Identification of the protein folding transition state from molecular dynamics trajectories

SUPPORTING INFORMATION

S. Muff and A. Caflisch

Department of Biochemistry,
University of Zurich,
Winterthurerstrasse 190,
CH-8057 Zurich, Switzerland
tel: +41 44 635 55 21,
fax: +41 44 635 68 62,
e-mail: caflisch@bioc.uzh.ch

Keywords: molecular dynamics; transition state ensemble; p_{fold}; free-energy profile; denatured state ensemble

A. Z_A/Z as progress coordinate

The cFEP projected onto the relative partition function Z_A/Z has the advantage that the first basin on the left (reference basin, usually the folded one) is isolated with its population quantified by the x-axis value at the first barrier on the left. Other progress coordinates can be used, e.g., the mean first passage time (mfpt).The advantage of the cFEP projection onto mfpt is that rates of folding from individual basins are readable from the x-axis[1]. Note that the cFEP is invariant with respect to arbitrary transformations of the reaction coordinate[2].

B. cFEPs with other progress variable than mfpt

The progress *coordinate* (Z_A/Z or mfpt) is used to project the cFEP, while the progress *variable* is required to sort the nodes for the cFEP. For each node in the equilibrium transition network (ETN) two progress variables can be evaluated: mfpt and p_{fold}. Mfpt calculations require the selection of only one node, i.e., the native node[1]. Alternatively, an extra node, which is connected to all nodes in the network by a link weighted proportionally to a Lagrange multiplier λ, is needed in the pfoldf procedure to represent the unfolded state[3]. The introduction of the extra node is a stratagem to circumvent the arbitrary selection of a node as representative of the unfolded state. The results of this work are robust upon the choice of the progress variable (see Figure S3).

C. Supplementary Figures

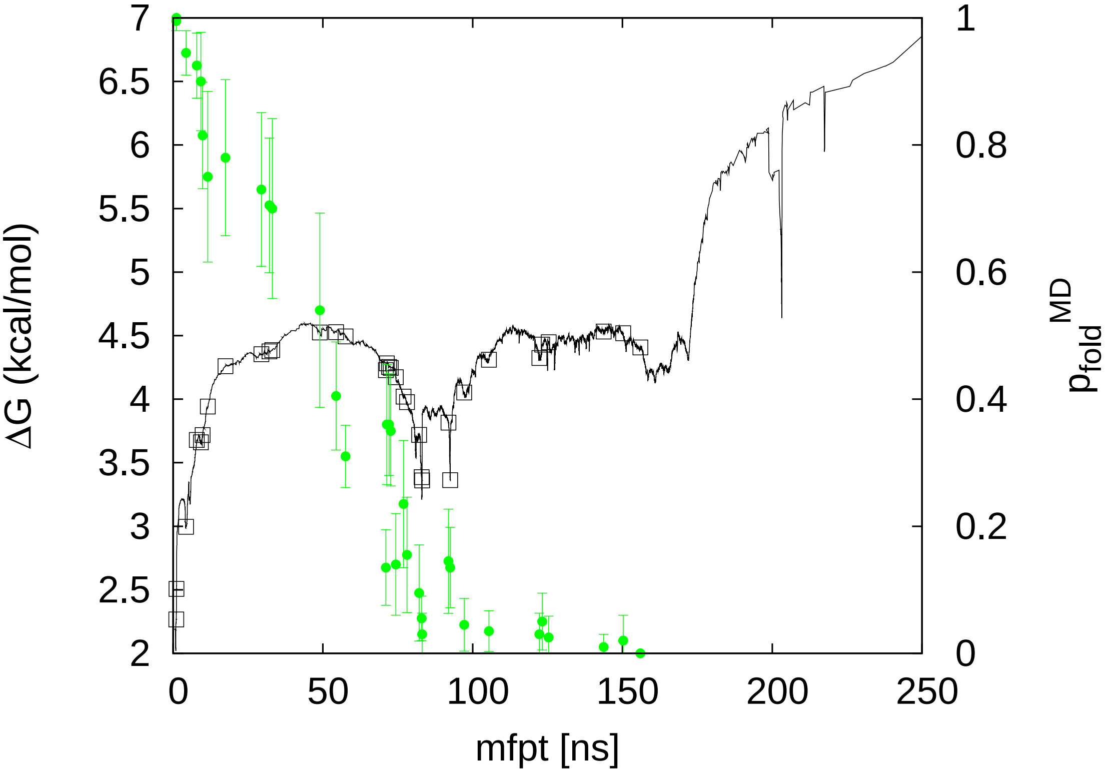

FIG. S1: cFEP with x-axis transformed into mfpt (black line). The p_{fold}^{MD} values (green circles) refer to the right y-axis and are given for the same 34 nodes as in Fig. 4 of the main text (black squares). The decay of p_{fold}^{MD} appears not as sharp as for Z_A/Z because only few nodes populate the region around 50ns, which is the average folding time of TSE structures (because half contribute about 100 ns, and the remaining less than 10 ns)

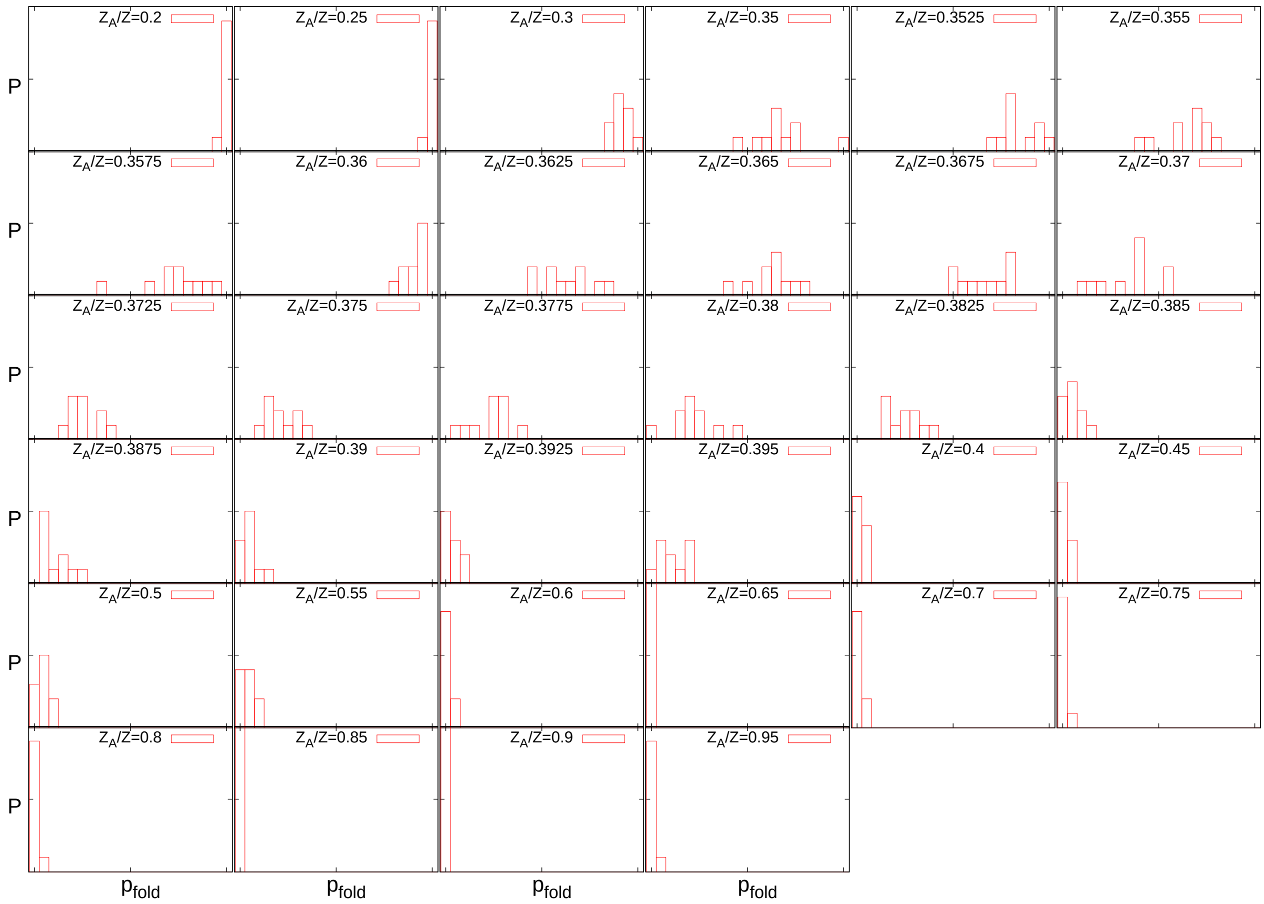

FIG. S2: Normalized histograms of p_{fold}^{MD} for the 34 nodes used for folding simulations. According to these plots, p_{fold}^{MD} values of individual snapshots are peaked around the average value of the respective node, indicating that the coarse-graining procedure applied here groups snapshots in a kinetically homogeneous way.

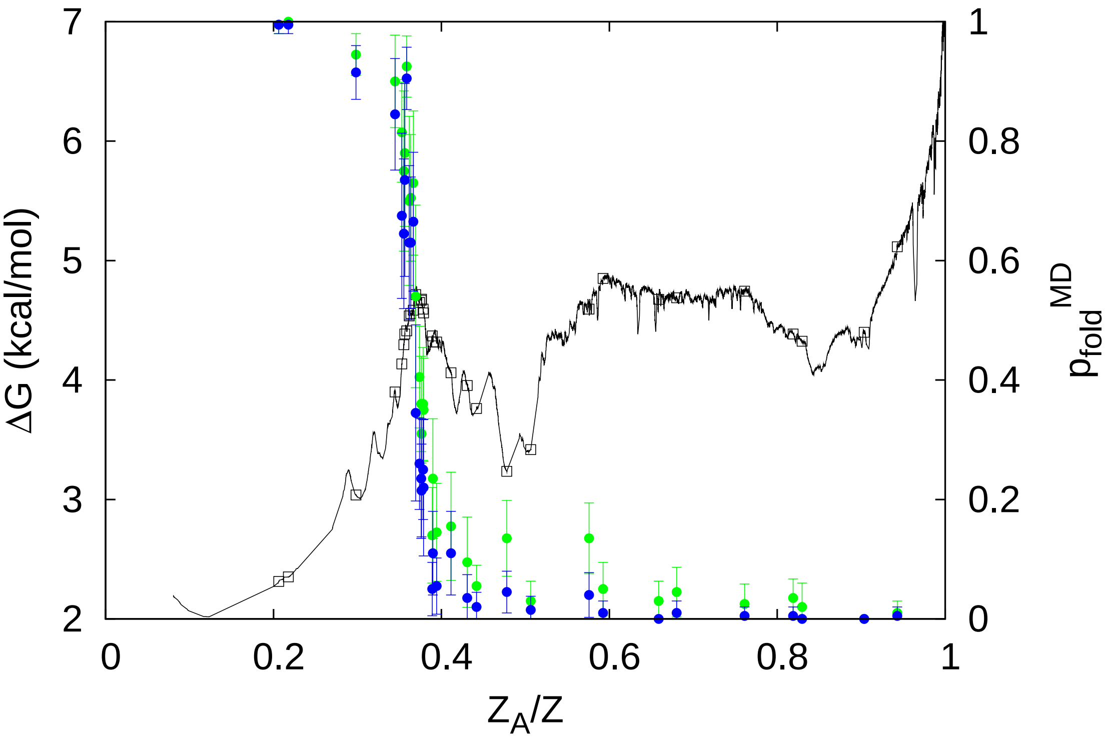

FIG. S3: Pfoldf-cFEP with an extra-node connected by a capacity λ =0.001[3] (black line) and the same selection of nodes as in Fig. 4 of the main text chosen for additional simulations (black squares). p_{fold}^{MD} results for τ_{commit} = 5 ns (blue) and τ_{commit} = 10 ns (green) are essentially identical. The similarity to the corresponding mfpt cFEP (Fig. 4A of the main text) indicates that the procedure is robust upon variation of the progress variable. The similarity of mfpt and p_{fold} profiles is expected, because both encode for kinetic distance to the native state and the equation system for analytical calculation of mfpt and p_{fold} from the ETN differs only in the explicit time dependence of the former[1].

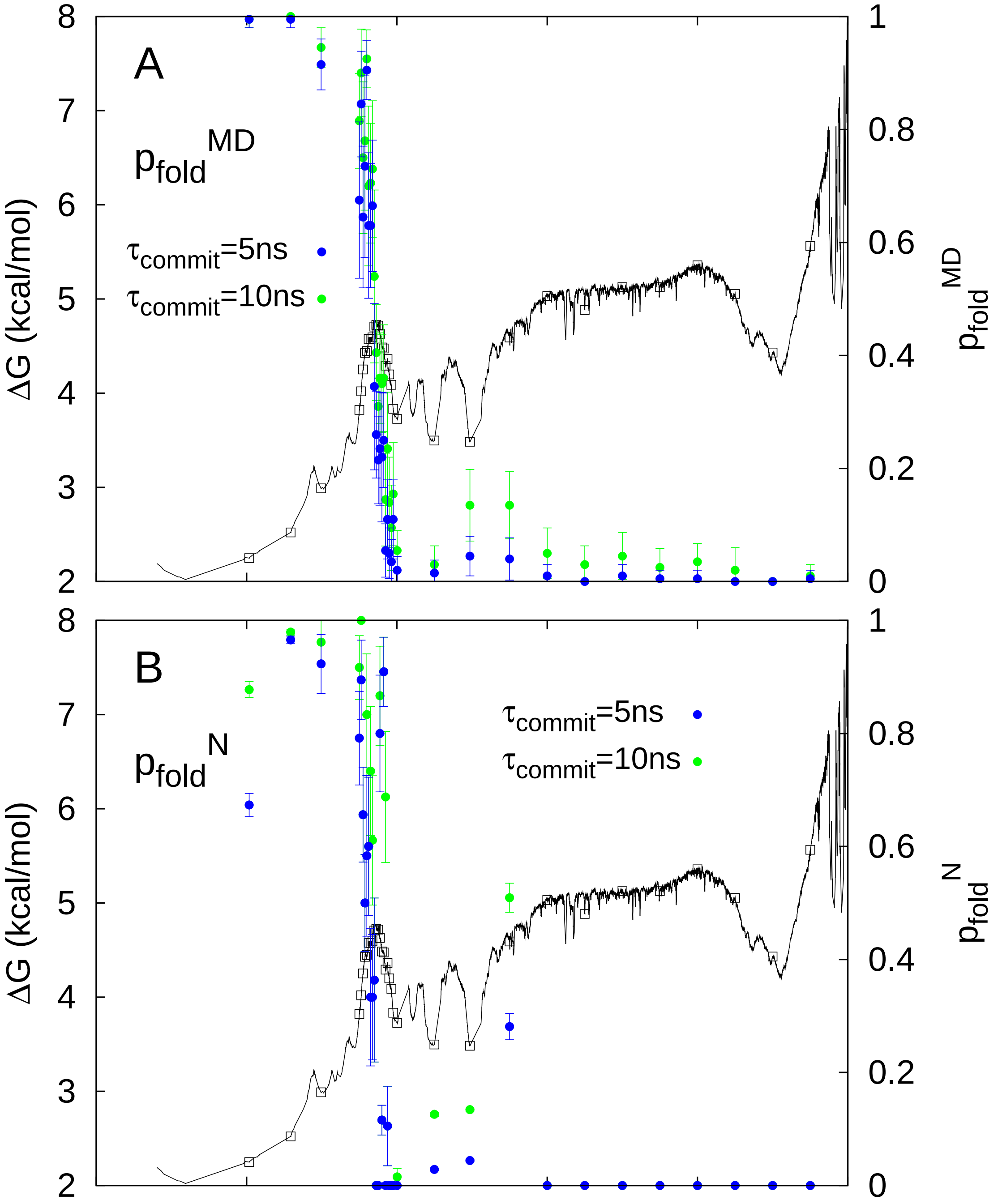

FIG. S4: Dependency of p_{fold}^{MD} (A) and p_{fold}^{N} (B) on τ_{commit}. The same plot as Fig. 4 in the main text, but with τ_{commit} = 5 ns (blue circles) and τ_{commit} = 10 ns (green circles). The results are almost identical.

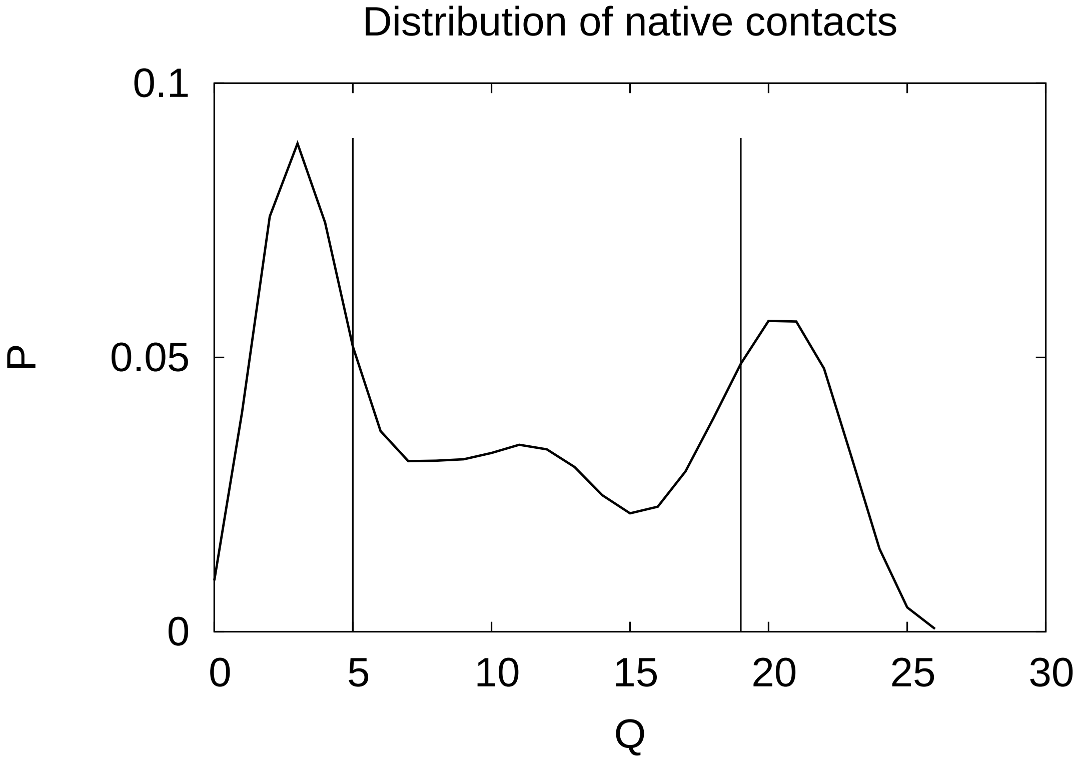

FIG. S5: The 26 native contacts were defined in Ref.[4]. Nodes whose structures have Q $>$ 19 in average were defined as folded, those with Q $<$ 5 as unfolded.

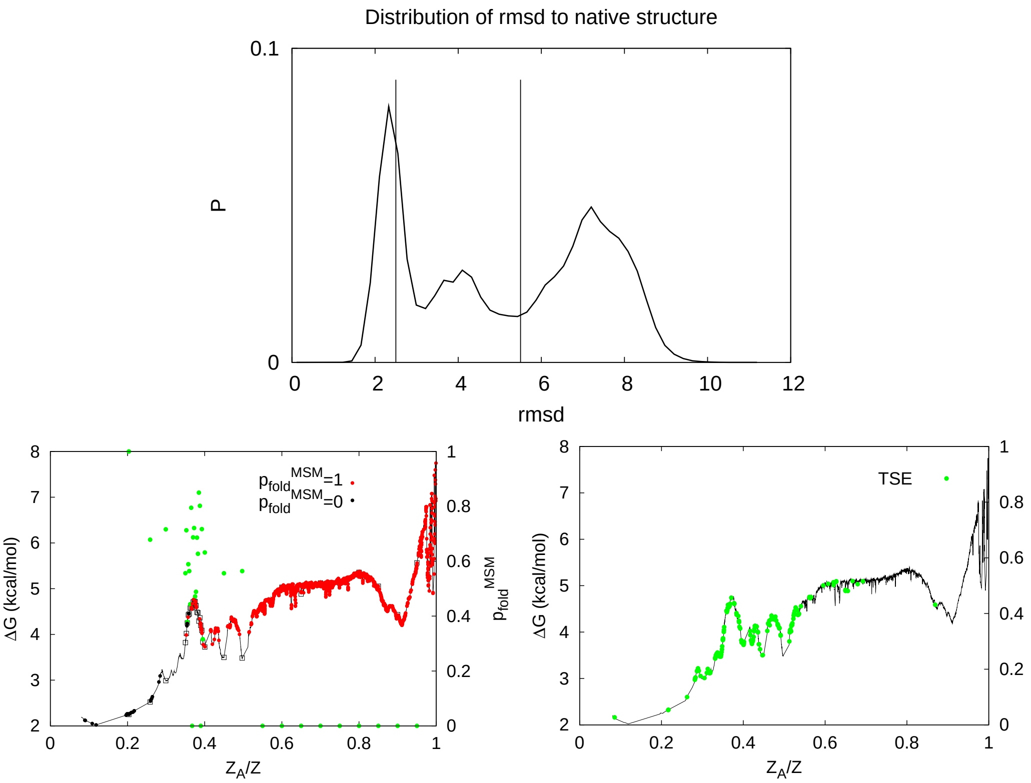

FIG. S6: Results of the Markov state model with rmsd-based definition of boundary states. (Top) Nodes with an average of rmsd $<$ 2.5 Å from the native structure were defined as folded, those with rmsd $>$ 5.5 Å as unfolded. (Bottom) The plots corresponding to Figure 4D (left) and 7D (right) of the main text show that some of the unfolded nodes have $p_{fold}^{\mathrm{MSM}} > 0.5$ and putative TSE structures are suggested far away from the barrier, respectively.

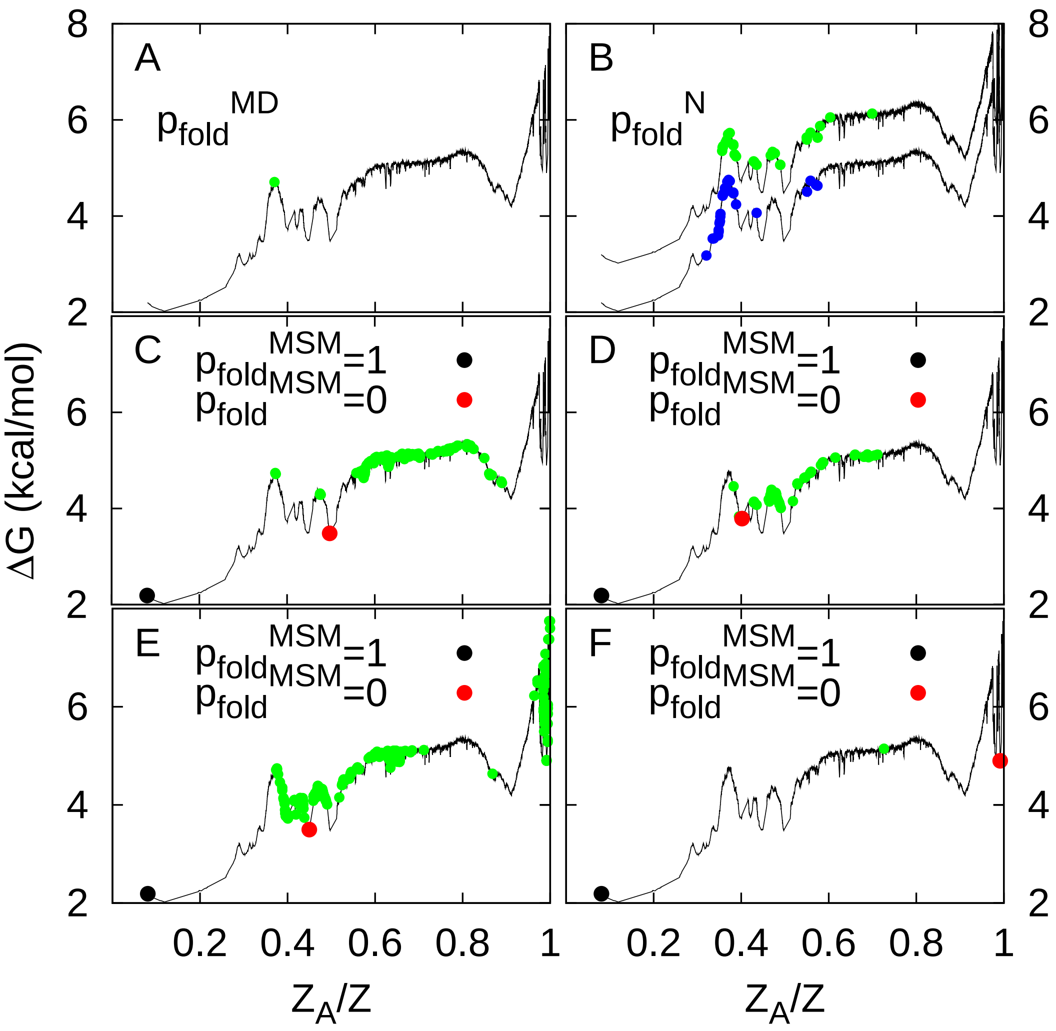

FIG. S7: Correct TSE (A) and putative TSE determined by p^{N}_{fold} (B), and p^{MSM}_{fold} (C-F). Nodes with $0.45 < p_{fold} < 0.55$ and 20 or more snapshots are shown (green circles). (A) Values of p^{MD}_{fold} were calculated using $\tau_{commit} = 10$ ns. (B) Values of p^{N}_{fold} were calculated using $\tau_{commit} = 5$ ns (blue circles) or $\tau_{commit} = 10$ ns (green circles). One of the two profiles is shifted vertically for visual clarity. (C-F) Different representatives of the denatured state (red circles) are used as boundary condition $p^{\mathrm{MSM}}_{fold} = 0$. The profiles are shown to illustrate that most of the putative TSE structures suggested by the p^{MSM}_{fold} approach do not belong to the TSE.

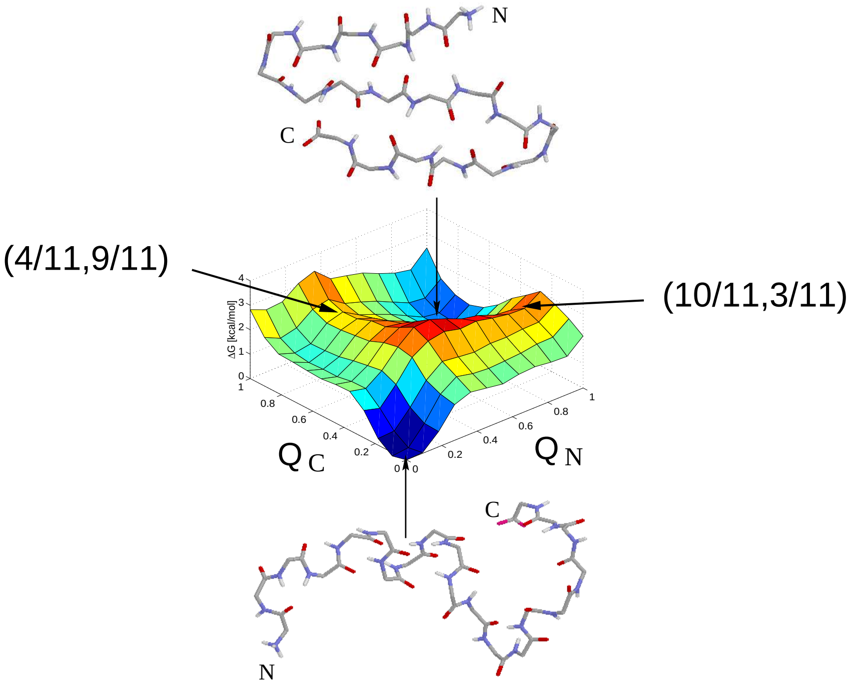

FIG. S8: It is not possible to extract the TSE from conventional histogram-based projections of the free energy onto geometric progress variables. The arrows show the location on the surface of the snapshots used for p_{fold}^{MD} calculations (Figure adapted from [5]).

[1] S. V. Krivov, S. Muff, A. Caflisch, and M. Karplus, J. Phys. Chem. B **112**, 8701 (2008).

[2] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. USA. **105**, 13841 (2008).

[3] S. V. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689 (2006).

[4] P. Ferrara and A. Caflisch, Proc. Natl. Acad. Sci. USA. **97**, 10780 (2000).

[5] A. Cavalli, P. Ferrara, and A. Caflisch, Proteins: Structure, Function, and Bioinformatics **47**, 305 (2002).