One-Dimensional Barrier Preserving Free-Energy Projections of a β-sheet Miniprotein: New Insights into the Folding Process SUPPLEMENTARY MATERIAL

Sergei Krivov,1[†] Stefanie Muff,2[†] Amedeo Caflisch,2[*] and Martin Karplus,1,3[*]

1 Laboratoire de Chimie Biophysique,
ISIS F-67000 Strasbourg, France

2 Department of Biochemistry,
University of Zurich,
Winterthurerstrasse 190,
CH-8057 Zurich, Switzerland

and

3 Department of Chemistry & Chemical
Biology Harvard University Cambridge,
Massachusetts 02138 U.S.A. [†]

(Dated: December 18, 2007)

Keywords:

A. Iterative calculation of $p_{fold}(\tau_{commit})$ on the EKN

The calculation of p_{fold} values basing on the equilibrium kinetic network (EKN) has been described previously[1] and in the main text. The calculation of $p_{fold}(\tau_{commit})$ values is based on a different system of equations and therefore requires additional considerations. Let p_i be the p_{fold} of node i. Then:

$$p_i = P[\tau_f(i) \leq \tau_{commit}] \ ,$$

with $\tau_f(i)$ representing the first passage time to the native node, starting in node i. Given a simulation with saving frequency Δt, the system of equations to be solved is

$$\begin{aligned} P[\tau_f(i) \leq \tau_{commit}] &= \sum_j p_{ji} P[\tau_f(j) \leq \tau_{commit} - \Delta t] \\ &= \sum_j p_{ji} \left(P[\tau_f(j) \leq \tau_{commit}] - P[\tau_f(j) = \tau_{commit}] \right) \ , \end{aligned} \qquad (1)$$

where p_{ji} is the transition probability from i to j and the sum runs over all nodes of the EKN. The system is bound by the condition $p_A = 1$. Let us first evaluate $P[\tau(j) = k], \Delta t \leq k \leq \tau_{commit}$, where $P[\tau(j) = k] = P[T_k = A|T_0 = j]$ with T_K equal to the probability to be in the native node A after k steps, starting from node j (not necessarily the first passage time). To avoid costly multiplication of the whole transition matrix, it is easier to evaluate the "reverse" probability to be in node j after k steps starting in A, $P[T_k = j|T_0 = A]$, because this can be calculated at once by iterative multiplication of the starting configuration $P[T_0 = j|T_0 = A] = \delta_{j,A}$ by the transition matrix:

$$P[T_{k+\Delta t} = j|T_0 = A] = \sum_i p_{ji} P[T_k = i|T_0 = A] \ .$$

Since the EKN fulfills detailed balance, the probability of the $j \rightarrow A$ transition can be calculated by

$$\underbrace{P[T_k = A|T_0 = j]}_{=P[\tau(j)=k]} = P[T_k = j|T_0 = A] \cdot \frac{P[A]}{P[j]} \ ,$$

where $P[A], P[j]$ are the relative populations of the nodes. The probability for the *first* passage time τ_f to node A can thus be calculated by

$$P[\tau_f(j) = \tau_{commit}] = \left(\prod_{n=1}^{(\tau_{commit} - \Delta t)/\Delta t} (1 - P[\tau(j) = n\Delta t]) \right) \cdot P[\tau(j) = \tau_{commit}] \ ,$$

i.e., the probability *not* to return within $\tau_{commit} - \Delta t$, but within exactly τ_{commit}. Inserting this expression into equation (1) and solving the system of equations yields the correct folding probabilities.

Figure S1 shows the FEP of Beta3s for different values of τ_{commit} and makes clear that too short commitment times are not suitable to fully resolve the unfolded state.

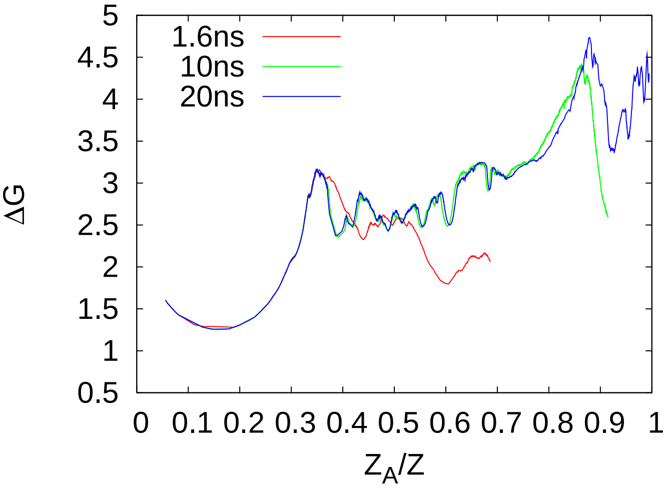


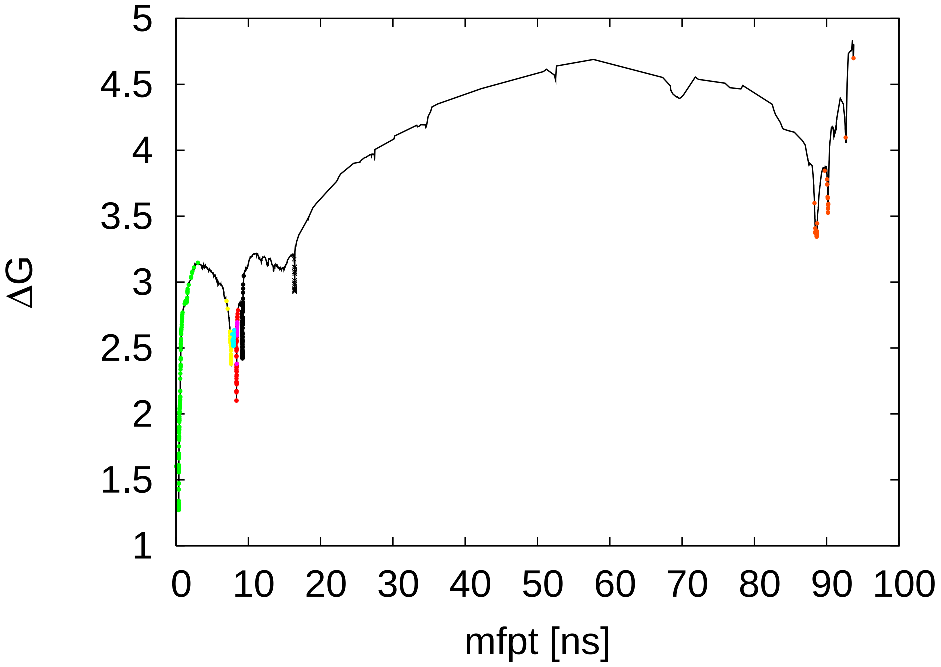
FIG. S1: FEP with $p_{fold}(\tau_{commit})$ calculations for τ_{commit}=1.6ns, 10ns and 20ns. It is important to choose the commitment time long enough to resolve the unfolded state. In fact, using a τ_{commit} value of only 1.6 ns (red curve) about 30% of the conformations have $p_{fold} = 0$ so that the profile stops at Z_A/Z=0.7.

B. Differences between pfoldf and mfpt FEPs

The deviations between the FEPs obtained by the two procedures originate from at least two points. First, p_{fold} calculations are bound by two conditions (p_A=1, p_B=0), and mfpt calculations only by one (τ_A=0). Second, the p_{fold} values are calculated on a slightly different (biased) underlying EKN due to the extra node that is used in the pfoldf procedure. When the nodes are sorted according to decreasing p_{fold} or increasing mfpt, the Spearman correlation coefficient of the noderanks (ρ) decreases with increasing λ (for λ=0.0001: ρ=0.9997; for λ=0.01: ρ=0.988), because a larger λ enhances the bias. If the mfpt and pfoldf FEPs are calculated on the same underlying EKN with the extra node connected with capacity λ=0.0001, pfoldf and mfpt are still not identical, although very similar with ρ=0.9999.

C. Mfpt as progress coordinate

The progress coordinate of the FEPs is the relative partition function of the EKN Z_A/Z, so that no information on the the underlying progress variable (p_{fold}, $p_{fold}(\tau_{commit})$ and mfpt) is present in the final plot. It is, however, straightforward to project the profile onto the original variable. In this way, the progress coordinate and the underlying progress variable are the same. Such a transformed profile shows ΔG as a function of the kinetic distance (in time units) from the native state (Figure S2) and provides supplementary information to the Z_A/Z projection. A disadvantage of the projection onto mfpt is that the non-native enthalpic basins are very close together in the profile because most of them have similar mfpt values (especially on the secondary structure network, where most values are around 10 ns for the mfpt values calculated by numerical solution of the equation $\text{mfpt}_i = \Delta t + \sum p_{ji} \cdot \text{mfpt}_j$, as detailed in the Methods section of the main text). Note that for the network with nodes coarse-grained according to secondary structure the numerically calculated mfpt values are smaller than those calculated directly from the trajectory (i.e., if one would follow the trajectory each time a node is visited), which arises from the fact that the secondary structure coarse-graining is too generous (see below) and because the solution of the mfpt equation system is equivalent to running a very long (infinite) Monte-Carlo (MC) simulation. On the other hand, performing the same analysis on the network obtained by the 2.5 Å RMSD coarse-graining, the mfpt values are very close to those found directly from the trajectory.



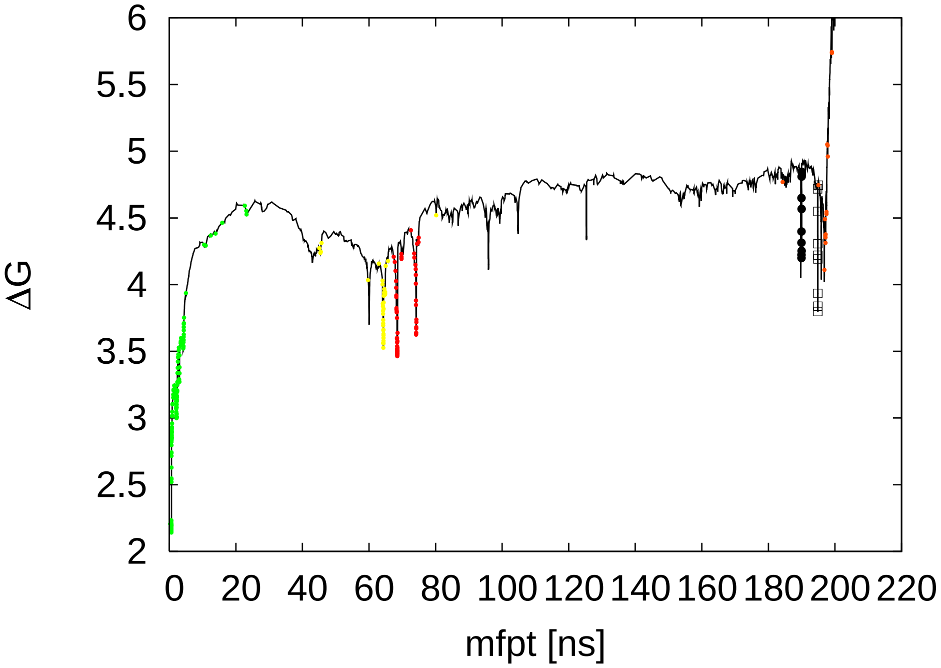


FIG. S2: Beta3s unfolding FEP calculated for the EKN (see Methods) using mfpt as a progress coordinate *and* progress variable for the secondary structure (top) and 2.5 Å RMSD coarse-graining (bottom). As in Figure 4 of the main text, individual basins are colored according to the basins extracted by the pfoldf procedure. Note that values of mfpt for individual basins are larger, and the barrier separating the native basin from the rest is higher for 2.5 Å RMSD than for secondary structure coarse-graining because of pseudo-tunneling affecting mainly the latter. Importantly, the similar mfpt values for the enthalpic traps, and a spread of only about three between mfpt values of enthalpic traps and the helical basin, is consistent with the single-exponential behavior of folding (see Results section in the main text).

D. Coarse-graining and Monte Carlo simulations

The main disadvantage of RMSD coarse-graining (clustering is used here as a synonymous) is the required computer time. For the one million snapshots of the 20 μs trajectory, all-atom RMSD clustering with a cutoff of 2.5 Å and 2.0 Å requires 10 days and (an estimate) 40 days, respectively, on a 2.8 GHz Intel Xeon. On the other hand, the disadvantage of secondary structure coarse-graining is revealed if MC simulations are performed on the resulting directed network. The folding time decreases from 100 ns to about 10 ns, whereas MC simulations on the directed network obtained by all-atom RMSD coarse-graining with 2.5 Å cutoff yield the correct value of 100 ns. Interestingly, a finer graining (RMSD 2.0 Å) increases the folding time to 137 ns, whereas the coarser RMSD of 3.0 Å decreases it to 84 ns. This phenomenon can be explained as follows: A very fine grained clustering yields low populations even for clusters in the native state. With a non-neglectable probability it can then happen that a folding event is not accounted because the trajectory does not visit the most populated (native) node before it unfolds, because the cluster is too small. On the other hand, a very coarse assignment of nodes as for RMSD 3.0 Å or secondary structure reduces the folding time in the MC simulation. The reason for the latter is that, due to the lax restriction of nodelimits, pseudo-tunneling happens frequently between nodes that are in reality separated by a significant barrier. Each pseudo-tunneling event introduces a ”shortcut” into the network which is taken into account in the MC simulation, even though folding never really proceeds via such a shortcut in the molecular dynamics (MD) simulation. This property leads to non-Markovianity. It has been observed earlier that the equivalence between MC and MD kinetics does not follow automatically and depends on the coarse-graining procedure[2] (Figure S3).

Interestingly, despite the considerable differences between the two methods used for coarse-graining, the basins isolated by pfoldf with secondary structure or 2.5 Å RMSD clustering are almost identical (Table S-I). Each snapshot belongs to a coarse-grained conformation, so it is grouped to the basin of the respective conformation. Basins are therefore comparable snapshot by snapshot and a similarity can be calculated analogous to Table II in the main text. Both the KGA and pfoldf procedures are not noticeably affected by the shortcuts (i.e., by the non-Markovian character) in the secondary structure

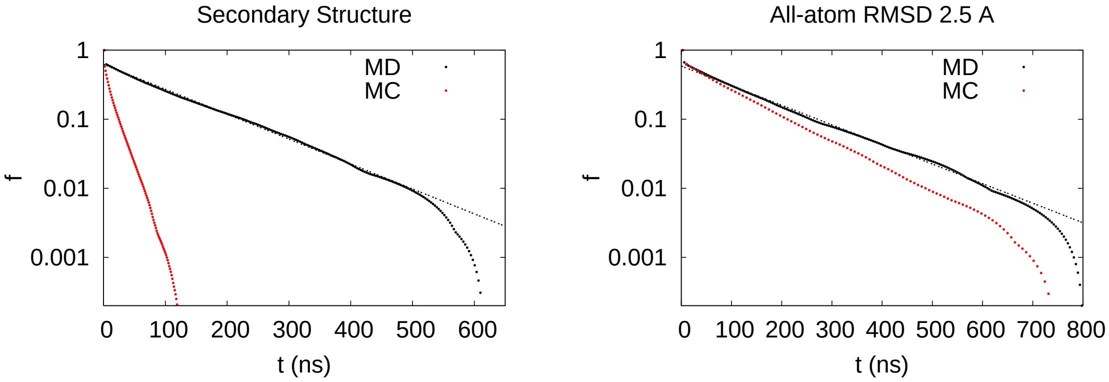


FIG. S3: Cumulative distribution of first passage times for the secondary structure (left) and the 2.5 Å RMSD coarse-graining (right). Black dots are extracted from the MD simulation, red dots from a 200μs MC simulation on the directed network. The folding kinetics of the secondary structure-based MC trajectory differ considerably from MD kinetics, whereas with 2.5 Å RMSD clustering almost identical MC and MD folding kinetics are observed.

coarse-graining. However, the free-energy barrier in the unfolding FEP of the native state (obtained by pfoldf) is about 0.5 kcal/mol higher using the all-atom 2.5 Å RMSD clustering (Figure S4) than the secondary structure clustering (Figure 2 of the main text), which shows that, in contrast to the isolation of basins, the extraction of barriers is very sensitive on the coarse-graining.

Basin		Weight (%)		Number of nodes		
Heaviest node	Name	sstruct	RMSD	sstruct	RMSD	Similarity^a
`-EEEESSEEEEEESSEEEE-`	Native	35.0	36.4	2672	6457	99.5
`-EEEESTTEEEEESSEEEE-`	Ns-or	6.2	3.2/2.9	1278	220/798	98.4/95.8
`-EEEESSEEEEESSSEEEE-`	Cs-or	2.6	3.8	967	5167	98.5
`-HHHHHHHHHHHHS------`	Helix	11.6	11.2	57134	49049	95.4
`---SSGGG---EESSEETT-`	Ch-curl$_1$	2.8	2.8	2153	430	95.0
`---SSGGG-EESSTTTTEE-`	Ch-curl$_2$	2.1	2.0	1675	119	98.8

TABLE S-I: Comparison of most populated basins of Beta3s obtained by pfoldf using either secondary structure or all-atom 2.5 Å RMSD clustering. Ns-or is split into two basins of almost equal size for RMSD, but the partitioning is also visible in the one-dimensional FEP generated using secondary structure clustering (Figure 3 of the main text). ^{a}The similarity value is calculated as the intersection of two corresponding basins, normalized to the lower population.

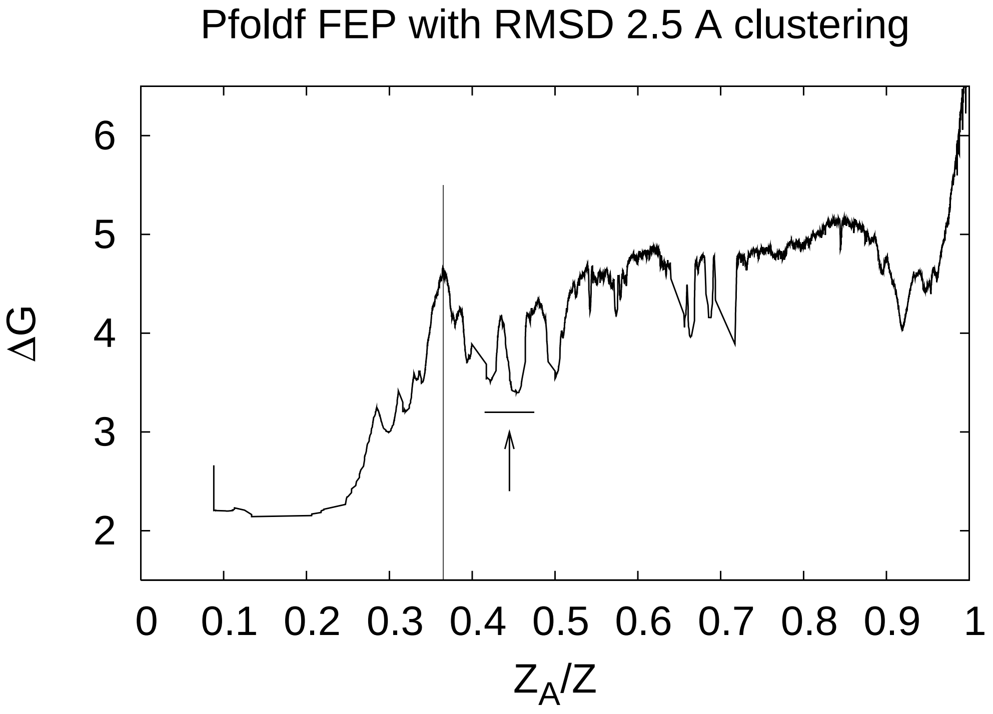


FIG. S4: Pfoldf-FEP of Beta3s using the snapshots (from the MD trajectories) clustered according to all-atom 2.5 Å RMSD. The vertical line shows the position of the unfolding barrier as extracted from the pfoldf procedure. The arrow and horizontal segment indicate the Ns-or basin which is split into two using 2.5 Å RMSD clustering. Note that this profile is very similar to the one obtained using secondary structure coarse-graining (Figure 2 top of the main text), but the barrier of the native basin is higher for 2.5 Å RMSD. In both the most distant basin from native is the helical basin.

A main difference between secondary structure and all-atom RMSD coarse-graining is that the former lacks the information about the position and orientation of the sidechains. Therefore, it is possible that conformations belonging to the same secondary structure string are separated by barriers that arise from differences in the orientation of sidechains. To exemplify the concern, Figure S5 shows two structures belonging to the native secondary structure node (`-EEEESSEEEEEESSEEEE-`), one with the Tyrosine19 sidechain pointing upward and one pointing down. The 2.5 Å RMSD coarse-graining correctly separates these two structures into two different clusters.

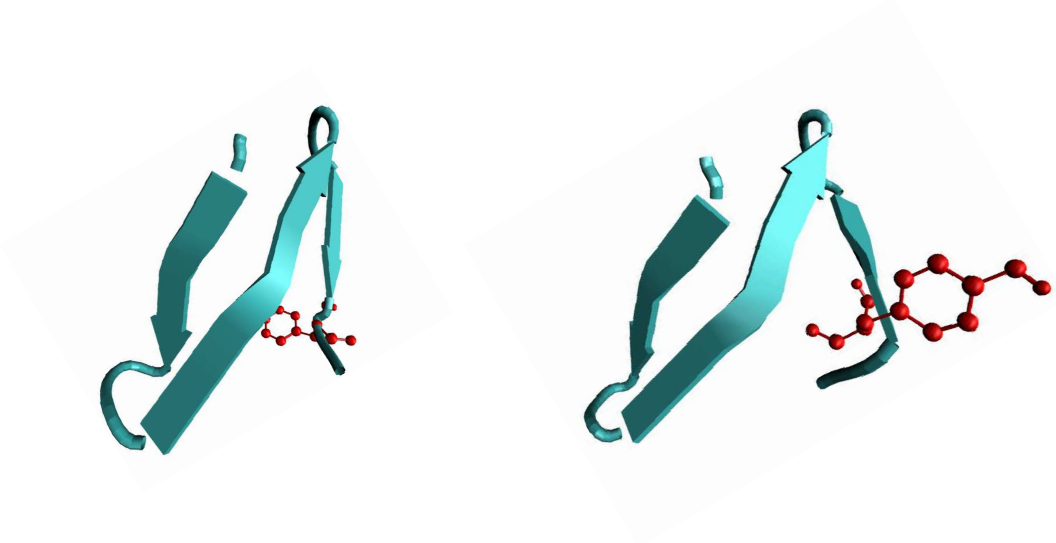


FIG. S5: Two snapshots belonging to the native secondary structure string, despite a completely different orientation of the Tyrosine19 sidechain.

E. Barriers in the entropic region

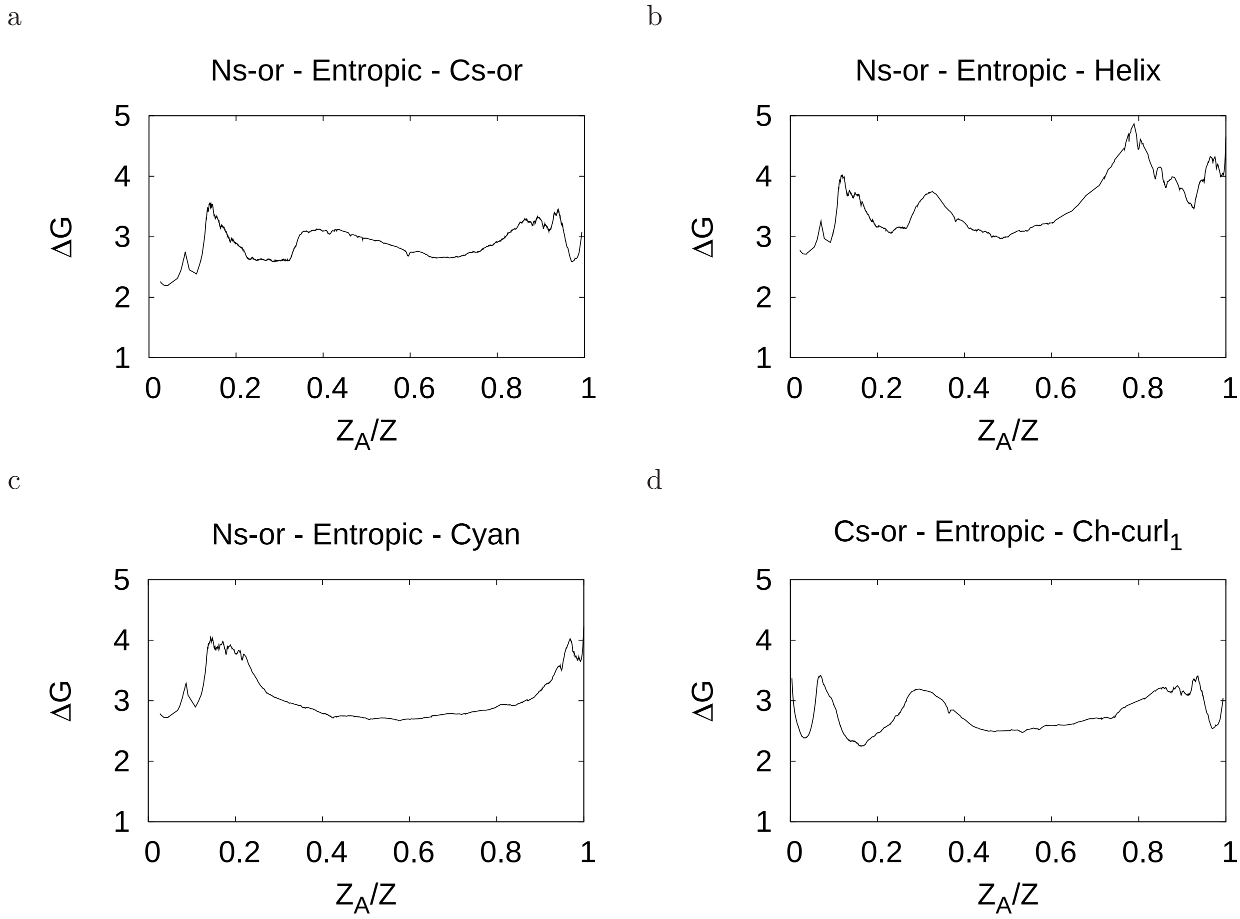


FIG. S6: Reduced pfoldf profiles. Only two enthalpic basins plus the entropic region are used to plot these FEPs. The entropic region, which stretches between the first (second in the case of Ns-or) and the last barrier, reveals barriers (a, b, d) that are otherwise invisible. The pairs of basins were chosen such that very few (or no) direct transitions between them were observed in the simulation except for Ns-or/cyan. Secondary structure-based coarse-graining was used for these profiles.

[*] corresponding authors, tel: +33 390 24 5123 fax: +33 390 24 5124, e-mail: `marci@tammy.harvard.edu,caflisch@bioc.uzh.ch`

[†] SK and SM have made equal contributions to this study.

[1] S. V. Krivov and M. Karplus, *J. Phys. Chem. B*, **2006**, *110*, 12689–12698.

[2] S. V. Krivov and M. Karplus, *Proc. Natl. Acad. Sci. USA.*, **2004**, *101*, 14766–14770.