value again. According to such a definition, the transition state ensembles are found out as from $t=495$ to 530 ps for CI2 and $t=190$ to 212 ps for barnase as indicated in Fig. 2. The transition state of barnase corresponds to the beginning of the disruption of one of the hydrophobic cores.

Since the value of $U$ represents a fraction of the native structure, the intermediates can be identified more easily based on the fluctuation of $U$. When the values of $U$ fluctuate within a certain extent around the average value for a certain time period, we consider the conformations of this phase composing the intermediate state. Here we set the upper and lower limit of the fluctuation of $U$ to be $\pm 0.02$, and the lasting time period does not less than 150 ps. So the phase from $t=1720$ to 1890 ps is considered to be the intermediate state $I$ for CI2; the phase from $t=810$ to 1070 ps is considered to be $I_1$, and $I_2$ is from $t=1350$ to 1560 ps for barnase. The intermediate $I$ of CI2 and the intermediate $I_2$ of barnase last for a shorter period of time, so it is difficult to be observed in experiment. However, the intermediate $I_1$ of barnase, called the major intermediate, lasts for longer time, and is easy to be observed in experiment. $I_1$ is the early unfolding intermediate which corresponds to the late folding intermediate in the folding direction, such as observed in the folding simulations and the protein engineering experiments. These intermediates are all dynamical ones, and they do not represent the thermodynamic folding behaviors. Among the above identified intermediates, only the intermediate $I_1$ of barnase can be observed in experiment, which corresponds to the thermodynamic intermediate. The other two dynamical intermediates will not be identified by defining longer lasting time period. This confirms the two-state folding behavior of CI2 and the three-state folding behavior of barnase.

To check the validity and make a comparison, the conformational cluster analysis is used for CI2 and barnase, respectively (cf. Fig. 3). Obviously, for CI2 the time points are distributed in three clusters. When the time points pass through the transition state at about $t=510$ ps, they do not come back again. After that, an intermediate appears. In Fig. 3(b), four clusters are observed. After passing through the transition state, the time points populate to form the major intermediate $I_1$. Then, another cluster lasting for a short time period is observed. All these are consistent with the results from the analyses based on the factor $U$.

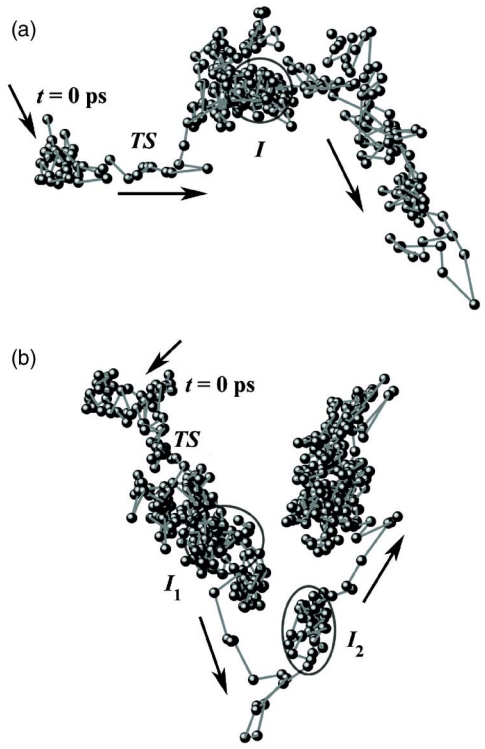


FIG. 3. (Color online) Conformational cluster analysis of protein structures during the unfolding simulations of CI2 (a) and barnase (b). The points are connected sequentially in time (15 ps/point for CI2 and 5 ps/point for barnase).

Note that for the identification of the intermediates, the selection of the upper and lower limit of the fluctuation of $U$ is based on the result of our detailed computations and analyses. To identify the intermediates more accurately, nine different upper and lower limits (or intervals) are set as ±0.010, ±0.015, ±0.020,…, ±0.050, respectively. For each interval, the computations are performed on different trajectories of protein barnase to find out the intermediates (with the lasting time 150 ps). Then, the theoretical $\Phi_{MD}$ values for the major intermediate of barnase are calculated by combining all the trajectories using Eq. (5), and the related correlation coefficients between theoretical and experimental $\Phi$ values are derived for these nine intervals, respectively. It is found that the highest correlation is related to the upper and lower limit ±0.02 for protein barnase. However, such a method based on the correlation is not workable for protein CI2, since there are no experimental data of $\phi$ values for the intermediate of protein CI2 (note that protein CI2 is a two-state folder, but it has an implicit and transient kinetic intermediate. Such a kinetic intermediate has been mentioned in Li and Daggett's studies [43]). For our case, the intermediate of protein CI2 could be well found with the upper and lower limit ±0.02 based on the same computations as for protein barnase, and this intermediate is the best one with respect to that identified by conformational clustering method. Thus the upper and lower limit ±0.02 is selected for identifying the intermediates in our simulations. In addition, the time interval is set as a value at least 150 ps since the kinetic intermediate (including implicit kinetic intermediate) can be sampled within such a time interval.

It is also worthy to note that the definition of transition state is based on the $S$-type transition of the unfolding $U$ curve. The $S$-type transition of the unfolding $U$ curve has two baselines, one is before the transition (namely the lower baseline), and the other after the transition (namely the upper baseline). Similar to the calculation of the Fermi radius in condensed matter physics, the starting point of the transition state is defined as the point with a value of 10% higher than the lower baseline, and the ending point of the transition state as the point with a value of 10% lower than the upper baseline. Thus the width (or the interval) of the transition state is defined as from the starting point to the ending point of the transition. The midpoint of the transition is based on the first-order derivatives of the curve, i.e., the midpoint of the transition state should have the maximal first-order derivative.

## IV. DISCUSSION

### A. Comparison of the dynamic behaviors of CI2 and barnase

Since CI2 is a single domain protein and contains only one hydrophobic core, after the transition state region, the value of $U$ changes sharply until reaching the second plateau in the above simulations. Thus no stagnation is observed in the process. As described in Sec. III A, the transition state of CI2 corresponds to the beginning of the disruption of the hydrophobic core. However, the situation is not the same for barnase. Since barnase is a multidomain protein and contains three hydrophobic cores, the values of $U$ do not rise up to the second plateau directly after the transition state, and two transient breaks are observed at about $t=210$ and 400 ps. This indicates that the disruption of three hydrophobic cores do not occur simultaneously, but sequentially.

Following the processes mentioned above, both CI2 and barnase begin to perform their unfolding furthermore. Then, larger fluctuations for barnase, but relatively small fluctuations for CI2 are observed in the profile of $U$ [cf. Figs. 2(a) and 2(f)]. Even in the intermediate regions, larger fluctuations of $R_g$ and SASA can also be observed for barnase [cf. Figs. 2(d), 2(e), 2(i), and 2(j)]. In addition, more obvious surge is observed in the profile of $R_g$ for barnase. This implies that more cragged barriers block the latter folding in the folding direction of barnase. We have calculated the linear correlation coefficients between $U$ and the time $t$ for CI2 and barnase. The values are 0.66 and 0.89, respectively, and the lower of the value shows the more complicated behavior and the more rugged landscape. In a word, barnase experiences a more complicated and rugged pathway during the folding and unfolding process.

### B. Comparison with experiment

The average $\Phi_{MD}$ values and the experimental $\Phi_F$ values for CI2 and barnase are shown in Figs. 4 and 5, respectively. All the data in these two figures are for hydrophobic deletion mutations. The crystal structures from the Brookhaven Protein Data Bank are used as the reference native structures. The correlation coefficient $R$ between the experimental $\Phi_F$ values in pure water [30–37] and the average $\Phi_{MD}$ values

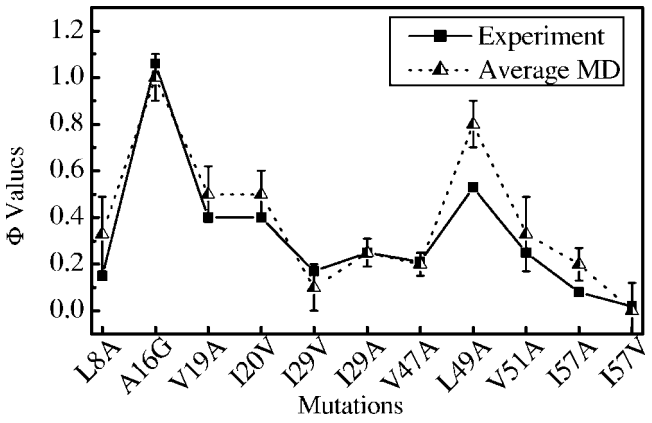


FIG. 4. Comparison of the calculated $\Phi_{MD}$ values for the transition state of CI2 with the experimentally obtained $\phi_F$ values. The $\Phi_{MD}$ values have been averaged over all simulations, and the crystal structure is used as the reference native structure.

from our simulations for the transition state of CI2 is 0.92, and the correlation coefficients for the transition state and major intermediate of barnase are 0.93 and 0.92, respectively. Note that if the mutations V45T and V45A are not included, the correlation coefficients are even larger than 0.96 (cf. Table I).

As can be seen from the two figures, the overall unfolding behaviors and the trends of structure changes for CI2 and barnase from the simulations are in good agreement with those from experiment. Nevertheless, it is worthy to note that for the mutations Val→Thr45 (V45T) and Val→Ala45 (V45A) of barnase the number of contacts of Val45 in native wild-type state equals to the number of contacts in the native mutant state, i.e., $N_{N,wt}-N_{N,mut}=0$. For such mutations the $\Phi_{MD}$ values could not be calculated correctly, thus are set as 0 (cf. Fig. 5).

These results indicate that the transition states and intermediates determined by our method are consistent with experiment. Although it is a simplified approach, it can well characterize and identify the folding and unfolding transition states and intermediates. It is worthy to note that in our six different simulations, although distinct trajectories are yielded, the transition states and intermediates identified from the trajectories are similar. For example, at the initial stage of unfolding for CI2, all simulations show a rapid expansion of the protein. After the transition state, the intermediate is observed, though it arises at different time in different simulations [cf. Fig. 2(a)]. The structures of the transition state of CI2 from different simulations are shown in Fig. 6. As can be seen, these six structures are almost the same. Likewise, the same features are observed for the transition state and intermediate of barnase [cf. Fig. 2(f)], and the corresponding structures are shown in Figs. 7(a) and 7(b), respectively.

Experimentally, the derived quantities come from the averaged behavior of a large number of molecules, while it is impossible to perform large number of simulations in theoretical computing. Here although only six simulations are performed and the simulation conditions may not be identical with the complicated environmental conditions in experiment, the higher values of the correlation coefficients between the $\Phi$ values from simulation and experiment $R \geqslant 0.92$ are still yielded. Such a good correlation really implies that our simulations have captured the essence of the folding and unfolding pathways of CI2 and barnase in the transition state and intermediate regions.

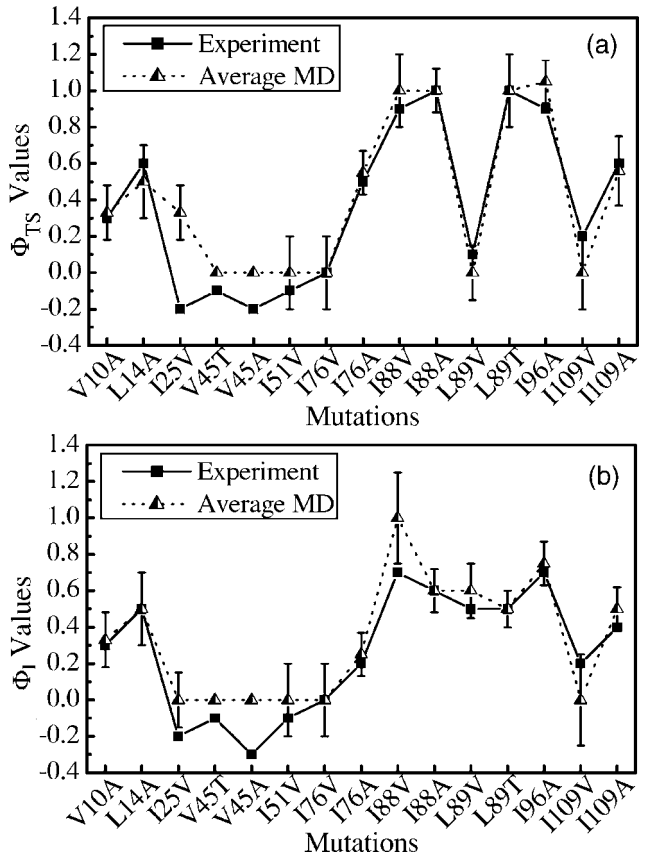


FIG. 5. (a) Comparison of the $\Phi_{MD}$ values for the transition state (labeled $\Phi_{TS}$) of barnase with the experimentally obtained $\Phi_{TS}$ values. (b) Comparison of the $\Phi_{MD}$ values for the major intermediate (labeled $\Phi_I$) of barnase with the experimentally obtained $\Phi_I$ values. The crystal structures are used as the reference native structures, and both of the $\Phi_{MD}$ values in (a) and (b) have been averaged over all simulations. It is noted that, for the mutations V45T and V45A, due to $N_{N,wt}-N_{N,mut}=0$, the corresponding $\Phi_{MD}$ values are set as 0, thus no error bars are demarcated.

TABLE I. Differences among the unfolding traces for $U$, $Q$, and RMSD. The values in the parentheses are derived by excluding the mutations V45T and V45A. The interrogation marks (?) represent that the transition and intermediate state cannot be well distinguished or the value cannot be calculated. The slashes (/) indicate that the value cannot be obtained due to the lack of experimental data $\Phi_F$.

| | Correlation coefficients between $\Phi_{MD}$ and $\Phi_F$ | | | | | |
|---|---|---|---|---|---|---|
| | Transition state | | | Major intermediate | | |
| Protein | $U$ | $Q$ | RMSD | $U$ | $Q$ | RMSD |
| CI2 | 0.92 (0.96) | 0.87 | 0.84 | / | / | / |
| Barnase | 0.93 (0.96) | 0.89 | ? | 0.92 (0.95) | ? | 0.81 |

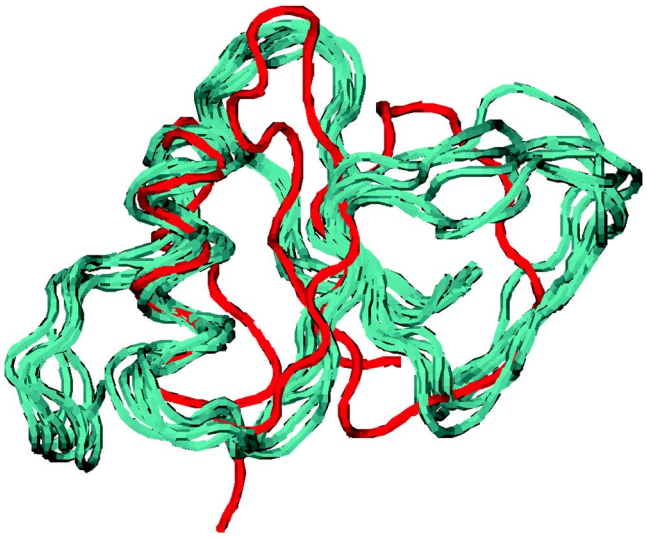

FIG. 6. (Color online) Conformational heterogeneity of the transition state for CI2. The backbone of the crystal structure is shown in red and the six transition states from different simulations are in cyan. Both Figs. 6 and 7 are constructed by the software VMD [50].

### C. Comparison with other methods

In Fig. 2, it is obvious that the transition states cannot be identified effectively using only the factor $Q$ or RMSD. From Fig. 2(g), it is clear that the transition state of barnase is hardly to be identified using $Q$ as the reaction coordination. At the same time, the identification of the intermediates for both CI2 and barnase is a bit optional [cf. Figs. 2(b) and 2(g)]. Take CI2 for example, if one only use RMSD as the criteria, the transition state region should be around $t=420$ ps [cf. Fig. 2(c)]. We have calculated the corresponding $\Phi_{MD}$ values of this region, the correlation coefficient between the $\Phi_{MD}$ values and the experimental $\Phi_F$ values only reaches 0.84. Similarly, due to large fluctuations in the values of RMSD, it is also difficult to identify the intermediates both in Figs. 2(c) and 2(h). That is why the rational way is to combine both $Q$ and RMSD for the identification of the transition states and intermediates.

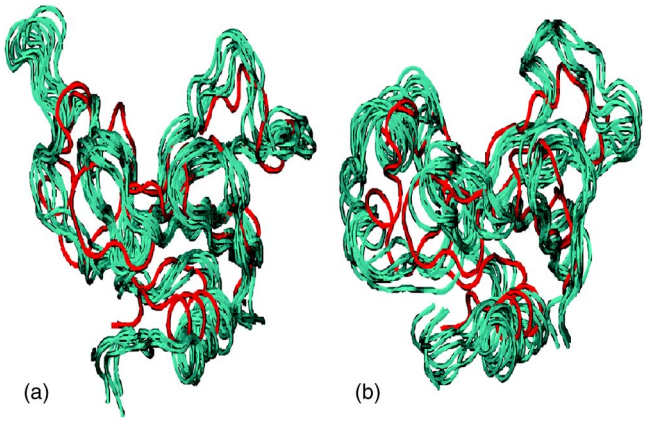


FIG. 7. (Color online) Conformational heterogeneity of the transition state (a) and intermediate (b) for barnase. The backbone of the crystal structure is shown in red, and both of the transition states in (a) and the intermediates in (b) from the six different simulations are in cyan.

Using other quantities, such as the radius of gyration and the solvent accessible surface area, can neither well identify and characterize the transition states and intermediates. The radius of gyration of an area with respect to a particular axis is the square root of the quotient of the moment of inertia divided by the area, i.e., it is related to the moment of inertia and the area. The radius of gyration denotes the compactness of a conformation. However, if the structure change during the unfolding process is large, the moment of inertia will also change unignorably accordingly. So $R_g$ cannot characterize the structure change finely, and using it as the criteria of transition states and intermediates might be problematical. As can be seen in Figs. 2(d) and 2(i), the values of radius of gyration around the intermediate regions do not keep stable at a certain value as it should be. The solvent accessible surface area is a quantity to measure the hydrophobic effect, while because of lots of noises in the time plot of SASA, it is difficult to distinguish the transition state and intermediate regions from others exactly. Nevertheless, since the solvent accessible surface area is related to the tertiary structure of proteins, it will be a good accessorial quantity to help characterize and identify the transition states and intermediates.

In addition, we have also applied the method for characterizing the transition state and the intermediates to other proteins, such as the C-terminal domains of troponin C [46]. With the same upper and lower limit for the fluctuations and lasting time, features of the transition state and the intermediates are also obtained and are relevant to the experimental observations [47] although a quantitative comparison with experiment cannot be provided due to the lack of experimental $\Phi$ values (unpublished results).

## V. CONCLUSIONS

In this work, we have introduced a convenient method for quickly identifying and nicely characterizing the transition states and intermediates based on the time evolution of fraction of the native structure $F$ (or $U$, fraction of the unfolded structure), which is a reaction coordinate for characterizing the folding (or unfolding) process. By studying the variation of $U$, the transition states and intermediates for proteins CI2 and barnase are determined. The results are consistent with those obtained from conformational cluster analysis. To verify the validity of our method, we also calculate the $\Phi$ values, i.e., $\Phi_{MD}$, for the hydrophobic deletion mutations, and make a comparison with the experimental $\Phi_F$ values. A correlation coefficient $R \geqslant 0.92$ is reached. This indicates that the identified transition states and intermediates are well relevant to those found in experiment [30–37]. We also find that using any one of the factors $Q$ and RMSD could not well identify the transition states and intermediates as our method works. A further comparison with different methods based on other geometrical variables is also made. All these imply that our method is more convenient, distinct, and exact than others. In addition, in our study, the time evolution of the radius of gyration and solvent accessible surface area is also used to characterize the transition states and intermediates.

As two remarks, it is worthy to note that the unfolding simulations at high temperature can speed up the searching in the conformational space and can reach many conformations. However, some of conformations related to the room temperature could not be reachable [26]. Thus the simulated trajectories should be used carefully to describe the dynamics at room temperature. To partially overcome such difficulty, simulated annealing is used after the productive simulations in this work. Additionally, the identification of the transition states and intermediates based on dynamic trajectories is an approximation, and may not ensure that all the trajectories will have well-defined transition states and intermediates. To make a better determination, some other features should be used simultaneously, like the various factors applied in this work. Thus the best way is to combine the thermodynamic quantities at the same time.

In conclusion, the two-state folding behavior for CI2 and the three-state folding behavior for barnase have been characterized clearly in this paper, which verifies that our method for identifying the transition states and intermediates is high up in the pictures of protein folding and unfolding.

## ACKNOWLEDGMENTS

This work was supported by the Foundation of NNSF (Grant Nos. 10474041, 90403120, 10021001), and the Nonlinear Project of the NSM (973).

[1] F. Gabel, D. Bicout, U. Lehnert, M. Tehei, M. Weik, and G. Zaccai, Q. Rev. Biophys. **35**, 327 (2002).
[2] M. Tehei and G. Zaccai, Biochim. Biophys. Acta **1724**, 404 (2005).
[3] J. Norberg and L. Nilsson, Q. Rev. Biophys. **36**, 257 (2003).
[4] C. D. Snow, E. J. Sorin, M. R. Young, and V. S. Pande, Annu. Rev. Biophys. Biomol. Struct. **34**, 43 (2005).
[5] L. Mirny and E. Shakhnovich, Annu. Rev. Biophys. Biomol. Struct. **30**, 361 (2001).
[6] P. J. Thomas, B. H. Qu, and P. L. Pedersen, Trends Biochem. Sci. **20**, 456 (1995).
[7] S. E. Radford and C. M. Dobson, Cell **97**, 291 (1999).
[8] C. M. Dobson, Philos. Trans. R. Soc. London, Ser. B **356**, 133 (2001).
[9] P. T. Lansbury, Jr., Proc. Natl. Acad. Sci. U.S.A. **96**, 3342 (1999).
[10] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995).
[11] J. N. Onuchic, Z. L. Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[12] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[13] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. I. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[14] N. D. Socci, J. N. Onuchic, and P. G. Wolynes, Proteins: Struct., Funct., Genet. **32**, 136 (1998).
[15] H. S. Chan and K. A. Dill, Proteins: Struct., Funct., Genet. **30**, 2 (1998).
[16] S. S. Plotkin and J. N. Onuchic, Q. Rev. Biophys. **35**, 111 (2002).
[17] S. S. Plotkin and J. N. Onuchic, Q. Rev. Biophys. **35**, 205 (2002).
[18] J. Chen, J. Wang, and W. Wang, Proteins: Struct., Funct., Genet. **57**, 153 (2004).
[19] M. Karplus, Biopolymers **68**, 350 (2003).
[20] S. E. Radford, Trends Biochem. Sci. **25**, 611 (2000).
[21] V. Daggett and A. Fersht, Nat. Rev. Mol. Cell Biol. **4**, 497 (2003).
[22] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[23] Y. Duan, L. Wang, and P. A. Kollman, Proc. Natl. Acad. Sci. U.S.A. **95**, 9897 (1998).
[24] V. S. Pande, I. Baker, J. Chapman, S. P. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic, Biopolymers **68**, 91 (2003).
[25] U. Mayor, N. R. Guydosh, C. M. Johnson, J. G. Grossmann, S. Sate, G. S. Jas, S. M. V. Freund, D. O. V. Alonso, V. Daggett, and A. R. Fersht, Nature (London) **421**, 863 (2003).
[26] *Computational Biochemistry and Biophysics*, edited by O. M. Becker, A. D. MacKerell, Jr., B. Roux, and M. Watanabe (Marcel Dekker, New York, 2001).
[27] K. Fan, J. Wang, and W. Wang, Phys. Rev. E **64**, 041907 (2001).
[28] D. D. Jong, R. Riley, D. O. V. Alonso, and V. Daggett, J. Mol. Biol. **319**, 229 (2002).
[29] A. Li and V. Daggett, J. Mol. Biol. **275**, 677 (1998).
[30] L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, J. Mol. Biol. **254**, 260 (1995).
[31] S. E. Jackson and A. R. Fersht, Biochemistry **30**, 10428 (1991).
[32] S. E. Jackson and A. R. Fersht, Biochemistry **30**, 10436 (1991).
[33] S. E. Jackson, N. elMasry, and A. R. Fersht, Biochemistry **32**, 11270 (1993).
[34] A. R. Fersht, A. Matouschek, and L. Serrano, J. Mol. Biol. **224**, 771 (1992).
[35] L. Serrano, A. Matouschek, and A. R. Fersht, J. Mol. Biol. **224**, 805 (1992).
[36] A. Matouschek, L. Serrano, and A. R. Fersht, J. Mol. Biol. **224**, 819 (1992).
[37] A. R. Fersht, FEBS Lett. **325**, 5 (1993).
[38] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[39] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[40] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[41] D. A. C. Beck and V. Daggett, Methods **34**, 112 (2004).
[42] C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[43] A. Li and V. Daggett, J. Mol. Biol. **257**, 412 (1998).
[44] V. Daggett, A. Li, L. S. Itzhaki, D. E. Otzen, and A. R. Fersht,

J. Mol. Biol. **257**, 430 (1996).
[45] M. Levitt, J. Mol. Biol. **168**, 621 (1983).
[46] G. S. Shaw, R. S. Hodges, and B. D. Sykes, Science **249**, 280 (1990).
[47] G. S. Shaw, L. F. Golden, R. S. Hodges, and B. D. Sykes, J. Am. Chem. Soc. **113**, 5557 (1991).
[48] K. B. Wong, J. Clarke, C. J. Bond, J. L. Neira, S. M. V. Freund, A. R. Fersht, and V. Daggett, J. Mol. Biol. **296**, 1257 (2000).
[49] R. Koradi, M. Billeter, and K. Whrich, J. Mol. Graphics **14**, 51 (1996).
[50] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).