Identification of the protein folding transition state from molecular dynamics trajectories

S. Muff and A. Caflisch[a)]
Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

(Received 9 December 2008; accepted 23 February 2009; published online 27 March 2009)

The rate of protein folding is governed by the transition state so that a detailed characterization of its structure is essential for understanding the folding process. *In vitro* experiments have provided a coarse-grained description of the folding transition state ensemble (TSE) of small proteins. Atomistic details could be obtained by molecular dynamics (MD) simulations but it is not straightforward to extract the TSE directly from the MD trajectories, even for small peptides. Here, the structures in the TSE are isolated by the cut-based free-energy profile (cFEP) using the network whose nodes and links are configurations sampled by MD and direct transitions among them, respectively. The cFEP is a barrier-preserving projection that does not require arbitrarily chosen progress variables. First, a simple two-dimensional free-energy surface is used to illustrate the successful determination of the TSE by the cFEP approach and to explain the difficulty in defining boundary conditions of the Markov state model for an entropically stabilized free-energy minimum. The cFEP is then used to extract the TSE of a β-sheet peptide with a complex free-energy surface containing multiple basins and an entropic region. In contrast, Markov state models with boundary conditions defined by projected variables and conventional histogram-based free-energy profiles are not able to identify the TSE of the β-sheet peptide. © *2009 American Institute of Physics*.
[DOI: 10.1063/1.3099705]

I. INTRODUCTION

Proteins fold from the heterogeneous set of denatured conformations to the structurally well-defined native state by a complex conformational transition governed by the free-energy surface.[1] In remarkable contrast to the complexity of the folding process, a simple two-state description, i.e., folded and denatured free-energy minima separated by the transition state ensemble (TSE), is often used to describe the experimental measurements on single-domain fast-folding proteins.[2] As a consequence, little information concerning the details of the folding pathways is obtained, although experimental approaches based on mutagenesis have played a key role in providing a description of the residue interactions at the TSE.[3] Also, studies supplementing the kinetic measurements by probes sensitive to structural details[4,5] have shed some light into the folding pathways, particularly when intermediates are present.[6] However, none of the experimental studies can provide a detailed description of the structures that are visited along the folding pathways. In particular, it is difficult to determine the structures of the folding TSE because of their transient character and the many degrees of freedom of the polypeptide chain. In a nutshell, the TSE is elusive and complex.

Several approaches have been proposed to identify putative TSE structures along molecular dynamics (MD) trajectories by calculating the probability of folding before unfolding (p_{fold}) and selecting the ensemble with $p_{\mathrm{fold}} \approx 0.5$ (Table I and Refs. 7–11). It is important to note that the definition of p_{fold} is the origin of many difficulties when it comes to practical applications because in contrast to the folded state, which is well defined by structural criteria, it is all but simple to define the usually very heterogeneous denatured state. An efficient but approximate approach to calculate p_{fold} for any snapshot directly from the original MD trajectory (termed p_{fold}^{N} hereafter) follows the trajectory segment of length τ_{commit} of a given snapshot, wherein it is checked if the folding condition is met.[11] Upon coarse-graining into mesostates (nodes are used as synonymous), the fraction of snapshots of a mesostate that fold within τ_{commit} corresponds to the p_{fold}^{N} of that node. Therefore, the statistics of p_{fold}^{N} are bounded by the number of visits along the original MD trajectory to the respective node. A more accurate way to determine the same quantity is by analytical calculation on the equilibrium transition network[12,13] (ETN) with a procedure termed pfoldt.[13] The ETN is the capacitated graph whose nodes and links represent coarse-grained mesostates and transitions, respectively, sampled by MD simulations. The evaluation of pfoldt is based on the complete information about p_{fold} for a given commitment time τ_{commit} as contained in the ETN, i.e., $p_{\mathrm{fold}}(\tau_{\mathrm{commit}})$ is the solution of an equation system operating on the ETN.[13]

Other p_{fold}-based procedures require additional short MD simulations[14] (termed $p_{\mathrm{fold}}^{\mathrm{MD}}$ in the following) and have been used for *validating* putative TSE structures.[9,15–17] In analogy to p_{fold}^{N}, $p_{\mathrm{fold}}^{\mathrm{MD}}$ corresponds to the fraction of MD trajectories that fold within τ_{commit}. In contrast to the direct evaluation on the original MD trajectory for the former, the large number of additional MD runs required to calculate the $p_{\mathrm{fold}}^{\mathrm{MD}}$ value is computationally expensive. Therefore, $p_{\mathrm{fold}}^{\mathrm{MD}}$ calculations are

[a)]Author to whom correspondence should be addressed. Tel.: +41 44 635 55 21. FAX: +41 44 635 68 62. Electronic mail: caflisch@bioc.uzh.ch.

TABLE I. Procedures used to validate or identify the folding TSE. Abbreviations: cFEP, cut-based free-energy profile; ETN, equilibrium transition network; MSM, Markov state model; and τ_{commit}, commitment time.

Procedure	Data used	Advantages	Disadvantages[a]	Ref.
cFEP	ETN	Fast Exact on ETN	Requires coarse-graining	12 and 13
p_{fold}^{MD}	Additional MD runs	Exact, used for validation No coarse-graining	Computationally expensive Requires τ_{commit}	14
p_{fold}^{N}	Original trajectory	Fast	Requires τ_{commit} Strong dependency on sampling Requires coarse-graining	11
pfoldt	ETN	Fast Exact on ETN	Requires τ_{commit} Requires coarse-graining	13
p_{fold}^{MSM}	ETN	Fast	Most nodes with $p_{fold}^{MSM}=0.5$ do not belong to TSE Requires unfolded state definition Requires coarse-graining	10 and 25

[a]The cFEP, p_{fold}^{N}, pfoldt, and p_{fold}^{MSM} methods rely on sufficient sampling and a meaningful coarse-graining of the trajectories. Note that the p_{fold}^{N} procedure has a stronger dependency on sampling than cFEP, pfoldt, and p_{fold}^{MSM}. The latter procedures use the ETN, which is much more informative than the original trajectory itself because the ETN represents the complete connectivity information of all states.

useful mainly as validation tool, while the p_{fold}^{N} procedure is helpful for an initial guess of the TSE.

Note that p_{fold}^{N}, pfoldt, and p_{fold}^{MD} all rely on the concept of a τ_{commit}-based definition of p_{fold}, which is comfortable because a detailed definition of the unfolded state is circumvented. This advantage is especially useful if the unfolded state is heterogeneous and/or entropic. By applying $p_{fold}(\tau_{commit})$, unfolding becomes a "negative concept," i.e., the nonobservation of a folding event within τ_{commit} is set equivalent to an unfolding event. However, $p_{fold}(\tau_{commit})$ is a good approximation to the original p_{fold} definition (i.e., folding before unfolding) only if there is a clear separation of timescales between intra- and interbasin transitions, which ensures that the choice of τ_{commit} is robust. Therefore, choosing τ_{commit} can be difficult and rather arbitrary, especially if the system is not known in detail.

Other approaches to calculate p_{fold} values for the TSE determination are embedded within the framework of a Markov state model (p_{fold}^{MSM}). Such calculations require the definition of initial (unfolded, $p_{fold}=0$) and final (folded, $p_{fold}=1$) regions. However, as shown in this paper, even for relatively simple systems these boundary states are not determined adequately by the selection of unfolded state representative(s) or by means of geometric variables like the number of native contacts and the root mean square deviation (rmsd) from the folded structure. As a consequence, p_{fold}^{MSM} fails to isolate the TSE.

In contrast, we show here that the folding TSE structures can be identified accurately by the cut-based free-energy profile[12] (cFEP) obtained from MD simulations. The cFEP is a barrier-preserving projection onto a progress coordinate that takes into account all routes to leave or enter the free-energy basin chosen as reference and uses as only input the ETN,[12,13] i.e., no additional parameters such as τ_{commit} or the knowledge of the unfolded state ensemble are needed. In particular, the unfolding barrier, which is the barrier to leave the folded state, can be determined exactly by the cFEP,[12,13] and the ensemble on top of the barrier corresponds to the TSE. The procedure to isolate the TSE by the cFEP is validated here on a simple and illustrative two-state free-energy surface as well as on a complex system with 645 degrees of freedom. The latter is a structured peptide (20-residue three-stranded antiparallel β-sheet called Beta3s) for which several folding-unfolding events can be sampled by implicit solvent MD simulations.[18,19] Finally, we show that conventional, i.e., histogram-based, free-energy projections onto apparently appropriate geometrical variables are not useful for determining the TSE of Beta3s.

II. METHODS

Table I gives a short description of the procedures that were employed to isolate and validate the TSE as well as advantages and disadvantages of the approaches (cFEP, p_{fold}^{MD}, p_{fold}^{N}, pfoldt, and p_{fold}^{MSM}) used to determine or validate putative TSE structures. Other approaches to bias simulations toward structures in the TSE (Refs. 20 and 21) or to isolate them from unfolding simulations[9] require experimental data (ϕ-values[3]) and are therefore not directly comparable with the procedures used in this paper.

A. Transition state identification from the cFEP

Projected free-energy surfaces are most useful if they preserve the barriers and minima in the order that they are met during folding/unfolding events. Using an analogy between the system kinetics and equilibrium flow through a network, Krivov and Karplus[12] introduced the cFEP method and a progress coordinate that have most of these properties.

The input for the cFEP calculation is the ETN, which is derived from the transitions between coarse-grained mesostates, sampled during a MD simulation at equilibrium.

For a node i in the ETN the partition function is $Z_i=\Sigma_j c_{ij}$, where $c_{ji}=(n_{ji}+n_{ij})/2$, and n_{ji} is the absolute number of transitions from node i to node j observed during the trajectory, i.e., c_{ji} are the entries of the symmetrized transition matrix that fulfills the detailed balance criterion. The transition probabilities can then be calculated as $p_{ij}=c_{ij}/\Sigma_k c_{kj}$. If the nodes of the network are partitioned into two sets $\mathcal{A}$ and $\mathcal{B}$, where set $\mathcal{A}$ contains the reference node A, then $Z_A=\Sigma_{i\in\mathcal{A}}Z_i$, $Z_B=\Sigma_{i\in\mathcal{B}}Z_i$, $Z_{AB}=\Sigma_{i\in\mathcal{A},j\in\mathcal{B}}c_{ij}$, and the free energy of the barrier between the two sets of nodes (or regions) is $\Delta G=-kT\log(Z_{AB}/Z)$, with Z being the partition function of the full network (Fig. 1). The progress coordinate then is the normalized partition function Z_A/Z of the reactant region containing the native node A, but other progress coordinates can be used because the cFEP is invariant with respect to arbitrary transformations of the reaction coordinate.[22] A pair $(Z_A/Z,\Delta G)$ can be calculated for every division into two sets of nodes and the result is a one-dimensional profile that preserves the barriers between the free-energy basins; given the barriers, the minima can be determined.[12] The method was applied to the β-hairpin of protein G (Ref. 12) and Beta3s.[13]

During the procedure of cFEP calculation, all nodes of the system are assigned a value of the progress variable. Here, the mean first passage time (mfpt) to the folded node (A) was used as progress variable. The latter can be calculated analytically for every node i if the ETN fulfils the Markov property by solving the equation system

$$\mathrm{mfpt}_i=\Delta t+\sum_j p_{ji}\cdot\mathrm{mfpt}_j,$$
$$\mathrm{mfpt}_A=0, \tag{1}$$

where Δt is the sampling interval (which is 20 ps in this work). The occurrence of the nodes in the profile is sorted in ascending order of mfpt from the native state (i.e., from the first node along Z_A/Z) because the data points of the profile are calculated for all possible divisions of nodes into sets with $\mathrm{mfpt}\le\mathrm{mfpt}_c$ and $\mathrm{mfpt}>\mathrm{mfpt}_c$ for any value of mfpt_c (Fig. 1). Therefore, every node i can be localized along the cFEP according to its progress variable mfpt_i because the Z_A/Z coordinate that corresponds to node i can be calculated as $Z_A/Z(i)=\Sigma_{\mathrm{mfpt}(j)\le\mathrm{mfpt}(i)}Z_j/Z$, where Z_j is the partition function of node j and Z is the total partition function. Since the cFEP takes into account all routes present in the ETN to and from the initial state[12,13] without any prejudgment as to the geometric coordinates or pathways involved, the TSE is situated on top of the unfolding barrier. In this way, nodes corresponding to the top of the first barrier to exit the native state can automatically be identified as TSE structures. Importantly, the calculation of cFEPs requires as only input the choice of a node as representative of the folded state. All cFEPs in this work were calculated using the program WORDOM.[23]

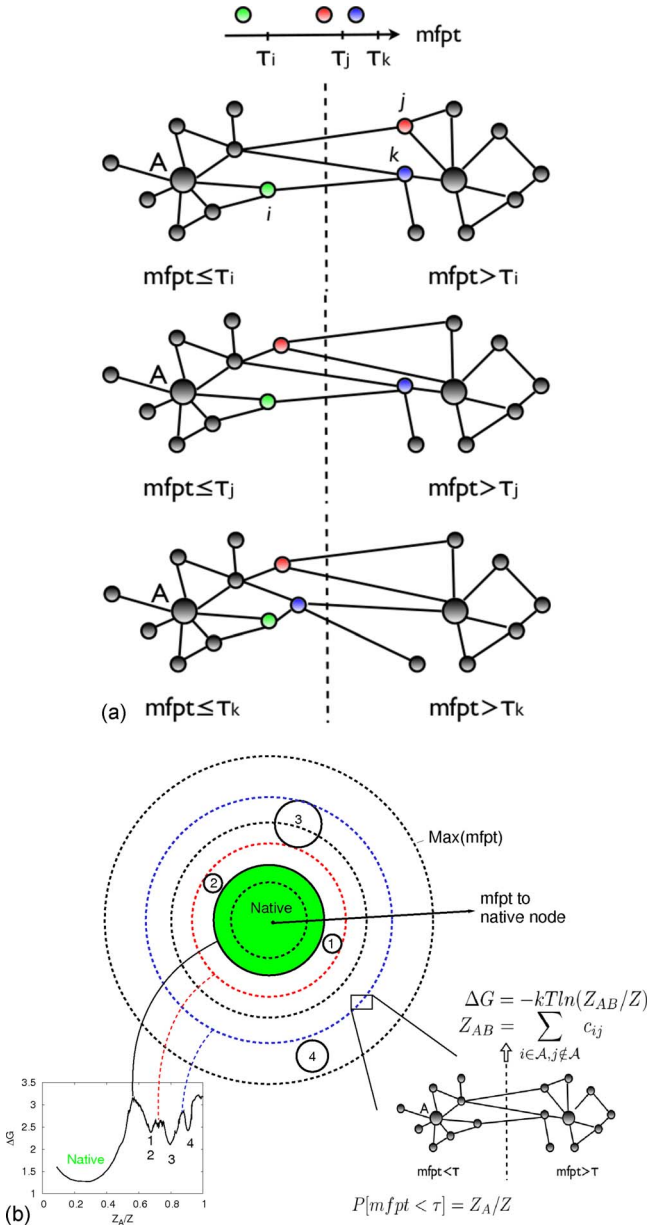

FIG. 1. (Color online) Schematic illustration of the cFEP procedure (Refs. 12 and 13). (Top) Nodes of the ETN are first sorted according to increasing value of mfpt. For each mfpt_c value in the range from 0 (node A) to max(mfpt), a value of the cut Z_{AB} is calculated. The set $\mathcal{A}$ of nodes on the left of the cut contains node A and all nodes with $\mathrm{mfpt}\le\mathrm{mfpt}_c$, and Z_A/Z is its relative partition function. The green, red, and blue nodes have increasing values of mfpt in this simplified illustration of the ETN. (Bottom) Relationship between free-energy basins and the cFEP. Solid circles represent basins, while concentric dashed circles represent values of mfpt. To plot the cFEP, $\Delta G=-kT\log(Z_{AB}/Z)$ is calculated as a function of Z_A/Z. Basins 1 and 2 overlap on the one-dimensional cFEP because they have the same mfpt distance from the native state and are therefore not separated; they are both located in the first minimum of the profile after the first, i.e., unfolding, barrier.

B. Evaluation of $p_{\mathrm{fold}}(\tau_{\mathrm{commit}})$ with additional simulations ($p_{\mathrm{fold}}^{\mathrm{MD}}$)

If snapshots saved along a trajectory are grouped into structurally homogeneous nodes during the coarse-graining procedure, nodes belonging to the TSE have equal probability to fold and to unfold, i.e., $p_{\mathrm{fold}}\approx 0.5$, whereas folded and unfolded regions correspond to $p_{\mathrm{fold}}\approx 1$ and $p_{\mathrm{fold}}\approx 0$, respectively. In order to validate the application of cFEPs to iden-

tify the transition, folded, and unfolded ensembles, a large number of MD trajectories from various structures with varying initial distribution of velocities can be started and the fraction of those that fold within a commitment time τ_{commit} (Refs. 14, 15, and 24) corresponds to the respective $p_{\text{fold}}^{\text{MD}}$. τ_{commit} has to be chosen much longer than the shortest time scales of conformational fluctuations and much shorter than the average folding time.[15]

The $p_{\text{fold}}^{\text{MD}}$ calculations are computationally very expensive since the error for a structure scales with $1/\sqrt{n}$, where n is the number of trajectories started from it. Therefore, the realization of many short trajectories from individual structures cannot be applied directly to identify the TSE from MD trajectories and is used in this work only to validate putative transition state structures as isolated by other approaches.

C. $p_{\text{fold}}(\tau_{\text{commit}})$ calculation directly from the trajectories (p_{fold}^{N}) or the ETN (pfoldt)

The calculation of $p_{\text{fold}}^{\text{MD}}$ is computationally very expensive and is feasible only for a small subset of nodes. In a previous work, a method was proposed for estimating folding probabilities for *all* structures visited in an equilibrium folding-unfolding trajectory.[11] The calculation does not require any additional simulations because the original MD trajectory is used to directly estimate the folding probabilities. The τ_{commit}-segment of the MD trajectory following each snapshot is analyzed to check if the folding condition is met (i.e., that the folded node, which usually is the most populated one, is visited). For each node, the ratio between the snapshots which lead to folding and the total number of snapshots in the node is defined as the node-$p_{\text{fold}}(p_{\text{fold}}^{N})$. This value is an approximation of the $p_{\text{fold}}(\tau_{\text{commit}})$ of any single structure in the node which is valid if the node consists of structurally and kinetically similar conformations. The error in p_{fold}^{N} scales with $1/\sqrt{W}$, where W is the number of structures in the node.

The analytical calculation of $p_{\text{fold}}(\tau_{\text{commit}})$ on the ETN of a Markov state model, termed pfoldt, is more accurate than p_{fold}^{N} because it uses the full connectivity of the ETN, thereby reducing the statistical error. The corresponding equation system was introduced previously[13] and the results were used as a progress variable for cFEP calculations (pfoldt procedure).

D. p_{fold} calculation with a Markov state model ($p_{\text{fold}}^{\text{MSM}}$)

Within the framework where the ETN corresponds to a Markov state model, folding probabilities can be calculated directly if the two regions $\mathcal{U}$ (unfolded) and $\mathcal{F}$ (folded) are known. Given these regions, the folding probability of a node i within the Markov state model is found as the solution of the equation system $p_i^{\text{MSM}}=\sum_j p_{ji} p_j^{\text{MSM}}$ with boundary conditions $p_{\kappa\in\mathcal{U}}^{\text{MSM}}=0$ and $p_{\kappa\in\mathcal{F}}^{\text{MSM}}=1$.[10,25] The equation system can be solved efficiently by iterative multiplication of the vector p_j^{MSM} by the matrix p_{ji}. According to the cFEP, the folded and unfolded states are defined as all nodes on the left and right of the folding barrier, respectively. However, use of this definition to determine $\mathcal{U}$ and $\mathcal{F}$ would be tautological because if the cFEP is known, there is no need for the Markov state model approach, and it is then trivial that the nodes on the barrier have no other choice than attaining $p_{\text{fold}}\approx 0.5$. Therefore, the Markov state model approach is applied in this work without the input of the knowledge from the cFEP in order to objectively compare $p_{\text{fold}}^{\text{MSM}}$ with the other methods.

III. TWO-STATE SYSTEM WITH ENTROPIC FREE-ENERGY MINIMUM: AN ILLUSTRATIVE MODEL

A simple two-dimensional, radially symmetric potential energy surface[26] is used here to illustrate the correct TSE isolation by the cFEP and the dependency of the $p_{\text{fold}}^{\text{MSM}}$ on the definition of initial and final regions (Fig. 2). The function of the potential energy surface is

$$z=f(x,y)=\begin{cases}\sqrt{x^2+y^2} & \text{if } \sqrt{x^2+y^2}\le 1\\ 1 & \text{otherwise,}\end{cases}$$

but the radial symmetry allows the unambiguous expression depending on a single variable $r=\sqrt{x^2+y^2}$ as

$$U(r)=\begin{cases}r & \text{if } r\le 1\\ 1 & \text{otherwise,}\end{cases}$$

where r is the radial coordinate.[26] Using a temperature of 125 K, the partition function of coordinate r is $Z(r)=r\exp(-U(r)/0.25)$ (with angle 1) and

$$F(r)=\begin{cases}r-0.25\ln(r) & \text{if } r\le 1\\ 1-0.25\ln(r) & \text{otherwise.}\end{cases}$$

The corresponding free-energy surface has only two minima: an enthalpic, funnel-like "folded" state and a purely entropic "denatured" state [Fig. 2(a)]. The discretization of a portion of the two-dimensional free-energy surface is shown in Fig. 2(b). The partition function of a node i with radial coordinate r_i can then be calculated as $z_i=\exp(-U(r_i)/0.25)$, and a Monte Carlo simulation yields the ETN.[26] There is a free-energy barrier at $r=0$ that separates the enthalpic ($r<1$) from the entropic basin ($r>1$), as indicated by the green line in Figs. 2(a) and 2(b).

In a first step, it was verified that the cFEP is able to identify the TSE. The TSE is correctly grouped around the barrier in the cFEP, and only the nine nodes in the neighborhood of the minimal cut (at $r\lessapprox 1$) lie between the first and the last green circle in the cFEP [Fig. 2(c)]. In a next step, $p_{\text{fold}}^{\text{MSM}}$ was calculated between the most populated node (black, $p_{\text{fold}}^{\text{MSM}}=1$) and an arbitrary representative of the entropic state (red, $p_{\text{fold}}^{\text{MSM}}=0$). The strong dependency of the $p_{\text{fold}}^{\text{MSM}}\approx 0.5$ region on the choice of the unfolded representative is remarkable [Figs. 2(d)–2(f)]. Furthermore, none of the choices is able to fully reveal the correct TSE, although the system is very simple. This result illustrates the difficulty of selecting a representative structure, which is in practice almost impossible in the case of an entropically stabilized state.

IV. APPLICATION TO BETA3S

In the previous example the complete knowledge about the free-energy surface is available and it is very simple to

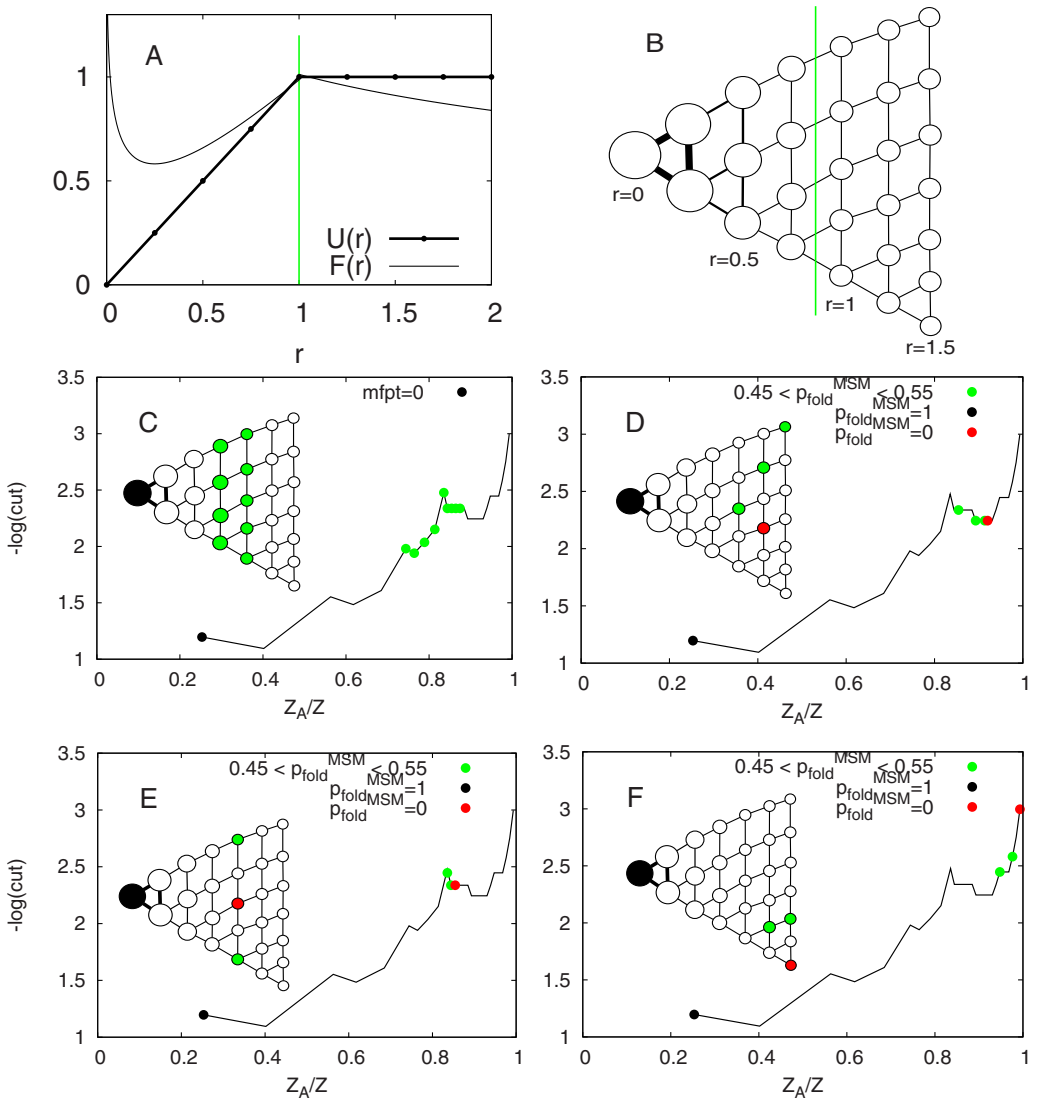

FIG. 2. (Color online) Simple two-state system with entropically stabilized "unfolded" state (Ref. 26). (a) Two-dimensional radially symmetric potential energy surface $U(r)$ and free-energy surface $F(r)$ at 125 K. There are two free-energy minima: one relatively deep representing the enthalpic ("native") state ($r<1$) and the other shallow representing the entropic unfolded) basin ($r>1$). The green line indicates the transition state. (b) Discretized portion of the corresponding free-energy landscape. The size of the nodes and links in the model network is proportional to the partition function of the nodes and transitions. The green line represents the minimal cut through the free-energy barrier, that is, the transition state. (c) The TSE is correctly identified by the cFEP approach, i.e., it consists of the nine nodes on the left and right of the minimal cut in the ETN. [(d)–(f)] The solution of the $p_{\mathrm{fold}}^{\mathrm{MSM}}$ calculations and identification of $p_{\mathrm{fold}}^{\mathrm{MSM}} \approx 0.5$ regions is strongly dependent on the node chosen as representative of the entropic region. By none of the three choices it is possible to isolate the complete TSE region correctly. This illustrative model shows that the arbitrary selection of representative nodes in the entropic state is not valid in general and that cFEPs are not affected by this problem because they require only the definition of the native node.

correctly determine the folded and unfolded regions and therefore also the TSE. However, this is an oversimplified and unrealistic case, and the following application to the structured peptide Beta3s illustrates the advantage of the cFEP approach for the analysis of complex systems.

A. MD simulations

Beta3s is a designed 20-residue sequence whose solution conformation has been investigated by NMR spectroscopy.[27] The NMR data indicate that Beta3s in water forms a monomeric (up to more than 1 mM concentration) triple-stranded antiparallel β-sheet, in equilibrium with the denatured state.[27] We have previously shown that in implicit solvent[28] MD simulations Beta3s folds reversibly to the NMR solution conformation, irrespective of the starting structure.[18] Recently, analysis of a 20 μs equilibrium MD simulation close to the melting temperature at 330 K revealed a very heterogeneous denatured state with a large entropic region and multiple enthalpic traps.[13,19,29] The same 20 μs of MD sampling was used here, and there are a total of 10^6 snapshots because coordinates were saved every 20 ps. The simulations were performed with the program CHARMM.[30,31] Beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field[30]). A mean field approximation based on the solvent accessible surface was used to describe the main effects of the aqueous solvent on the solute.[28] The two surface tensionlike parameters of the solvation model were optimized without using Beta3s. The same force field and implicit solvent model have been used in MD simulations of

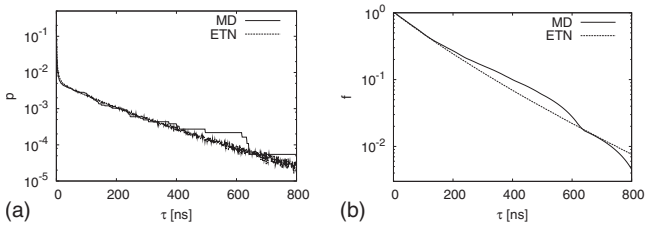

FIG. 3. Folding time distribution $p(\tau)$ (left) and cumulative distribution $f(t)=\int_t^\infty p(\tau)d\tau$ (right) as extracted directly from the 20 μs equilibrium simulation of Beta3s (MD) and from solving the system of Eq. (1) on the ETN, which is treated as a Markov state model. The folding dynamics from the non-native ensemble can be reproduced by the model, which is a strong indication that the Markov assumption is justified for the lagtime of 20 ps used here. The discrepancies between the two cumulative folding time distributions for long time scales are due to the rare sampling of slow folding events in the MD trajectory [see stepwise drop of $p(\tau)$], whereas the calculation of folding times from the ETN is less affected by statistical errors.

the early steps of ordered aggregation,[32] and folding of structured peptides,[18,28,33] as well as small proteins of about 60 residues.[34]

B. Coarse-graining

The leader algorithm[35] is used for coarse-graining the snapshots according to the all-atom rmsd and a cutoff value of 2.5 Å.[13] The current snapshot is grouped to the last (in time) visited node whose central snapshot has a rmsd from the current snapshot lower than the cutoff. This version of the leader algorithm accounts not only for structural but also for kinetic similarity because recently visited snapshots are more likely to be kinetically close than those that were visited with a large temporal delay. The importance of using transitions rather than only structures to assign states has been recently investigated for a small helical peptide.[36]

Note that nodes in the ETN with only one or two neighbors (i.e., one incoming and/or one outgoing neighbor) were grouped to their outgoing neighbor. This regrouping is justified because the future of such nodes within a trajectory is completely determined, i.e., no information is erased through their regrouping. Upon rmsd coarse-graining and regrouping 34 671 nodes and 151 819 links were visited. These nodes are the states of the Markov state model (i.e., the ETN), and the lagtime was set to $\Delta t=20$ ps. Figure 3 contains a comparison of folding dynamics from the MD simulations and from the corresponding Markov state model at 330 K. Essentially the same kinetics can be extracted, indicating that the Markov assumption holds and non-Markovian noise is negligible.

C. Correct identification of the TSE by the cFEP method

The native basin is bounded by the first local maximum in the unfolding cFEP of Beta3s, which is the cFEP with the native node as reference (Fig. 4). To show that the TSE is situated on top of the unfolding barrier, the folding probability $p_{\mathrm{fold}}^{\mathrm{MD}}$ with a τ_{commit} value of 5 ns was evaluated on 34 nodes by running additional MD simulations (see Sec. II). These nodes were selected equally spaced along the Z_A/Z, except for a higher density in a region bracketing the unfold-

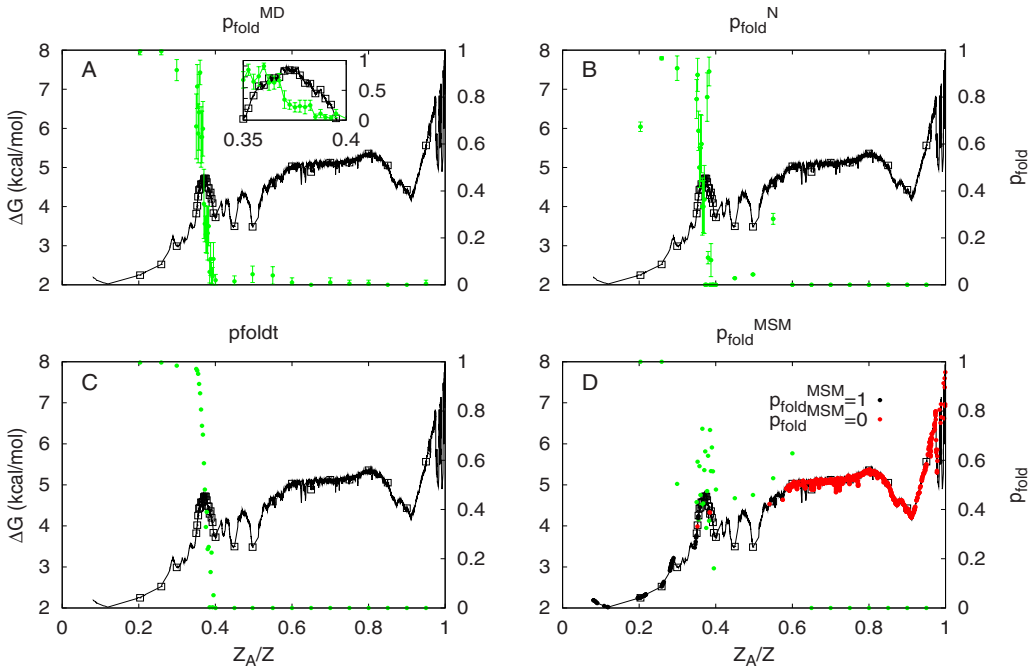

FIG. 4. (Color online) The TSE can be identified by the cFEP (solid line). The cFEP is shown together with the 34 nodes that were selected for $p_{\mathrm{fold}}^{\mathrm{MD}}$ calculations (black squares). They are equally spaced along the progress coordinate Z_A/Z, with a distance of 0.05 units except for the first barrier, i.e., $0.35<Z_A/Z<0.4$, where the spacing is 0.0025 units to obtain a higher resolution. Values of p_{fold} (green circles) refer to the y-axis on the right. (a) $p_{\mathrm{fold}}^{\mathrm{MD}}$. The error bars represent the standard deviation among the ten structures within a node. The inset illustrates the sharp decay at the unfolding barrier. (b) $p_{\mathrm{fold}}^{\mathrm{N}}$. The error bars represent the standard deviation if the calculations are considered as a Bernoulli experiment. (c) p_{fold} as calculated analytically on the ETN (pfoldt). (d) $p_{\mathrm{fold}}^{\mathrm{MSM}}$. The use of the number of native contacts Q for the definition of the folded ($Q>19/26$, black circles) and unfolded ($Q<5/26$, red circles) states as boundary conditions of the MSM results in incorrect values of $p_{\mathrm{fold}}^{\mathrm{MSM}}$ and no sharp transition can be observed at the barrier. A similar failure is observed when defining folded and unfolded ensembles by rmsd from the native structure (Supplementary Material, Fig. S8).[37]

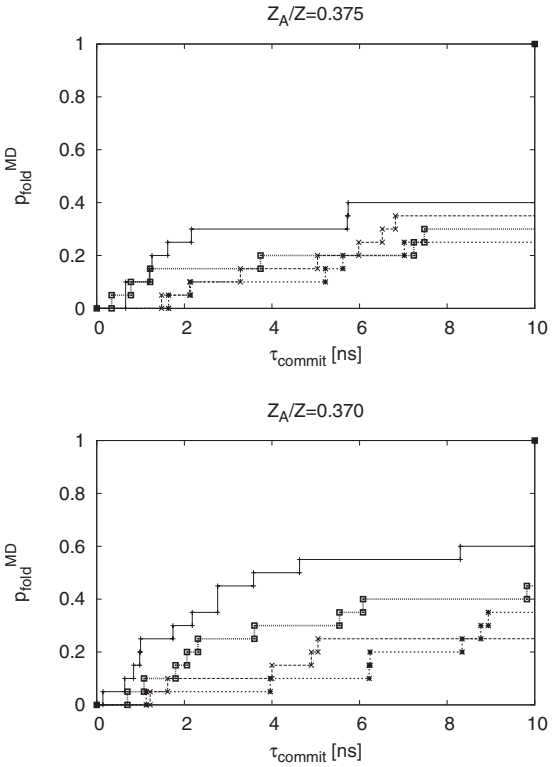

FIG. 5. Dependence of $p_{\mathrm{fold}}^{\mathrm{MD}}$ on the value of τ_{commit} (Ref. 17). Results are shown for 20 short runs from each of four structures of a node with $Z_A/Z=0.375$ (top) and $Z_A/Z=0.370$ (bottom). The curves are step functions and reach a plateau at about 5–10 ns.

ing barrier on the cFEP (results with mfpt as progress coordinate are shown in Supplementary Material, Fig. S1).[37] Ten structures were chosen randomly from every node, and 20 simulations of 10 ns each with different initial velocities were started from each structure, i.e., a total of 200 simulations was accumulated per node. Note that for this validation it is more informative to monitor, for individual snapshots, the monotonously growing behavior of $p_{\mathrm{fold}}^{\mathrm{MD}}$ as a function of τ_{commit} rather than selecting a single value of τ_{commit} (Ref. 17) (Fig. 5). Notably, the top of the first free-energy barrier in the cFEP corresponds to the $p_{\mathrm{fold}}^{\mathrm{MD}} \approx 0.5$ region, i.e., to the folding/unfolding TSE of Beta3s [Fig. 4(a)]. The accuracy of the TSE identification on the cFEP is striking. Nodes before the first barrier belong to the native basin and have $p_{\mathrm{fold}}^{\mathrm{MD}} \approx 1$, while nodes after the barrier have $p_{\mathrm{fold}}^{\mathrm{MD}} \approx 0$. Moreover, the distributions of $p_{\mathrm{fold}}^{\mathrm{MD}}$ values over the ten structures in each node are peaked around the respective average $p_{\mathrm{fold}}^{\mathrm{MD}}$ value, even for TSE nodes (Supplementary Material, Fig. S2).[37] These results show that the cFEP approach is able to correctly identify not only free-energy basins and barriers of complex systems but also the TSE to exit or enter the region of interest (usually the folded state). The very good correlation between the increasing values of Z_A/Z along the cFEP and $p_{\mathrm{fold}}^{\mathrm{MD}}$ can be explained because both mfpt and $p_{\mathrm{fold}}^{\mathrm{MD}}$ describe the kinetic distance from a state. Note that this correlation is not due to a tautology because the cFEP is calculated using the ETN as input, whereas $p_{\mathrm{fold}}^{\mathrm{MD}}$ is extracted from additional MD simulations. Moreover, the correlation is robust with respect to the choice of the progress variable of the cFEP procedure (see Supplementary Material, Fig. S3).[37]

Importantly, no additional parameter is needed to calculate the cFEP, which only requires the selection of the native node (or the representative node of any other free-energy basin). Note that the perfect match between the unfolding barrier on the cFEP and the sharp decay of the $p_{\mathrm{fold}}^{\mathrm{MD}}$ values justifies *a posteriori* the choice of a commitment time of 5 ns used to calculate the latter. Essentially identical results are obtained with a commitment time of 10 ns (Table II and Supplementary Material, Fig. S4).[37]

D. Approximation by p_{fold}^{N} and pfoldt

As mentioned in Sec. II, the calculation of $p_{\mathrm{fold}}^{\mathrm{MD}}$ is computationally very expensive and is feasible only for a small subset of nodes. On the other hand, the calculation of p_{fold}^{N} does not require any additional simulations.[11] The p_{fold}^{N} values, which were also calculated with $\tau_{\mathrm{commit}}=5$ ns, are close to the $p_{\mathrm{fold}}^{\mathrm{MD}}$ for most of the 34 nodes used to calculate the latter [Fig. 4(b) and Table II]. The error is due to low statistics harvested for the p_{fold}^{N} estimation, which is limited by the number of visits to the node within the trajectory. This problem is particularly severe for TSE nodes, which suffer from low sampling, whereas highly populated conformations can be classified more reliably (Fig. 6, left).

The results improve dramatically if p_{fold}^{N} is replaced by its equivalent calculated on the ETN, i.e., pfoldt [Fig. 4(c)], and a very sharp decay of folding probabilities at the cFEP unfolding barrier can be observed. pfoldt is significantly more accurate than p_{fold}^{N} because as mentioned above the ETN contains all connectivity and pathway information between regions, which is not present if only isolated sampling events from individual nodes during the simulation are considered (as for p_{fold}^{N}). Results for all nodes populated by a significant number of snapshots above a certain cutoff are given in Fig. 6, middle. Again, the comparison of p_{fold}^{N} (left panel) with pfoldt (middle panel) confirms the higher accuracy of the latter. Note that pfoldt, like p_{fold}^{N} and $p_{\mathrm{fold}}^{\mathrm{MD}}$, relies on the correct choice of τ_{commit}, a parameter that is usually not simple to determine.

E. Failure of TSE identification by a Markov state model with boundaries defined according to structural criteria

The p_{fold} calculation in the framework of a Markov state model ($p_{\mathrm{fold}}^{\mathrm{MSM}}$) involves the choice of representative regions for the folded and unfolded states with $p_{\mathrm{fold}}=1$ and $p_{\mathrm{fold}}=0$, respectively, as boundary conditions.[10,25,38] Like in many previously published applications, regions $\mathcal{U}$ and $\mathcal{F}$ were determined according to a simple structural criterion based on the number of native contacts Q, which is a commonly used geometrical variable. A node was assigned to the initial and final regions if the structures in that node had on average less than 5 or more than 19 of the 26 native contacts[18] formed, respectively (Supplementay Material, Fig. S5).[37] With this definition of boundary conditions, $\mathcal{U}$ and $\mathcal{F}$ consist of 31% and 27% of the total number of snapshots, respectively.

Calculation of $p_{\mathrm{fold}}^{\mathrm{MSM}}$ values from the equation system reveals that several nodes after the unfolding barrier have a $p_{\mathrm{fold}}^{\mathrm{MSM}} > 0.5$ [Fig. 4(d)]. Figure 7 contains as a supplement to

TABLE II. p_{fold} values of nodes used for the calculation of p_{fold}^{MD}. In the region of the first (i.e., unfolding) barrier of the cFEP, $0.35 \leq Z_A/Z \leq 0.4$, the correlation between p_{fold}^{N} and p_{fold}^{MD} is 0.70, and the correlation between pfoldt and p_{fold}^{MD} is 0.95. Within the same range, there is no correlation between p_{fold}^{MSM} and p_{fold}^{MD} (correlation coefficient of 0.01). Similar correlation coefficients are obtained for $\tau_{commit}=5$ and 10 ns.

	p_{fold}^{MD}		p_{fold}^{N}		pfoldt			
Z_A/Z	5 ns	10 ns	5 ns	10 ns	5 ns	10 ns	p_{fold}^{MSM}	Node population
0.2	0.995	0.995	0.673	0.878	0.997	0.998	1.000	539
0.25	0.995	1.000	0.966	0.979	0.997	0.998	1.000	726
0.3	0.915	0.945	0.923	0.962	0.984	0.990	0.504	26
0.3500	0.675	0.815	0.792	0.917	0.971	0.981	0.430	24
0.3525	0.845	0.900	0.895	1.000	0.964	0.975	0.594	19
0.3550	0.645	0.750	0.656	0.656	0.951	0.967	0.359	32
0.3575	0.735	0.780	0.500	0.500	0.910	0.942	0.576	34
0.3600	0.905	0.925	0.583	0.833	0.872	0.916	0.459	12
0.3625	0.630	0.700	0.600	0.600	0.806	0.874	0.417	16
0.3650	0.630	0.705	0.333	0.733	0.740	0.834	0.728	15
0.3675	0.665	0.730	0.333	0.611	0.704	0.805	0.425	18
0.3700	0.345	0.540	0.364	0.364	0.588	0.752	0.560	11
0.3725	0.260	0.405	0.000	0.000	0.481	0.687	0.636	27
0.3750	0.215	0.310	0.000	0.000	0.329	0.572	0.325	15
0.3775	0.235	0.360	0.800	0.867	0.390	0.603	0.475	15
0.3800	0.220	0.350	0.116	0.116	0.239	0.524	0.618	147
0.3825	0.250	0.360	0.909	0.909	0.248	0.542	0.354	22
0.3850	0.055	0.145	0.000	0.688	0.003	0.570	0.723	16
0.3875	0.110	0.235	0.105	0.105	0.146	0.459	0.556	19
0.3900	0.050	0.140	0.000	0.000	0.225	0.459	0.652	177
0.3925	0.035	0.095	0.000	0.000	0.003	0.463	0.555	10
0.3950	0.110	0.155	0.000	0.000	0.052	0.361	0.161	10
0.4000	0.020	0.055	0.000	0.015	0.000	0.460	0.481	66
0.45	0.015	0.030	0.029	0.126	0.000	0.271	0.446	8 584
0.5	0.045	0.135	0.044	0.134	0.000	0.251	0.459	14 918
0.55	0.040	0.135	0.281	0.509	0.000	0.000	0.527	377
0.6	0.010	0.050	0.000	0.000	0.000	0.000	0.628	13
0.65	0.000	0.030	0.000	0.000	0.000	0.000	0.000	16
0.7	0.010	0.045	0.000	0.000	0.000	0.000	0.000	36
0.75	0.005	0.025	0.000	0.000	0.000	0.000	0.000	14
0.8	0.005	0.035	0.000	0.000	0.000	0.000	0.000	25
0.85	0.000	0.020	0.000	0.000	0.000	0.000	0.000	17
0.9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	10
0.95	0.005	0.010	0.000	0.000	0.000	0.000	0.000	14

the cFEP the location of all putative TSE nodes as isolated with p_{fold}^{MD}, p_{fold}^{N}, pfoldt, and p_{fold}^{MSM}, i.e., those nodes with $0.45 < p_{fold} < 0.55$ and at least 20 snapshots for statistical significance. While all three τ_{commit}-based methods approximate the TSE quite well [Figs. 7(a)–7(c)], most structures identified by p_{fold}^{MSM} are far away from the unfolding barrier in the cFEP [Fig. 7(d)]. The incorrect determination of the TSE by p_{fold}^{MSM} is also shown for regions $\mathcal{U}$ and $\mathcal{F}$ defined by all-atom rmsd >5.5 Å (weight of 48%) and all-atom rmsd <2.5 Å (weight of 25%), respectively (Fig. S6 of Supplementary Material).[37]

The isolation of the correct TSE by p_{fold}^{MSM} can only be achieved if the selected regions are "true" representatives of the folded and unfolded states, i.e., if each time the polypeptide folds or unfolds (and only then), the folded or unfolded region is visited, respectively. It is important to emphasize that, except for a two-state system with well-defined native and non-native basins, the choice of such representative ensembles is very difficult and mostly impossible by geometrical criteria. This problem originates from the usually very heterogeneous character of the denatured state with multiple basins and/or an entropic region.[13,19,26] While the representation of the folded state by a single node may be legitimate if the basin is enthalpic, the denatured state cannot be represented by a single node. For instance, for each choice of the unfolded representative disparate $p_{fold}^{MSM} \approx 0.5$ regions are obtained for Beta3s (Supplementary Material, Fig. S7).[37]

F. Failure of TSE identification from free-energy projections onto geometric variables

In a previous work, the numbers of native contacts in the N-terminal hairpin (Q_N) and C-terminal hairpin (Q_C) of Beta3s were used as progress variables to investigate thermodynamics and folding pathways sampled by MD simulations close to the melting temperature.[18] Note that these variables are the most "natural" among the geometric coordinates, considering that a three-stranded antiparallel

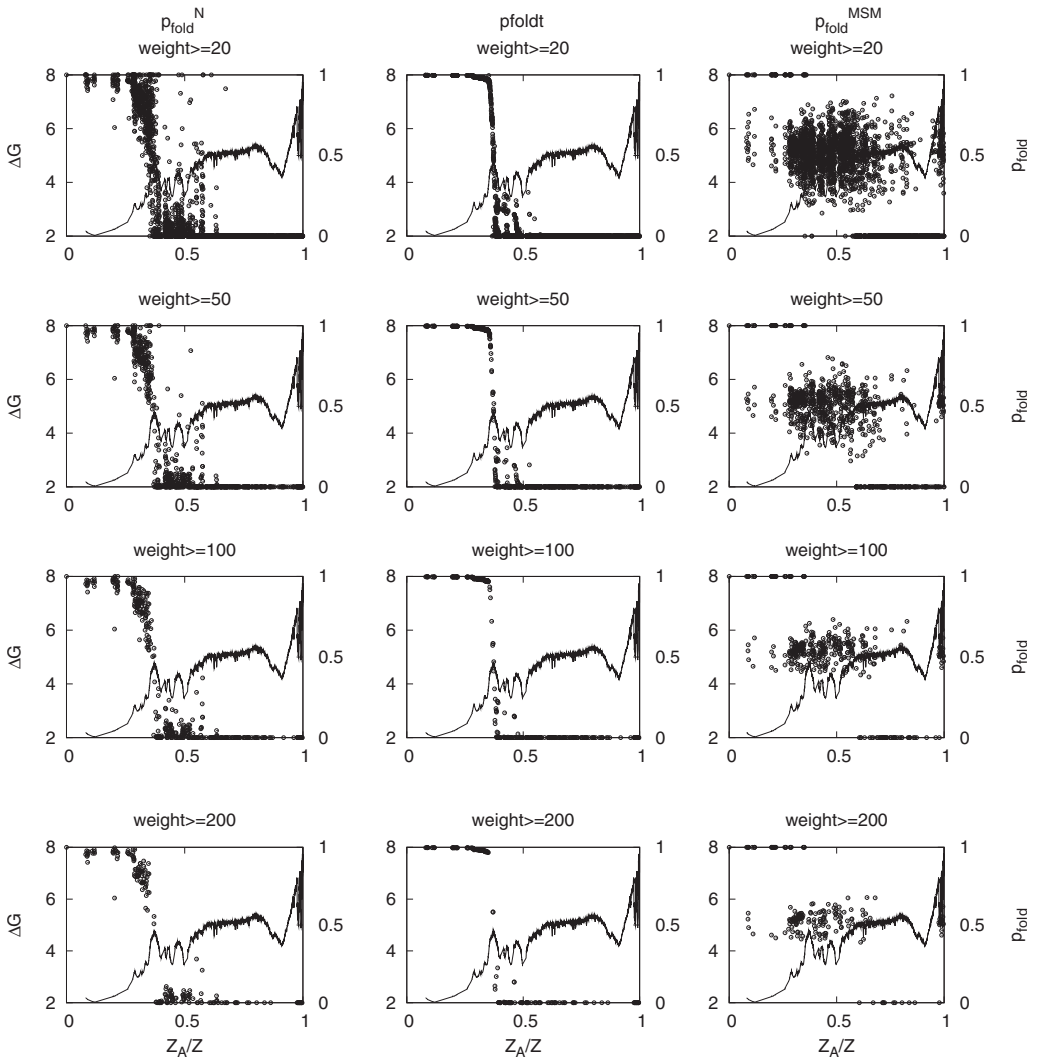

FIG. 6. The cFEP (solid line, left y-axis) is shown with the values of p_{fold} (circles, right y-axis) calculated by three different methods. (Left) Evaluation of p_{fold}^{N} directly from the original MD trajectories as explained in Sec. II C and in Ref. 11. (Middle) Analytical calculation of pfoldt on the ETN as explained in Sec. II C and in Ref. 13. (Right) Analytical calculation of $p_{\mathrm{fold}}^{\mathrm{MSM}}$ with boundary conditions, i.e., definition of folded and unfolded state ensembles, based on the number of native contacts Q as described in Secs. II D and IV E. The values are given for all coarse-grained mesostates that were visited during the MD trajectory by a significant number of snapshots. The significance cutoff increases from top (20) to bottom (200). Note that pfoldt shows the sharpest decay of p_{fold} at the unfolding barrier and that the accuracy of pfoldt and p_{fold}^{N} improves when higher populated coarse-grained mesostates are used. In contrast, the $p_{\mathrm{fold}}^{\mathrm{MSM}}$ values are wrong independent of the statistics because of the flawed definition of folded and unfolded state ensembles, which leads to incorrect boundary conditions of the Markov state model.

β-sheet has an inherent symmetry and consists of two β-hairpins sharing the central β-strand. The histogram-based projection of the free energy onto the $(Q_{\mathrm{N}}, Q_{\mathrm{C}})$-plane showed two barriers separating the folded from the denatured state at $(Q_{\mathrm{N}}=4/11, Q_{\mathrm{C}}=9/11)$ and $(Q_{\mathrm{N}}=10/11, Q_{\mathrm{C}}=3/11)$, with the former lower by about 0.5 kcal/mol than the latter, as shown in Supplementary Material, Fig. S8.[37] To calculate $p_{\mathrm{fold}}^{\mathrm{MD}}$, multiple short MD runs were started from ten structures with $(Q_{\mathrm{N}}=4/11, Q_{\mathrm{C}}=9/11)$ and ten structures with $(Q_{\mathrm{N}}=10/11, Q_{\mathrm{C}}=3/11)$. The value of $p_{\mathrm{fold}}^{\mathrm{MD}}$ was equal (or very close) to 1 or 0 for 19 of the 20 putative TSE structures (data not shown). This failure is not surprising considering the sharp decay of $p_{\mathrm{fold}}^{\mathrm{MD}}$ at the cFEP barrier [Fig. 4(a)], which suggests that the correct identification of the TSE is very sensitive and not possible at all if the choice of the progress variable(s) results in projections that do not preserve the barrier(s). Therefore, free-energy projections onto geometric variables are in general not appropriate to determine the folding TSE.

V. CONCLUSIONS

The accurate determination of the TSE is essential for understanding the protein folding reaction. This paper deals with the automatic extraction of folding TSE structures for a simple two-dimensional energy surface and from MD simulations of a structured peptide. The cFEP, a barrier-preserving projection able to fully quantify the kinetic and thermodynamic properties of a system at equilibrium,[12] is shown to successfully determine TSE conformations at the top of the transition region to enter or leave a free-energy basin. On the other hand, free-energy projections onto geometric coordinates like the fraction of native contacts or the rmsd from the native structure are shown to fail (for the structured peptide) as most of the conformations at the maxima of the projected surface do not belong to the TSE. This failure is a consequence of the sharpness of the folding transition barrier and the fact that such projections do not

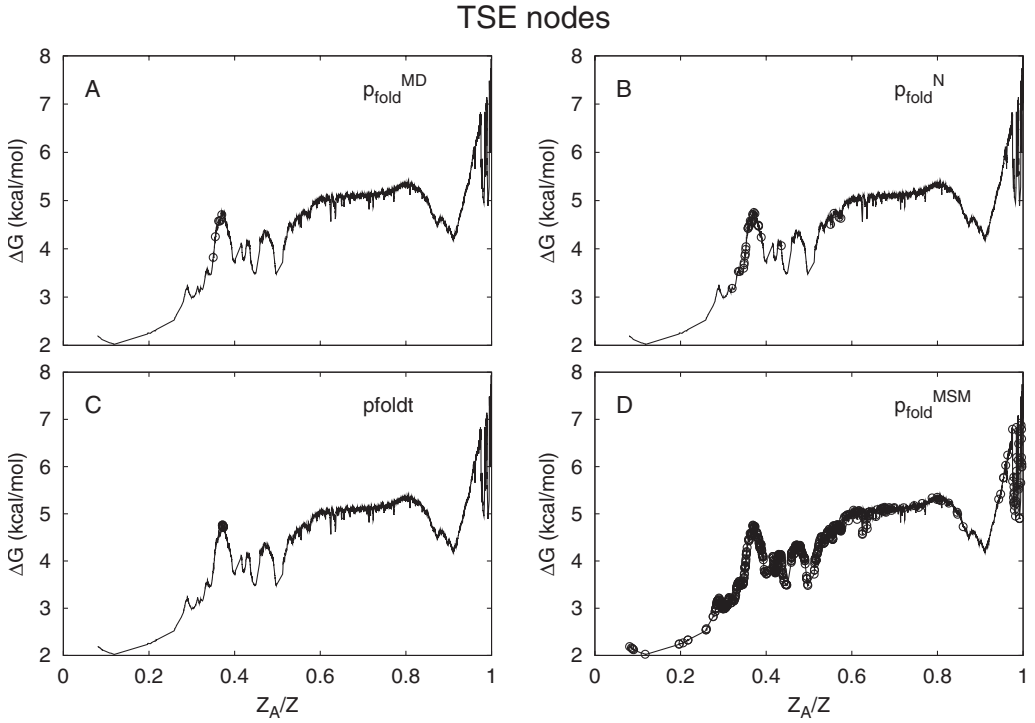

FIG. 7. Putative TSE determined by (a) p_{fold}^{MD}, (b) p_{fold}^{N}, (c) pfoldt, and (d) p_{fold}^{MSM}. Nodes with $0.45<p_{fold}<0.55$ and 20 or more snapshots are shown (empty circles). (a) Of the 34 nodes used for p_{fold}^{MD} calculations, only four nodes belong to the putative TSE region and are situated close to the top of the cFEP unfolding barrier. (b) Although some nodes with $0.45<p_{fold}^{N}<0.55$ are located close to the cFEP unfolding barrier, the TSE isolated by p_{fold}^{N} is affected by statistical error. (c) The TSE isolated by pfoldt is situated exactly on top of the barrier. (d) Most of the putative TSE nodes suggested by the p_{fold}^{MSM} approach do not belong to the TSE.

preserve the location of the barriers. The TSE determination has been attempted previously only for minimally frustrated systems[16,39,40] or for reactions involving a small and well-defined region of a protein.[41] For such reactions, an automatic procedure can identify reaction coordinates from an initial guess of several thousands physical variables, but requires the evaluation of commitment probabilities by additional simulations,[42] which is computationally prohibitive for a large set of structures.

In contrast to the automatic and parameter-free TSE determination by the cFEP, conventional p_{fold}-based methods involve the choice of a commitment time, or the arbitrary selection of representative regions for the native and the denatured state. The TSE isolation from the original MD trajectory (p_{fold}^{N}) (Ref. 11) or by analytical calculation on the ETN (pfoldt)[13] is very efficient and does not require any additional simulations, but the results depend on the choice of the commitment time. Moreover, p_{fold}^{N} values can be biased if insufficient amount of statistics is harvested, especially at the transition region, which is naturally sampled less than the free-energy minima.

More problematic is the p_{fold} calculation with a Markov state model (p_{fold}^{MSM}) because for a complex free-energy surface it is not possible to define the boundary conditions (i.e., $p_{fold}^{MSM}=0$ and 1) by simple structural criteria. This implies that most choices of such boundary regions lead to wrong p_{fold}^{MSM} results and thus to a flawed or incomplete isolation of the TSE. It is important to note that the same coarse-graining of the structures and ETN is employed in the p_{fold}^{MSM} calculation and the cFEP approach, but only the latter does not require that the denatured state is defined *a priori*.

The difficulty related to the calculation of p_{fold} lies in nuances of its definition: p_{fold} is the probability to fold before unfolding.[14] While p_{fold} calculated using a commitment time approximates this definition, p_{fold}^{MSM} between regions $\mathcal{F}$ and $\mathcal{U}$ is the probability to visit region $\mathcal{F}$ before $\mathcal{U}$, which corresponds to the original definition of p_{fold} only if the trajectory visits $\mathcal{F}$ and $\mathcal{U}$ each time it folds and unfolds, respectively, but not in between. Therefore, it is likely that p_{fold}^{MSM} calculations will be valid only in very special cases, e.g., in a two-state system with two enthalpic basins, where (simple) geometrical criteria are sufficient to separate the states. In contrast, the cFEP is able to isolate the TSE from a complex free-energy surface and does not necessarily require (long) equilibrium folding-unfolding simulations, as recently shown for an ETN obtained by short segments of replica exchange MD trajectories.[43]

ACKNOWLEDGMENTS

The authors thank Dr. R. Pellarin and P. Schütz for critical reading of the manuscript. They thank Dr. A. Cavalli for performing some of the simulations for p_{fold}^{MD} evaluation. The MD simulations were performed on the Etna and Matterhorn computer clusters at the University of Zurich. This work was supported by a Swiss National Science Foundation grant to one of the authors (A.C.).

[1] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes, Science **254**, 1598 (1991).
[2] S. E. Jackson and A. R. Fersht, Biochemistry **30**, 10248 (1991).

[3] A. Matouschek, J. T. Kellis, Jr., L. Serrano, and A. R. Fersht, Nature (London) **340**, 122 (1989).
[4] C. M. Dobson, A. Šali, and M. Karplus, Angew. Chem., Int. Ed. **37**, 868 (1998).
[5] J. A. Ihalainen, J. Bredenbeck, R. Pfister, J. Helbing, G. A. Woolley, and P. Hamm, Proc. Natl. Acad. Sci. U.S.A. **104**, 5383 (2007).
[6] S. E. Radford, C. M. Dobson, and P. A. Evans, Nature (London) **358**, 302 (1992).
[7] A. Li and V. Daggett, Proc. Natl. Acad. Sci. U.S.A. **91**, 10430 (1994).
[8] L. Li and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **98**, 13014 (2001).
[9] J. Gsponer and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **99**, 6719 (2002).
[10] N. Singhal, C. D. Snow, and V. S. Pande, J. Chem. Phys. **121**, 415 (2004).
[11] F. Rao, G. Settanni, E. Guarnera, and A. Caflisch, J. Chem. Phys. **122**, 184901 (2005).
[12] S. V. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689 (2006).
[13] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus, J. Phys. Chem. B **112**, 8701 (2008).
[14] R. Du, V. S. Pande, A. Y. Grosberg, T. Tanaka, and E. S. Shakhnovich, J. Chem. Phys. **108**, 334 (1998).
[15] I. Hubner, J. Shimada, and E. Shakhnovich, J. Mol. Biol. **336**, 745 (2004).
[16] S. S. Cho, Y. Levy, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **103**, 586 (2006).
[17] G. Settanni and A. Fersht, Biophys. J. **94**, 4444 (2008).
[18] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **97**, 10780 (2000).
[19] F. Rao and A. Caflisch, J. Mol. Biol. **342**, 299 (2004).
[20] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus, Nature (London) **409**, 641 (2001).
[21] G. Settanni, J. Gsponer, and A. Caflisch, Biophys. J. **86**, 1691 (2004).
[22] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **105**, 13841 (2008).
[23] M. Seeber, M. Cecchini, F. Rao, G. Settanni, and A. Caflisch, Bioinformatics **23**, 2625 (2007).
[24] D. Chandler, J. Chem. Phys. **68**, 2959 (1978).
[25] W. Swope, J. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[26] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **101**, 14766 (2004).
[27] E. De Alba, J. Santoro, M. Rico, and M. A. Jiménez, Protein Sci. **8**, 854 (1999).
[28] P. Ferrara, J. Apostolakis, and A. Caflisch, Proteins: Struct., Funct., Bioinf. **46**, 24 (2002).
[29] S. Muff and A. Caflisch, Proteins: Struct., Funct., Bioinf. **70**, 1185 (2008).
[30] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[31] B. R. Brooks, C. L. Brooks III, A. D. MacKerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, "CHARMM: The biomolecular simulation program," J. Comput. Chem. (in press).
[32] J. Gsponer, U. Haberthür, and A. Caflisch, Proc. Natl. Acad. Sci. U.S.A. **100**, 5154 (2003).
[33] J. A. Ihalainen, B. Paoli, S. Muff, E. Backus, J. Bredenbeck, G. A. Woolley, A. Caflisch, and P. Hamm, Proc. Natl. Acad. Sci. U.S.A. **105**, 9588 (2008).
[34] J. Gsponer and A. Caflisch, J. Mol. Biol. **309**, 285 (2001).
[35] J. Hartigan, *Clustering Algorithms* (Wiley, New York, 1975).
[36] N.-V. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[37] See EPAPS Document No. E-JCPSA6-130-069912 for the supplementary material including explanations on the progress coordinate Z_A/Z and cFEP with different progress variables, as well as the following supplementary figures: cFEP with x-axis transformed into mfpt (Fig. S1); normalized histograms of $p_{\mathrm{fold}}^{\mathrm{MD}}$ for the 34 nodes used for folding simulations (S2); pfoldf cFEP with an extra-node (S3); dependency of $p_{\mathrm{fold}}^{\mathrm{MD}}$ and p_{fold}^{N} on τ_{commit} (S4); distribution of number of native contacts (S5); results of the Markov state model with rmsd-based definition of boundary states (S6); correct TSE and putative TSE determined by p_{fold}^{N} and $p_{\mathrm{fold}}^{\mathrm{MSM}}$, respectively (S7); and histogram-based two-dimensional projection of the free energy (S8). For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[38] F. Noe and S. Fischer, Curr. Opin. Struct. Biol. **18**, 154 (2008).
[39] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **97**, 5877 (2000).
[40] R. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **102**, 6732 (2005).
[41] J. Hu, A. Ma, and A. R. Dinner, Proc. Natl. Acad. Sci. U.S.A. **105**, 4615 (2008).
[42] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).
[43] S. Muff and A. Caflisch, J. Phys. Chem. B **113**, 3218 (2009).