Free Energy Surfaces from Single-Distance Information

Philipp Schuetz,[†] René Wuttke, Benjamin Schuler,* and Amedeo Caflisch*

Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Received: June 11, 2010; Revised Manuscript Received: September 21, 2010

We propose a network-based method for determining basins and barriers of complex free energy surfaces (e.g., the protein folding landscape) from the time series of a single intramolecular distance. First, a network of transitions is constructed by clustering the points of the time series according to the short-time distribution of the signal. The transition network, which reflects the short-time kinetics, is then used for the iterative determination of individual basins by the minimum-cut-based free energy profile, a barrier-preserving one-dimensional projection of the free energy surface. The method is tested using the time series of a single $C_\beta - C_\beta$ distance extracted from equilibrium molecular dynamics (MD) simulations of a structured peptide (20 residue three-stranded antiparallel β-sheet). Although the information of only one distance is employed to describe a system with 645 degrees of freedom, both the native state and the unfolding barrier of about 10 kJ/mol are determined with remarkable accuracy. Moreover, non-native conformers are identified by comparing long-time distributions of the same distance. To examine the applicability to single-molecule Förster resonance energy transfer (FRET) experiments, a time series of donor and acceptor photons is generated using the MD trajectory. The native state of the β-sheet peptide is determined accurately from the emulated FRET signal. Applied to real single-molecule FRET measurements on a monomeric variant of λ-repressor, the network-based method correctly identifies the folded and unfolded populations, which are clearly separated in the minimum-cut-based free energy profile.

I. Introduction

The thermodynamics and kinetics of a variety of complex systems, ranging from spin glasses to proteins, have been investigated by energy landscape theory in the 40 years since the publication of the seminal idea.[1] Peptides and proteins have a multidimensional and very complex potential energy surface with a large number of conformations of similar energy.[2,3] Yet, fast folding is possible because of the natural selection of sequences that make the native (i.e., functional) structure a pronounced energy minimum.[4] Entropic contributions are relevant at physiological temperatures, and therefore the *free* energy surface governs the thermodynamics and kinetics of polypeptide chains. In the past five years, new methods based on complex networks have been proposed to analyze free energy surfaces of folding,[5−10] which govern the process by which structured peptides or proteins assume their well-defined three-dimensional structure.

In view of the large number of microscopic folding pathways and the conformational heterogeneity in the denatured state, single molecule methods are a promising new approach to experimentally determine free energy surfaces.[11] One of the most versatile approaches, single molecule Förster resonance energy transfer (FRET), allows intramolecular distances and distance dynamics of individual protein molecules to be monitored.[12−17] Since distance distributions in different free energy states often overlap, the separability of the different basins is not straightforward.[18] Baba and Komatsuzaki suggested an approach (termed BK procedure hereafter) to extract free energy basins from the time series of a single distance.[19] The BK procedure is able to resolve different basins even if the distance distributions overlap because the short-time behavior of the observable is considered. Applied to a simplified model of a protein with 46 beads of three types (hydrophobic, hydrophilic, and neutral), the authors identified four free energy basins, in good agreement with the free energy surface derived using the complete structural information of the reference simulation.

Here we present a procedure for the automatic determination of free energy surfaces from single-molecule time series (FESST). First, an equilibrium transition network (ETN) is constructed by clustering individual time windows according to similarity in the short-time distribution of the signal, whose usage was inspired by the BK procedure.[19] The ETN is then used as the input for the minimum-cut-based free energy profile (cFEP) method, which is able to determine free energy basins and barrier heights (Figure 1).[7] The FESST parameters are optimized using an intrinsic cost function, the height of the unfolding barrier in the cFEP. This self-consistent choice of optimal FESST parameters leads to a unique solution in an objective and autonomous way, which allows for complete automatization of the procedure.

The accuracy of FESST is assessed using molecular dynamics (MD) trajectories of the 20-residue peptide β-3s,[20,21] whose sequence was designed to favor the three-stranded antiparallel β-sheet conformation, that is, a double β-hairpin.[22] β-3s has been shown to fold reversibly to the native structure determined by NMR[22] in MD simulations with the CHARMM polar hydrogen molecular mechanics potential energy function supplemented by a simple implicit solvent model.[23] In these simulations, β-3s folds in about 0.1 and 8 μs at 330 and 286 K, respectively.[24] Since multiple folding and unfolding events at the melting temperature of about 330 K can be simulated in less than a week (on a commodity processor), the free energy surface and the folding pathways and mechanism of β-3s have been

* To whom correspondence should be addressed. E-mail: schuler@bioc.uzh.ch; caflisch@bioc.uzh.ch.
[†] Current address: Empa, Swiss Federal Laboratories for Materials Science and Technology, Überlandstrasse 129, 8600 Dübendorf, Switzerland.

10.1021/jp1053698
Published on Web 10/22/2010

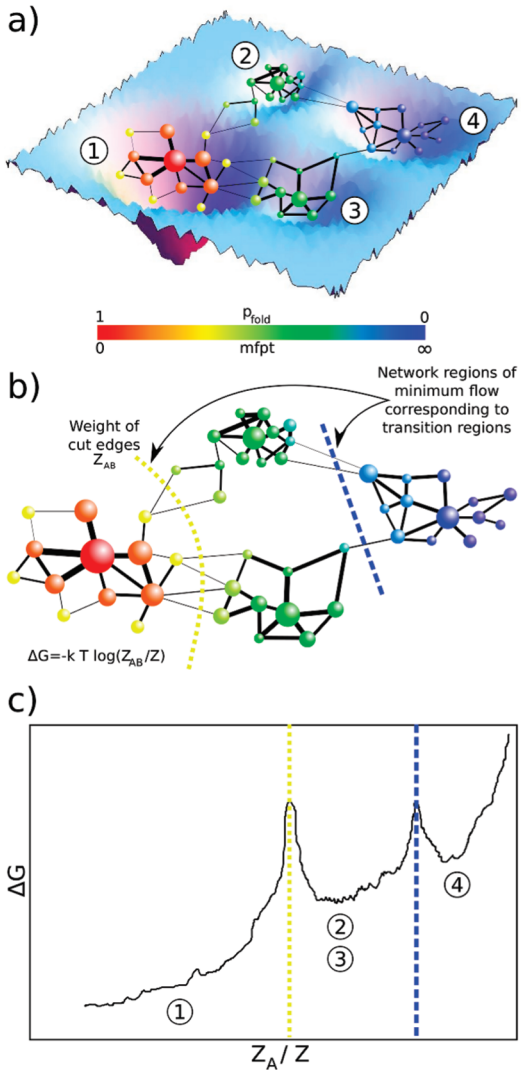

Figure 1. Illustration of the minimum-cut-based free energy profile (cFEP).[7] (a) The high-dimensional free energy surface is coarse-grained into nodes of the network. Two nodes are linked if the system proceeds from one to the other along the considered timeseries. The folding probability p_{fold} or the mean first passage time (mfpt) are calculated for each node analytically. Note that p_{fold} ranges from 1 (at the reference node) to 0 and mfpt from 0 to infinity. (b) For each value of p_{fold} (or mfpt), the set A of all nodes with a higher folding probability (or lower mfpt value) is defined. The free energy ΔG of the barrier between the two states formed by the nodes in A and the remainder of the network B can be calculated by the number of transitions Z_{AB} between nodes of either set.[7] (c) The cFEP is a projection of the free energy surface onto the relative partition function Z_A/Z, which includes all pathways to the reference node. For each value of p_{fold} (or mftp), the point (Z_A/Z, $-kT\,\log(Z_{AB}/Z)$) is added to the cFEP. The cFEP projects the free energy surface faithfully for all nodes to the left of the first barrier (basin 1). After the first barrier, two or more basins overlap (e.g., basins 2 and 3) if they have the same kinetic distance from the reference node.

investigated in detail.[6,20,21,25] The complexity of the free energy surface of β-3s[6] and its detailed characterization make it an ideal test system.

II. Methods

A. Free Energy Surface from Single-Molecule Time Series (FESST). FESST is a three-step procedure: construction of the ETN by clustering individual time windows using local kinetic information, identification of free energy basins by the cFEP approach, and removal of overlap from the non-native basins. The details of the three steps of FESST are presented in the next subsections and the Supporting Information (SI), while a schematic illustration is shown in Figure 2.

B. Coarse-Graining and Equilibrium Transition Network (ETN). Each time bin in the time series of the one-dimensional signal is assigned to a node of the ETN by the leader algorithm.[26] In the initialization step, the first bin is defined as the representative of the first node. At each successive bin t_n ($n > 1$), the distribution of the single-molecule observable within a short time window starting at time t_n (henceforth named the short-time distribution of t_n) is compared with the distributions of the previously visited representatives. The length of the time window is adjusted ideally such that it is about 1−5% of the characteristic time scale of the process monitored. To preserve the local kinetics (i.e., the actual dynamic evolution of the system), the comparison is carried out starting from the latest defined representative, that is, by parsing the list of representatives in inverse chronological order. A new node is defined whenever the short-time distribution of t_n deviates by more than a given threshold from the distributions of all previously defined representatives. In this way, one obtains a time series of nodes and a corresponding sequence of transitions between nodes, which is used to construct the ETN (Figure 2a,b).

C. Minimum-Cut-Based Free Energy Profile (cFEP). Krivov and Karplus have exploited an analogy between the kinetics of a complex process and equilibrium flow through a network to develop the cFEP, a projection of the free energy surface that preserves the barriers[7] and can be used for extracting folding pathways and mechanisms from MD simulations.[27] The input for the cFEP calculation is the ETN (Figure 1a), which is derived by the coarse-graining described above. For each node i in the ETN, the partition function is $Z_i = \sum_j c_{ij}$, that is, the number of times the node i is visited, where c_{ij} is the number of direct transitions from node i to node j observed along the time series. The transition probabilities can then be calculated as $p_{ij} = c_{ij}/\sum_k c_{ik}$. If the nodes of the ETN are partitioned into two groups A and B, where group A contains the reference node, then $Z_A = \sum_{i\in A} Z_i$ (the number of times a node in A is visited), $Z_B = \sum_{i\in B} Z_i$, and $Z_{AB} = \sum_{i\in A, j\in B} c_{ij}$ (the number of transitions between nodes in A and nodes in B). The free energy of the barrier between the two groups is $\Delta G = -kT\,\log(Z_{AB}/Z)$, where Z is the partition function of the full ETN (Figure 1b). The progress coordinate then is the normalized partition function Z_A/Z of the reactant region containing the reference node, but other progress coordinates can be used, because the cFEP is invariant with respect to arbitrary transformations of the reaction coordinate.[28]

In practice, the cFEP is calculated from the ETN in three steps: (1) The folding probability p_{fold} or the mean first passage time mfpt (Figure 1a) are calculated analytically for each node on the ETN by solving the system of transition rate equations.[7,27] (2) Nodes are sorted by decreasing values of p_{fold}, and for each of these values the relative partition function Z_A and the cut Z_{AB} are calculated (Figure 1b). (3) The individual points on the profile are evaluated as $[x = Z_A/Z, y = -kT\,\log(Z_{AB}/Z)]$ (Figure 1c). The result is a one-dimensional profile that preserves the barrier heights between the free energy basins; given the barriers, the basins can be determined.[7] It is important to note that the cFEP is a projection that preserves the heights of the barriers as long as the underlying coarse-graining does not group kinetically distant bins into the same node, that is, as long as the ETN captures the correct dynamics of the system. Time

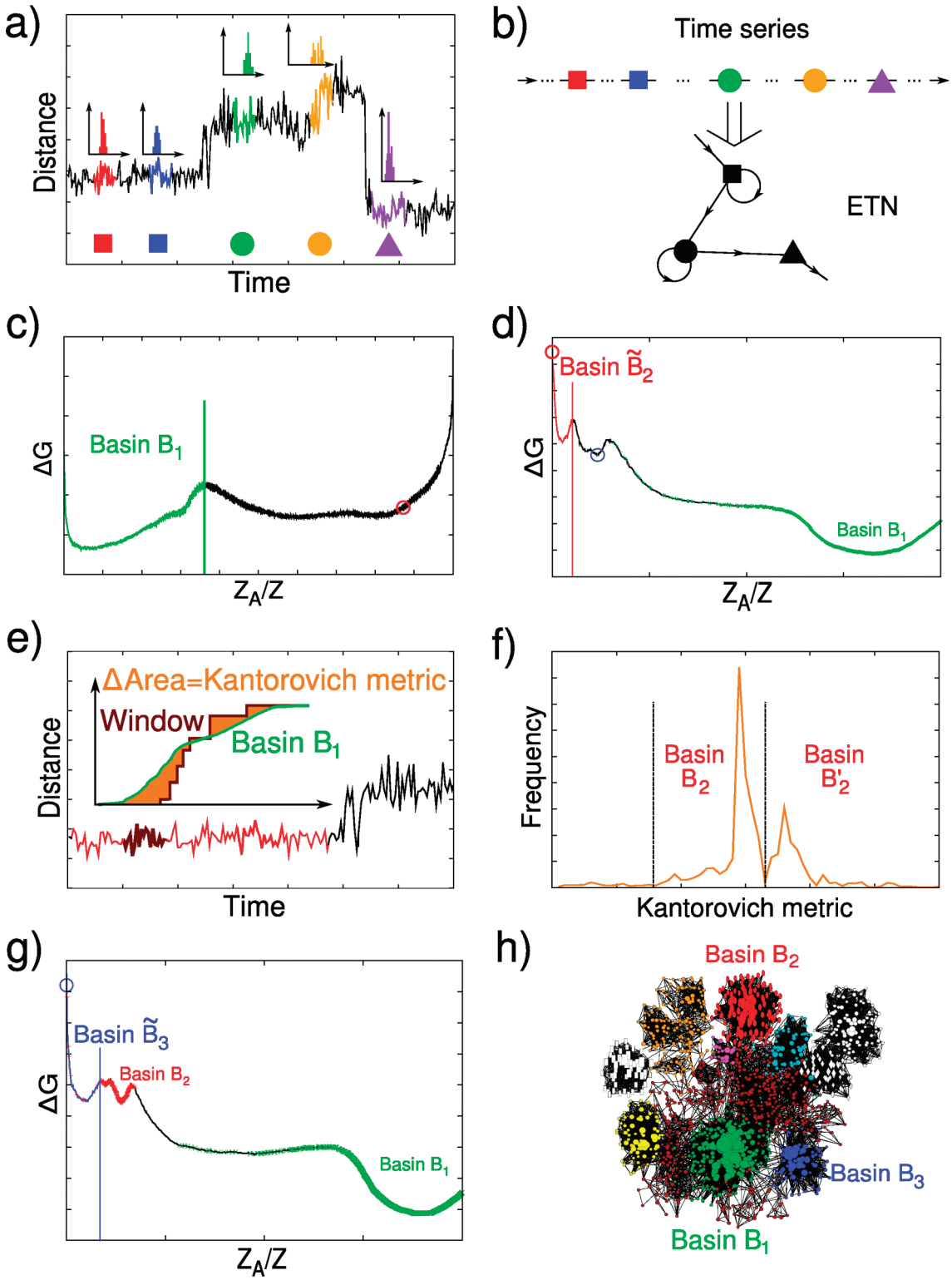

Figure 2. Schematic illustration of the FESST procedure. (a) The time series of the scalar signal is coarse-grained according to the short-time distribution of the distance. (b) The coarse-graining yields a time series of nodes and transitions that define the ETN. (c) The cFEP is plotted using the most populated node as a reference. The first free energy basin is isolated by cutting at the first barrier. The red circle indicates the most populated node outside the first basin, which is used to plot the cFEP for the determination of the second basin. (d) Because of the degeneracy of the short-term distance distribution, nodes from different free energy basins overlap in the second basin (see text). The tilde is used to denote a cFEP basin with overlap. The blue circle is the most populated node outside of $\tilde{B}_2$. (e,f) The overlap in $\tilde{B}_2$ is removed by comparing it with the entire distribution of the first basin (B_1). (g) The procedure is repeated for the next basin. (h) The basins extracted by FESST are illustrated on the conformation space network of β-3s with the native basin in green, and non-native basins Ch-curl$_1$ (B_2) and Ns-or$_1$ (B_3) in red and blue, respectively.[21]

bins misassigned by the coarse-graining result in spurious transitions between the basins and therefore lead to a lower barrier.[27]

All cFEPs in this paper are calculated with the software package WORDOM[29] using p_{fold} as the progress variable and an extra node for the $p_{fold} = 0$ boundary.[7]

D. Iterative Determination of Free Energy Basins. The most populated node is used to isolate the first basin by the cFEP approach. The barrier in the cFEP (Figure 2c) corresponds to the barrier leaving the basin identified first. For the remaining basins, the procedure is the same, except that the most populated unassigned node is used as a reference (Figure 2d). All nodes to the left of the cut at the first barrier make up the basin. Basins to the right of the first barrier are potentially overlapping (Figure 1c); thus, each basin requires a separate "exiting" profile.[27] Moreover, a FESST basin usually encompasses more than one of the true free energy basins, because the short-time distribution of the single distance can be degenerate. To remove this overlap, the signal's distribution in long time windows (ideally 10% of the time characteristic for the process monitored) starting from each bin assigned to the considered FESST basin is compared with the distribution of the signal in the entire basin identified previously (Figure 2e). Longer time windows are considered here for improved statistics and to exploit information complementary to the short-time distribution used in the construction of the ETN. Different subbasins in the basin to split are characterized by different ranges of the comparison metric (Figure 2f), because two distinct free energy basins differ in their similarity to a third one.

E. Static Model and Minimal-Kinetics Model. To investigate the importance of the system's kinetics and the signal's distribution in FESST, two simple procedures are tested for the identification of the native basin. They are called the static and the minimal-kinetics models as follows.

In the static model, a state is characterized by a range of observable values. To make the comparison with FESST as stringent as possible, the best possible static model is generated using the most accurate definition of the native basin.[27] For this purpose, the optimal observable range is determined for each completeness value (coverage of the native state by the basin identified, cf. Figure S3 of the SI) by testing multiple ranges and recording only the solution with the highest accuracy (fraction of the basin identified being native, cf. Figure S3).

In the minimal-kinetics model, each bin of the time series is assigned to the node defined by the discretized mean and standard deviation of the signal calculated over a short window around the bin considered. Subsequently, the resulting time series of nodes is analyzed by cFEPs as for FESST. This model incorporates local kinetic information by the consideration of the short-time evolution of the signal. In contrast to FESST, the minimal-kinetics model ignores the detailed structure of the signal's distribution. As for the static model, the parameters of the minimal-kinetics model, for example, the length of the window, are fine-tuned using an independent characterization of the native basin[27] as input.

III. Results. Application to MD Simulations of β-Sheet Folding

A. MD Simulations of β-3s. A total simulation time of 20 μs at 330 K was used for the FESST analysis. It has been shown previously that, in MD at 330 K, β-3s folds reversibly to the NMR conformation, irrespective of the starting structure; 23 of the 26 nuclear Overhauser effect constraints are satisfied.[20,21] All MD runs and most of the analysis of the trajectories were carried out with CHARMM.[30]

A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent.[23]

B. Intramolecular Distance and Metric Used for Coarse-Graining. The time series of the $C_\beta Gln_4 - C_\beta Thr_{16}$ distance is used in FESST, but the results are robust with respect to the choice of residue pairs as long as one of the two residues is in β-strand 1 and the other in β-strand 3. Two time windows $[t_1, t_1 + \tau]$ and $[t_2, t_2 + \tau]$ are grouped into the same node of the ETN if their distributions of the intramolecular distance (the short-term distribution) pass a Kolmogorov−Smirnov test,[31] which checks if two samples are picked from the same distribution. Each MD snapshot is used as a starting point of a time series bin, so that there are as many bins as coordinate frames along the MD trajectory. In other words, two successive bins are shifted by the MD saving interval of 20 ps. The length of the time window τ is chosen such that it is much shorter than the folding time, which is about 100 ns in MD simulations of β-3s at 330 K.[21] The dissimilarity of the two time windows is defined as the maximum difference of the cumulative distribution functions c_1, c_2 of the distance r (dissimilarity $= \max_{r>0}|c_1(r) - c_2(r)|$). The test is passed; that is, two time bins are grouped in the same node if

$$\text{dissimilarity} \leq \sqrt{\frac{2}{N}} \cdot \zeta$$

where N is the number of MD snapshots in each time window and ζ the acceptance cutoff that corresponds to a certain confidence level.[31] Note that the FESST results on β-3s are robust with respect to the choice of N in the range $30 \leq N \leq 250$ (i.e., 0.6 ns $\leq \tau \leq$ 5 ns) and ζ in the range $0.3 \leq \zeta \leq 1.5$ (Figure S1 of the SI). Values of $\tau = 2$ ns and $\zeta = 0.3$ are used in the following. To slightly improve on the sampling of the MD simulation, a detailed balance is imposed to the FESST-ETN by averaging the numbers of transitions c_{ij} and c_{ji} between nodes i and j.

C. Native Basin and Unfolding Barrier. The free energy basins of β-3s have been determined previously by the cFEP procedure using information on all 645 coordinates.[27] Since the full information of the peptide dynamics was taken into account, those free energy basins and barriers are used here as a reference for a critical evaluation of FESST and the comparison with other approaches.

Using the time series of the $C_\beta Gln_4 - C_\beta Thr_{16}$ distance, the native basin of β-3s is determined by FESST with remarkable accuracy (96% of the FESST native basin is part of the native state as determined by the 645 coordinates of cFEP, cf. Figure S3 of the SI) and completeness (95% of the native state of the 645 coordinates of the cFEP is captured by the FESST native basin, cf. Figure S3 of the SI). Moreover, the FESST unfolding barrier (defined as the free energy difference between the bottom of the first basin on the left in the cFEP and the top of the first barrier in the cFEP[7]) has a height (10.7 kJ/mol) very similar to the one obtained by the 645-coordinates cFEP (10.6 kJ/mol, Figure 3a and Figure S5 of the SI).

To investigate the influence of the choice of the residue pair monitored, each of the 154 $C_\beta - C_\beta$ pairs was tested in FESST. Remarkably, for 32 of these pairs the native basin is identified with an accuracy greater than 80% and at the same time a completeness of more than 90% (Figure 4b). Interestingly, the larger the separation along the sequence, the better the score. A notable exception is the 5−7 distance, which reflects the formation of the β-turn at the N-terminal hairpin. The distances yielding the best score are those between residues in β-strands 1 and 3 (top left part of the matrix in Figure 4b), which is likely to be a consequence of the β-sheet topology. Moreover, the $C_\beta - C_\beta$ distances involving the N-terminal β-strand show a higher score than those involving the C-terminal β-strand, which is consistent with the higher structural stability of the C-terminal

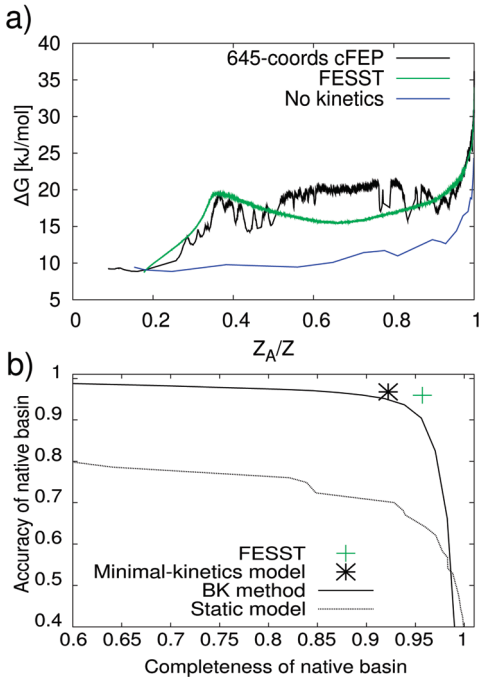

Figure 3. (a) Determination of the native basin by taking into account all structural information (black) or only a single distance (green). The cFEP is shown with the most populated node as a reference (for details, see the SI). If the short-time kinetics of the system is ignored, (only the network of transitions between the coarse-grained values of the single distance is analyzed), the cFEP (blue) displays no discernible barrier, and no meaningful basin can be extracted. (b) Comparison of FESST with a previously published single distance approach (BK = Baba and Komatsuzaki[19]) and two simple models (see the Model section for details). Note that FESST and the minimal-kinetics model yield a single data point rather than an accuracy versus completeness curve, because there is a unique optimum that can be determined by maximizing the barrier height.

hairpin.[20,21] In other words, the fully folded state can be better separated from non-native conformers by taking into account the N-terminal β-strand, because the C-terminal hairpin is folded correctly in the most populated non-native conformers.

FESST performs much better than the static model (Figure 3b and Figure S8 of the SI), which shows the importance of exploiting information about the short time kinetics. The cFEP calculated from the time series of coarse-grained distance values without taking into account the kinetic information displays no barrier, which indicates that a single distance value does not discriminate the native state of β-3s (Figure 3). If, in FESST, the time series was coarse-grained based on statistical properties (such as mean and standard deviation) instead of the signal's distribution, a native state of comparable quality would result (minimal-kinetics model in Figure 3b), but non-native basins are detected significantly less accurately (e.g., only 72% accuracy and 77% completeness for Ch-curl$_1$, which is defined below).

Although both approaches make use of short-time distributions of the signal, FESST has two advantages compared to the BK procedure.[19] First, FESST exploits the local kinetic information for the coarse-graining, while the BK procedure iteratively removes the time windows least similar to the distribution of the whole distance time series, thus ignoring the chronological order of the windows. Second, the optimal values of parameters required by FESST (size of the time window k and acceptance cutoff ζ used in the coarse-graining) can be determined automatically using the cFEP barrier height as a cost function, because the barrier height is the main determinant of the interconversion rates between the free energy basins. Therefore, the most accurate determination of the native basin yields the highest barrier.[27] Correspondingly, the parameter set yielding the highest barrier achieves the highest score (defined as the product of accuracy and completeness, Figure S2 of the SI). Therefore, FESST yields a single data point in the accuracy versus completeness plot (Figure 3b), whereas the basins extracted by other procedures depend on the cutoffs chosen for their iterative refinement, so that it is not possible to automatically identify the optimal solution.

D. Identification of Non-native Basins. The most populated node outside of the native basin is used as a reference to plot the cFEP profile for identifying the first non-native basin (termed $\tilde{B}_2$ in Figure 2). Because of the degeneracy of the short-time distribution of the distance, multiple free energy basins may overlap on the cFEP. Such overlap can be removed by comparing the long-time distance distribution of each time window with the distance distribution in a previously identified basin. In practice, for each time window $[t_2, t_2 + T]$ in basin $\tilde{B}_2$, the distribution of the distance is compared with the histogram of the entire native basin. Time windows of length

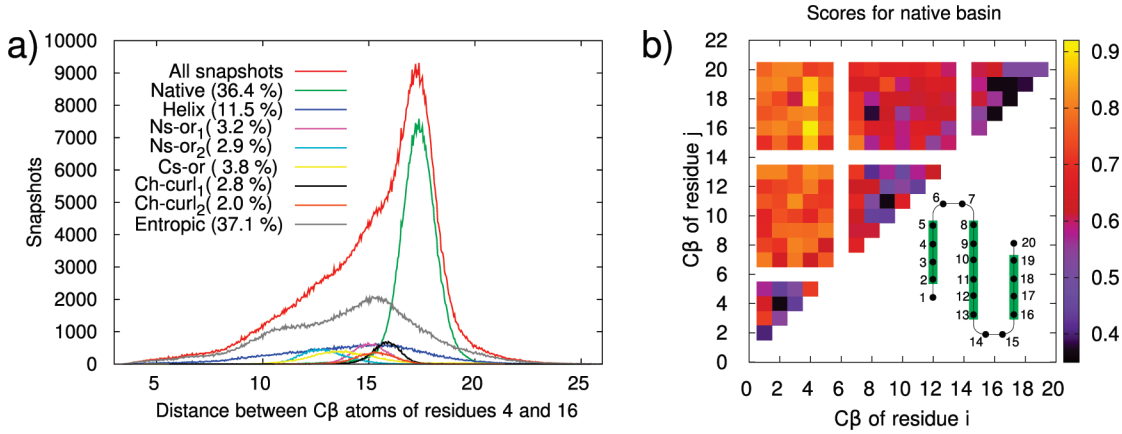

Figure 4. Robustness of FESST with respect to the choice of the distance. (a) Histograms of distance between the C_β atoms of residues 4 and 16 for the snapshots in each free energy basin determined by cFEP using all 645 degrees of freedom of β-3s.[27] (b) Matrix of scores for native state detection. Each (i,j) value of the score was calculated by applying FESST to the time series of distance between the C_β-atoms of residues i and j (Gly_6 and Gly_{14} have no C_β atom). The inset shows a schematic representation of β-3s with the three native β-strands (green rectangles). In the coarse-graining step of FESST, $N = 100$ distance values are compared, and the acceptance cutoff $\zeta = 0.3$ is used in the Kolmogorov−Smirnov test (cf. Section III B).

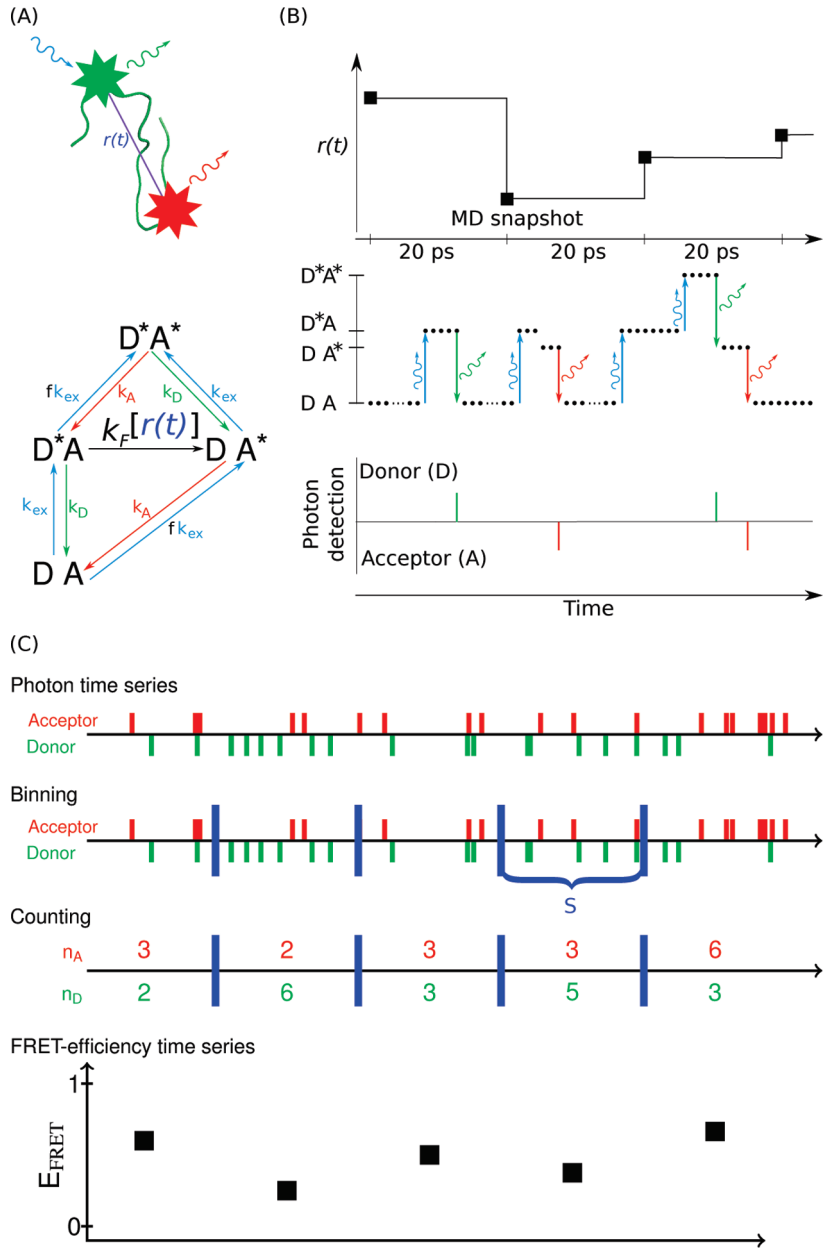

Figure 5. Markov state model used to generate the photon time series from the MD time series and illustration of the transformation of the photon time series into the FRET-efficiency time series. Red, green, and blue curled arrows represent acceptor photon emission, donor photon emission, and photon absorbance. (A, top) Schematic illustration of the emulated FRET experiment, where $r(t)$ represents the distance between the C_β-atoms of residues 4 and 16. (A, bottom) State diagram of the Markov process used to simulate the FRET experiment. (B) Illustration of the simulation of the FRET experiment, which uses the time series of the distance r, measured along the MD trajectory, to generate the photon time series. For each MD saving interval of 20 ps, 100 steps (black dots) of a random walker on the Markov state model are carried out with a constant value of the distance $r(t)$, i.e., the constant Förster rate $k_F[r(t)]$. Note that each excitation leads to an emitted photon in the FRET emulation. Direct excitation of the acceptor is set to $f = 5\%$ of the donor excitation rate.[40] (C) Transformation of the photon time series into the FRET-efficiency time series. The initial photon time series is binned with binning time S. In each bin, the number of acceptor photons n_A and donor photons n_D is counted. The FRET efficiency is then calculated using the formula $E_{FRET} = n_A/(n_A + n_D)$.

$T \approx (10 \text{ to } 20)\ \tau$, that is, significantly larger than those used for the construction of the ETN, are considered here for better statistics. The comparison consists of calculating the Kantorovich metric[32] between the two distributions (the area between the two cumulative histograms, Figure 2e). Finally, each peak in the histogram of the Kantorovich values is assigned to a new subbasin (Figure 2f). The window size T can be chosen by optimizing the separation of the different peaks in the Kantorovich histogram (Figure S9 of the SI).

With this procedure, the basin B_2 derived from $\tilde{B}_2$ corresponds to the 645 coordinates of the free energy basin Ch-curl$_1$ (curl-like conformation with folded C-terminal hairpin[27]) with 92% accuracy and 85% completeness. Further, the third FESST basin $\tilde{B}_3$ encompasses two free energy basins and can be split by comparing with the distance distribution in $\tilde{B}_2$. The free energy basin Ns-or$_1$ (N-terminal strand out of register and folded C-terminal hairpin[27]) can be extracted with 77% accuracy and 68% completeness. The second subbasin detected in B_3 contains 56% of MD snapshots in Ch-curl$_2$ (curl-like conformation 2[27]) covering 77% of these MD snapshots. These non-native conformers are stabilized mainly enthalpically.[27] Entropically stabilized conformations such as those in the "helical basin" and the "entropic region"[27] show a very broad distribution of distances (blue and gray curves in Figure 4a). These broad

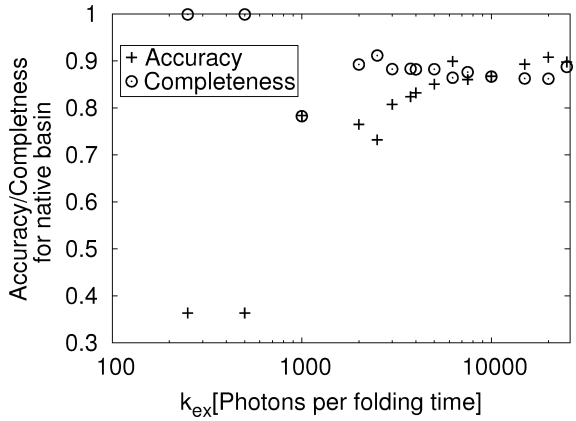

Figure 6. FESST performance on an emulated FRET experiment. The time series of FRET efficiencies calculated for 0.4 ns bins is used for the analysis by FESST with a window size of 25 bins (the effect of other window sizes is illustrated in Figure S11 of the SI) and acceptance cutoff $\zeta = 0.3$. The accuracy and completeness for the identification of the native basin is calculated for the set of snapshots in all bins of the first FESST basin (cf. Figure S3 for the definition of accuracy and completeness). The excitation rate k_{ex} is expressed as the average number of photons per folding time (about 100 ns for β-3s[21] in MD simulations at 330 K).

distributions overlap strongly with those of other basins and therefore are distributed over multiple FESST basins. In other words, both $\tilde{B}_2$ and $\tilde{B}_3$ contain time windows of the entropic region that can be removed by the procedure illustrated in Figure 2e,f. (For the native basin this step is not performed because the overlap of the distance distributions is much smaller than for $\tilde{B}_2$ and $\tilde{B}_3$, and no basin for comparison is available.)

IV. Application to an Emulated FRET Signal

A promising experimental method for obtaining intramolecular distance information in heterogeneous systems is single-molecule FRET.[12,15,17] To elucidate the applicability of FESST to such data, a FRET experiment is mimicked by generating a photon time series using a Markov state model (Figure 5). In this model, the rate of energy transfer $k_F(r)$ between the two "virtual" chromophores depends on the inverse sixth power of the distance r between the C_β-atoms of β-3s residues 4 and 16 as recorded along the MD trajectory (for details, see SI). From such photon time series, the native state of β-3s can be detected with 78% accuracy and 78% completeness from 1000 photons per folding time (Figure 6). This detection quality is obtained by comparing intervals of the time series of FRET efficiencies (for details, cf. Figure 5C and SI) as long as 10 ns, which corresponds to about one tenth of the folding time of β-3s in the MD simulation at the melting temperature. The detection quality depends only weakly on the size of each FRET bin (Figure S10 of the SI) and the length of the time series interval (Figure S11 of the SI). However, simulations of a simple two-state model indicate that the required number of photons is significantly reduced if the individual populations are slightly better separated in distance space (details in SI) than the free energy basins of β-3s (Figure 4a), suggesting that the application of the method to experiments is feasible.

V. Application to Single Molecule FRET Experiments

In a first attempt to apply FESST to real experimental data, we chose a protein whose folding dynamics at the unfolding midpoint (where both folded and unfolded state are populated) are in the range of a few milliseconds. This allows us to use experiments on freely diffusing molecules to maximize the excitation rate, while there is at the same time a large probability of observing folding or unfolding transitions during the diffusion time through the confocal volume of about a millisecond. As a result, we obtain a large number of short observations, similar in spirit to short simulations from parallel or distributed computing.[33] Even though these short observations will largely be independent, they can still be used to reconstruct the free energy surface if they are locally equilibrated, representative of the relevant conformational space, and provide a sufficient time resolution for the process investigated.[24,33]

We used a variant of monomeric λ-repressor, a well-established fast-folding protein[34,35] with a folding relaxation time of about 1 ms at the unfolding midpoint,[36] labeled it with Alexa Fluors 488 and 594 as FRET donor and acceptor, respectively, and investigated it with confocal single molecule experiments at a guanidinium chloride concentration of 0.68 M (see SI for details). Every protein molecule diffusing through the confocal volume will then result in a burst of photons. By selecting bursts with a duration of at least 2.5 ms and by working at high excitation rate, the average number of photons in the 48461 events collected during 4 h measurement time used was 575, which allows us to apply the FESST analysis with a data binning of 0.1 ms. FESST correctly identifies the folded and unfolded subpopulations (Figure 7a). Because of the high excitation power used, a significant contribution of acceptor photobleaching is present, which results in a third apparent population at lower transfer efficiencies.

The corresponding cFEP plotted from the unfolded state (Figure 7b) exhibits a well-defined folding barrier at $Z_A/Z = 0.42$, whose height can be maximized by node merging as for the analysis of the β-3s simulations (details in the SI, Figure S5). Even though the position of the barrier agrees with the relative populations, the resulting barrier height for folding (and for unfolding, see Figure S16 of the SI) is in the range of 1 kT, significantly lower than expected from the folding rate of 225 s^{-1} extracted from the frequency of transitions identified in the bursts. A factor that contributes to the reduction in barrier height is the imperfect separation of populations due to shot noise, resulting in spurious transitions in the FESST analysis. Another limitation is the time resolution achievable with photon detection rates in the range of about 0.14 MHz currently available in our free diffusion experiments, with which it is not possible to resolve the nanosecond diffusive dynamics of the polypeptide chain[16] from individual fluorescence bursts. Even though FESST is still limited by the photon rates in current single molecule experiments, the clear identification of subpopulations and the existence of a barrier in the resulting free energy profile illustrate its feasibility and potential for the analysis of experimental data.

VI. Discussion

FESST is a method for determining free energy basins and barriers from the time evolution of a scalar observable. The accuracy and range of possible applications of FESST have been investigated using the scalar time series derived from atomistic MD simulations of the reversible folding of a structured peptide. First, FESST was applied to the time series of a single interresidue distance of β-3s, a 20-residue peptide with native three-stranded β-sheet topology. The native state of β-3s, three subbasins in the denatured state, and the free energy barrier for unfolding can be determined with high accuracy. Importantly, FESST is robust to the choice of the residue pair. In fact, 20% of the 154 pairs of $C_\beta - C_\beta$ distances can be used in FESST for

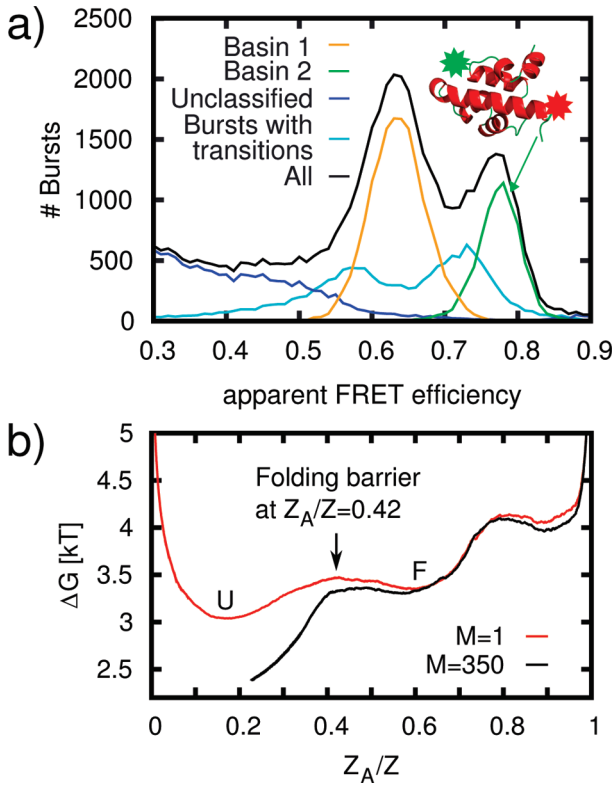

Figure 7. FESST analysis of single-molecule FRET measurements on a monomeric λ-repressor. (a) Histogram of apparent FRET efficiency of all identified bursts (black). The histograms of the apparent FRET efficiency for bursts containing only bins attributed by FESST to the first (unfolded) and the second (folded) basin are shown in orange and green, respectively. Those bursts not assigned to basins 1 or 2 are shown in blue. Because of averaging, the apparent FRET efficiency of the bursts with transitions between the FESST basins cumulates between the mean FRET efficiency of the other FESST basins (cyan curve). The inset shows the structure of the folded λ-repressor fragment with the FRET labeling sites. For the FESST coarse-graining, a window size $N = 25$ and an acceptance cutoff $\zeta = 0.4$ are used (the effect of other coarse-graining parameters is shown in Figure S15 of the SI). (b) One-dimensional projection of the free energy surface of λ-repressor. The cFEP is plotted using the most populated node in the FESST-ETN as the reference, in this case a representative of the unfolded basin (U). The location of the barrier for folding is indicated (black arrow). The highest barrier for folding is found when $M = 350$ nodes in the U basin are merged (black). The second barrier at $Z_A/Z = 0.8$ originates from the transitions to the bleached state.

determining the native state of β-3s, and in particular distances between residues in β-strands 1 and 3 are optimal. Furthermore, the basin assignment by FESST is robust to changes of the parameters used for coarse-graining, which can be determined self-consistently.

In a second test, FESST was applied to a time series of FRET efficiencies generated from the MD trajectory. An accurate identification of the native basin of β-3s is possible with FRET efficiencies calculated from about 1000 photons emitted during the folding time.

A first application to single-molecule FRET experiments on a freely diffusing monomeric λ-repressor with folding dynamics in the millisecond range shows that FESST is able to correctly identify the folded and unfolded subpopulations and yields a free energy profile that captures this separation. This result clearly demonstrates the feasibility of applying FESST to experimental data. However, the height of the folding barrier in the corresponding free energy profile is lower than expected, an effect that is presumably dominated by the current limitations in photon rates. Recent developments in the use of additives that reduce photobleaching and increase fluorescence emission rates[37−39] are expected to contribute strongly to an improvement of this situation both for experiments on freely diffusing and immobilized molecules.

The present analysis focused on the FRET efficiency, because it is one of the most commonly used observables. Additional information, for example, interphoton times, polarization, or fluorescence lifetimes, is expected to further increase the discriminatory power of FESST. In conclusion, FESST can be applied to the time series of any type of scalar observable as long as the short-time distribution of the single-molecule signal contains enough information to allow FESST to remove the signal's degeneracy.

Acknowledgment. We thank Dr. D. Nettels for help with the data analysis and Drs. S. Muff, I. Gopich, D. Nettels, and S. V. Krivov for interesting discussions. This work was supported by grants of the Swiss National Science Foundation to A.C. and B.S. and a Starting Investigator Grant of the European Research Council (FP7) to B.S. Most of the simulations were carried out on the Matterhorn computer cluster and on the Schrödinger computer cluster of the University of Zurich.

Supporting Information Available: FESST applications to atomistic simulations of β-sheet folding, to an emulated FRET signal, and to real experimental data and the FESST resolution limit. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Goldstein, M. *J. Chem. Phys.* **1969**, *51*, 3728–3739.
(2) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598–1603.
(3) Dill, K. A.; Chan, H. S. *Nat. Struct. Biol.* **1997**, *4*, 10–19.
(4) Shakhnovich, E. I. *Phys. Rev. Lett.* **1994**, *72*, 3907–3910.
(5) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.
(6) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299–306.
(7) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.
(8) Caflisch, A. *Curr. Opin. Struct. Biol.* **2006**, *16*, 71–78.
(9) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154–162.
(10) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.
(11) Bai, C.; Wang, C.; Xie, X. S.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11075–11076.
(12) Ha, T.; Enderle, T.; Ogletree, D. F.; Chemla, D. S.; Selvin, P. R.; Weiss, S. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 6264–6268.
(13) Deniz, A. A.; Laurence, T. A.; Beligere, G. S.; Dahan, M.; Martin, A. B.; Chemla, D. S.; Dawson, P. E.; Schultz, P. G.; Weiss, S. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5179–5184.
(14) Schuler, B.; Lipman, E. A.; Eaton, W. A. *Nature (London)* **2002**, *419*, 743–747.
(15) Haran, G. *J. Phys.: Condens. Matter* **2003**, *15*, R1291–R1317.
(16) Nettels, D.; Gopich, I. V.; Hoffmann, A.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 2655–2660.
(17) Schuler, B.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 16–26.
(18) Li, C.-B.; Yang, H.; Komatsuzaki, T. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 536–541.
(19) Baba, A.; Komatsuzaki, T. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19297–19302.
(20) Ferrara, P.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10780–10785.
(21) Muff, S.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1185–1195.
(22) De Alba, E.; Santoro, J.; Rico, M.; Jiménez, M. A. *Protein Sci.* **1999**, *8*, 854–865.
(23) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 24–33.
(24) Muff, S.; Caflisch, A. *J. Phys. Chem. B* **2009**, *113*, 3218–3226.
(25) Cavalli, A.; Haberthür, U.; Paci, E.; Caflisch, A. *Protein Sci.* **2003**, *12*, 1801–1803.

(26) Hartigan, J. *Clustering Algorithms*; Wiley: New York, 1975.
(27) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701–8714.
(28) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13841–13846.
(29) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. *Bioinformatics* **2007**, *23*, 2625–2627.
(30) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
(31) Smirnov, N. V. *Mat. Sb.* **1939**, *6*, 3–24.
(32) Vershik, A. *J. Math. Sci.* **2006**, *133*, 1410–1417.
(33) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.
(34) Huang, G. S.; Oas, T. G. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6878–6882.
(35) Yang, W. Y.; Gruebele, M. *Nature (London)* **2003**, *423*, 193–197.
(36) Ghaemmaghami, S.; Word, J. M.; Burton, R. E.; Richardson, J. S.; Oas, T. G. *Biochemistry* **1998**, *37*, 9179–9185.
(37) Rasnik, I.; McKinney, S. A.; Ha, T. *Nat. Methods* **2006**, *3*, 891–893.
(38) Widengren, J.; Chmyrov, A.; Eggeling, C.; Löfdahl, P.-A.; Seidel, C. A. M. *J. Phys. Chem. A* **2007**, *111*, 429–440.
(39) Vogelsang, J.; Kasper, R.; Steinhauer, C.; Person, B.; Heilemann, M.; Sauer, M.; Tinnefeld, P. *Angew. Chem., Int. Ed.* **2008**, *47*, 5465–5469.
(40) Schuler, B. *Methods Mol. Biol.* **2007**, *350*, 115–138.

JP1053698

Supporting Information

Contents

List of Figures

1 FESST application to atomistic simulations of β-sheet folding

Most of the data presented in the first subsection of this supporting information (SI) refer to the time series of the single distance C_β Gln_4−C_β Thr_{16} in the Beta3s peptide whose time series was generated by atomistic simulations (cf. Sec. 1.1). Robustness tests are presented in subsections 1.2-1.5.

1.1 Molecular dynamics (MD) simulations

Beta3s is a designed 20-residue peptide (Thr_1-Trp_2-Ile_3-Gln_4-Asn_5-Gly_6-Ser_7-Thr_8-Lys_9-Trp_{10}-Tyr_{11}-Gln_{12}-Asn_{13}-Gly_{14}-Ser_{15}-Thr_{16}-Lys_{17}-Ile_{18}-Tyr_{19}-Thr_{20}) that folds to a three-stranded anti-parallel β-sheet [1, 2]. Multiple folding and unfolding events have been sampled by molecular dynamics (MD) simulations [3, 4] with an implicit solvent model [5]. In these MD simulations, Beta3s was modeled by explicitly considering all heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms (PARAM19 force field [3, 6] with the default cutoff of 7.5 Å for the nonbonding interactions). A mean field approximation based on the solvent accessible surface (SAS) was used to describe the main effects of the aqueous solvent [5]. More explicitly, the screening of the electrostatic interactions is approximated by the distance-dependent dielectric function $\epsilon(r) = 2r$, while the remaining solvation effects are approximated by replacement of the monopole moment of charged groups by strong dipole moments and a linear function of atomic SAS values. The latter requires only two surface-tension like parameters and takes into account both polar and apolar solvation effects by a negative (i.e., favorable) value of the surface-tension parameter for nitrogen and oxygen atoms, and a positive (unfavorable) value for carbon and sulfur atoms. Ten MD runs of 2 μs each with different initial distributions of velocities were performed with the Berendsen thermostat (coupling constant of 5 ps) at 330 K, which is slightly above the melting temperature of Beta3s [7]. A time step of 2 fs was used and the coordinates were saved every 20 ps for a total of 10^6 MD snapshots. This required three weeks on a 10-CPU cluster.

1.2 Robustness of FESST upon variation of window size and cutoff for coarse-graining

It is important to evaluate the performance of FESST upon changes in the parameters used for coarse-graining. This analysis shows that the determination of the native basin is robust (Fig. S1). The height of the unfolding barrier as determined in the cFEP can be used to find the optimal coarse-graining parameters (Fig. S2).

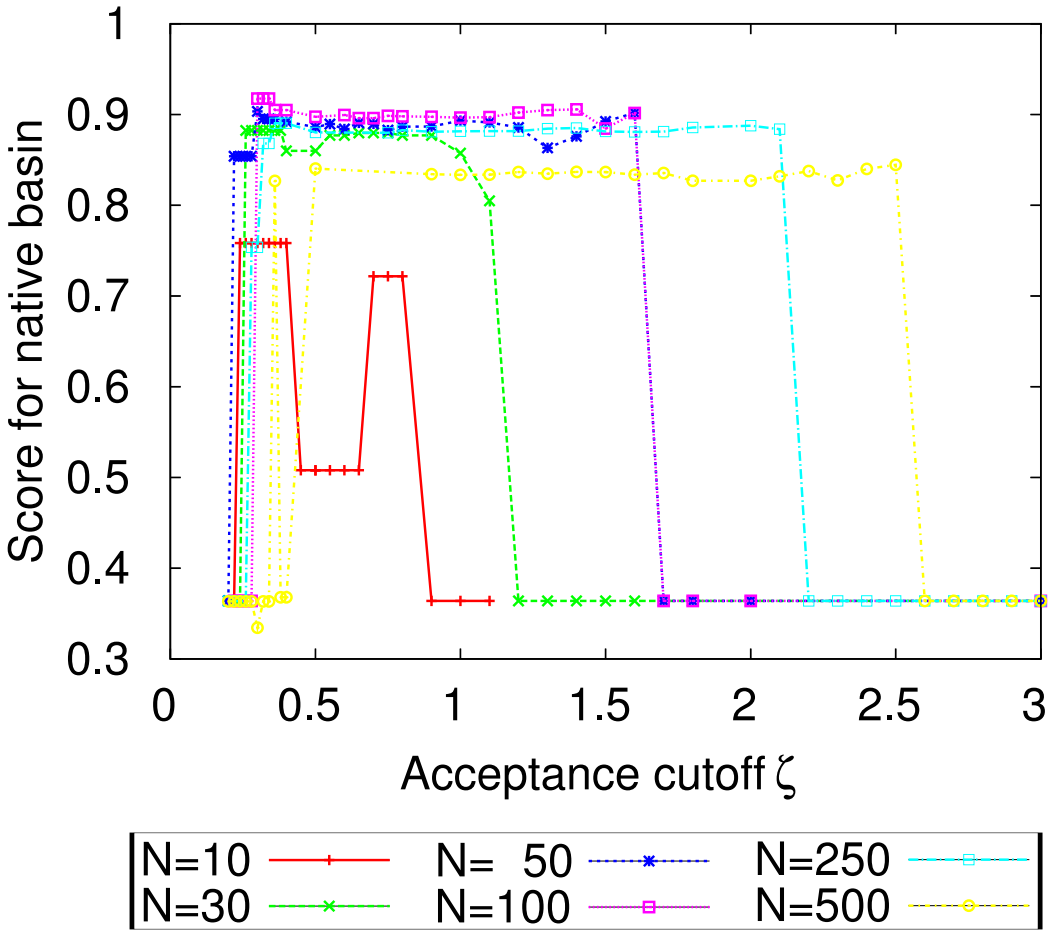

Figure S1: Robustness of FESST upon variation of the parameter used for coarse-graining. The signal is the single distance C_β Gln_4−C_β Thr_{16} of Beta3s obtained by implicit solvent MD. The score is the product of accuracy (i.e., fraction of the FESST native basin belonging to the native state as determined using all 645 coordinates of Beta3s, cf. Fig. S3) and completeness (i.e., fraction of the native state captured by FESST, cf. Fig. S3). The range of values tested for the size of the time window is $10 \leq N \leq 500$, i.e., 0.2 ns $\leq \tau \leq$ 10 ns as the number of MD snapshots N is equal to τ times the saving frequency of 1/20 ps^{-1}. Results in the main text are obtained for values of τ = 2ns (N=100) and $\zeta = 0.3$.

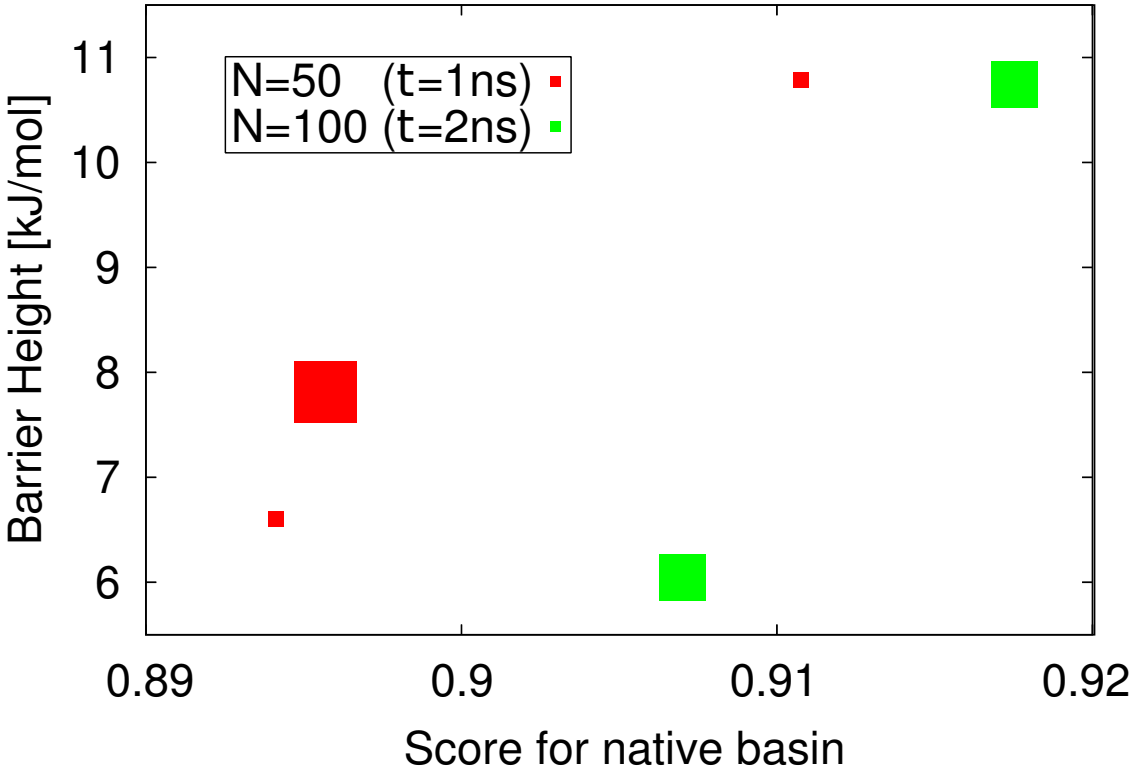

Figure S2: Effect of the parameters used for coarse-graining on the FESST determination of the native state on the height of the unfolding barrier in the cFEP. Note that multiple values of the threshold ζ yield the same score and barrier height. The size of the symbol is proportional to the number of values tested. As an example, the best result, i.e., the data point with highest score *and* barrier height (green square in the top right corner) is obtained with $\zeta = 0.30, \zeta = 0.32,$ and $\zeta = 0.34$ using a window size of N=100. The plot provides evidence that the parameters can be optimized with the height of the cFEP barrier as a cost function.

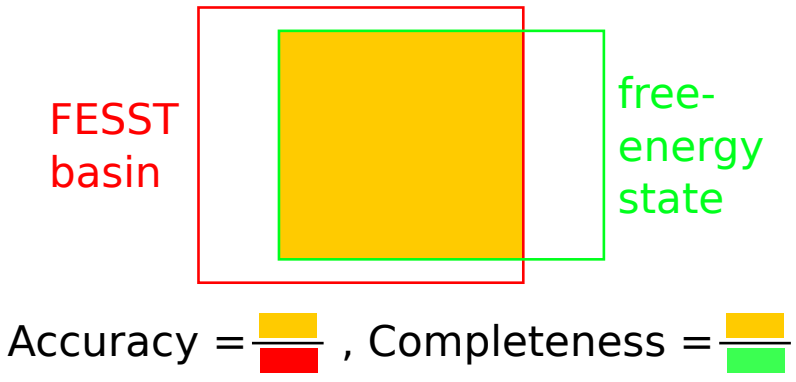

Figure S3: Definition of accuracy and completeness for the free-energy state as determined using the information of all 645 coordinates of Beta3s. The accuracy is the fraction of the FESST basin belonging to the free-energy state as determined using all 645 coordinates of Beta3s, and the completeness is the fraction of the free-energy state captured by FESST.

1.3 Merging of nodes in the native basin

In comparison to the 645-coordinates equilibrium transition network (ETN), the ETN derived from the single distance signal (termed FESST-ETN) lacks nodes with very large weight (Fig. S4). For instance, the most populated node in the latter (158 snapshots) is almost three orders of magnitude smaller than the most visited node of the 645-coordinate ETN (88022 snapshots). To render the most populated node in the FESST-ETN more representative of the native basin, the M heaviest native nodes are combined. The native basin consists of those nodes with a value of the progress variable Z_A/Z in the cFEP (calculated from the most populated node) smaller than the value at the first barrier in the cFEP[8]. The new ETN is constructed from the node sequence in the MD simulation with the heaviest M native nodes merged into one node. The merging step affects only the native basin (inset of Fig. S5), and mainly results in a lower (i.e., more favorable) free-energy value for the bottom of the native basin, i.e., a higher unfolding barrier. The highest value of the unfolding barrier is observed for M = 7000 nodes merged. The value of the barrier height is robust for $3000 \leq M \leq 12000$. For $M \geq 3000$ nodes merged, the weight of the most populated node in the FESST-ETN exceeds the weight of the most visited node in the unmodified 645-coordinates ETN (Fig. S6). As a basis of comparison, the merging procedure can also be applied to the 645-coordinates ETN. The highest barrier is found for only 107 nodes merged and exceeds the value for the unmodified network by only 0.7 kJ/mol (Fig. S5).

Another consequence of the reduced size of the reference node is the overestimation of the time needed to reach the reference node from the other nodes in the FESST-ETN, i.e., a folding time much longer than the one obtained from the 645-coordinates ETN (Fig. S7). For the FESST-ETNs with $M \geq 3000$, the distribution of the mean first passage times matches those of the 645-coordinates ETN (Fig. S7). The correspondence of the folding time distributions provides further evidence that the system dynamics is reliably captured by the FESST coarse-graining. The reliable representation of the system's dynamics in the ETN is a necessary condition for the correct operation of the cFEP approach [8].

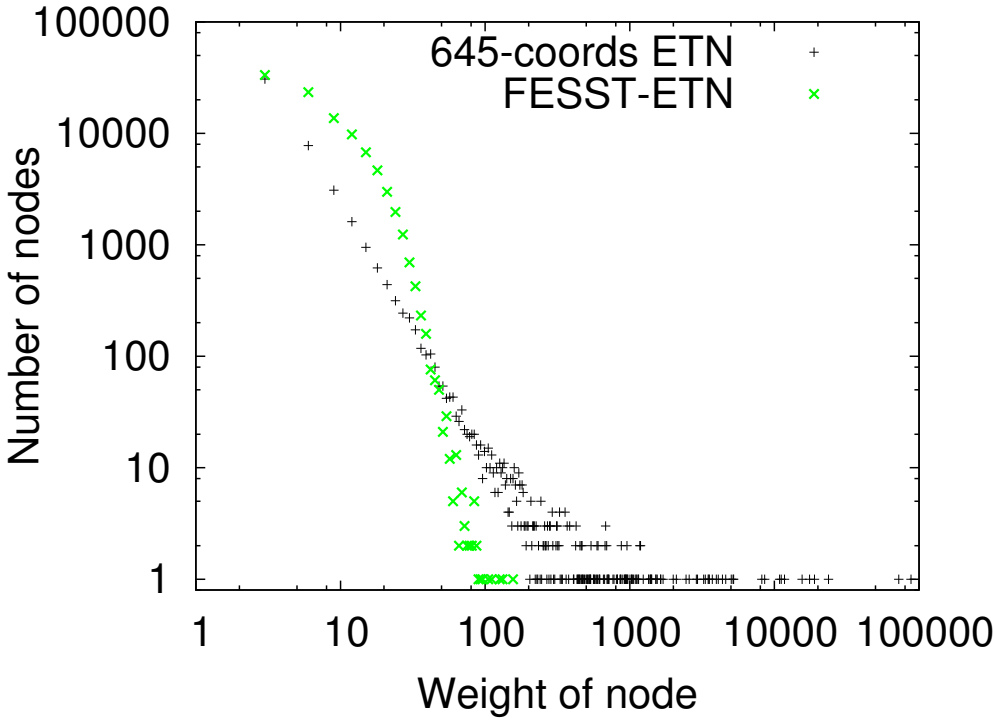

Figure S4: Distribution of node weights (number of snapshots) for the 645-coords ETN and the FESST-ETN. The distance C_β Gln_4−C_β Thr_{16} in Beta3s, window size N = 100, and acceptance cutoff $\zeta = 0.3$ are used. Note that the weight of the most populated node is 88022 and 158 for the 645-coords ETN and the FESST-ETN, respectively.

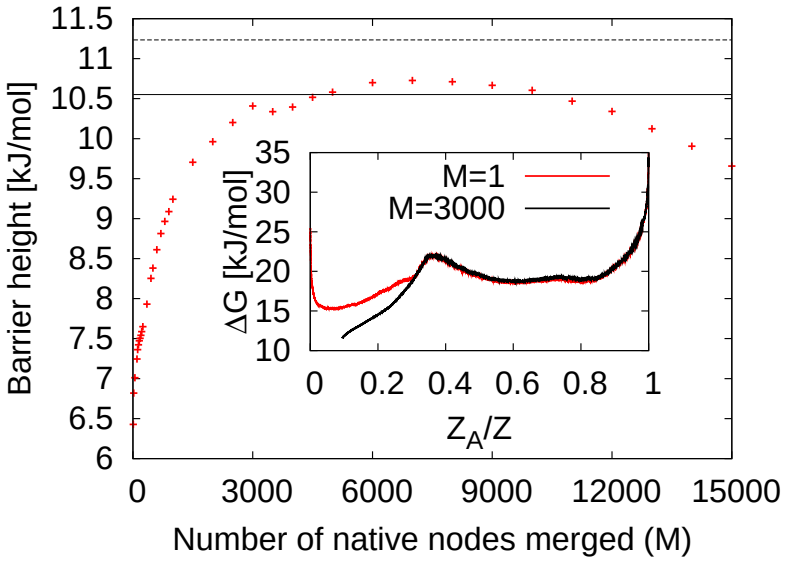

Figure S5: Dependence of barrier height on the merging of the heaviest nodes of the native basin. The red crosses show the barrier height of the FESST-cFEP, i.e. the cFEP of the ETN obtained from the application of FESST to the time series of the single distance C_β Gln_4–C_β Thr_{16} in Beta3s. The solid horizontal line indicates the barrier height of the 645-coords cFEP [9]. The dashed horizontal line displays the maximal barrier height found, which results when the heaviest 107 native nodes in the 645-coords cFEP are merged. The inset shows the FESST cFEPs with M= 3000 native nodes merged (black) and without merging (red).

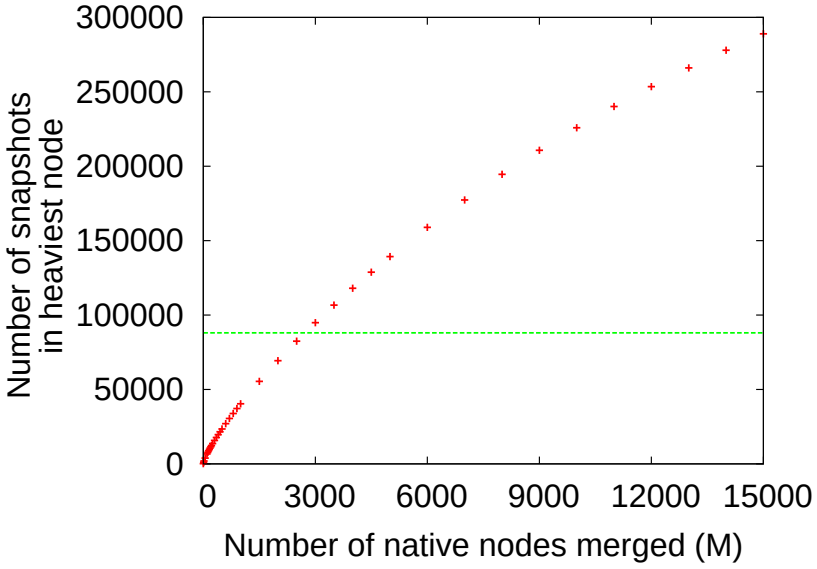

Figure S6: Number of time series bins (which corresponds to snapshots of the MD simulation) in the heaviest node of the FESST-ETN as a function of the number of merged nodes in the native basin. The window size is N = 100 and the acceptance cutoff is $\zeta = 0.3$. The horizontal line indicates the number of snapshots in the heaviest node of the 645-coordinates ETN.

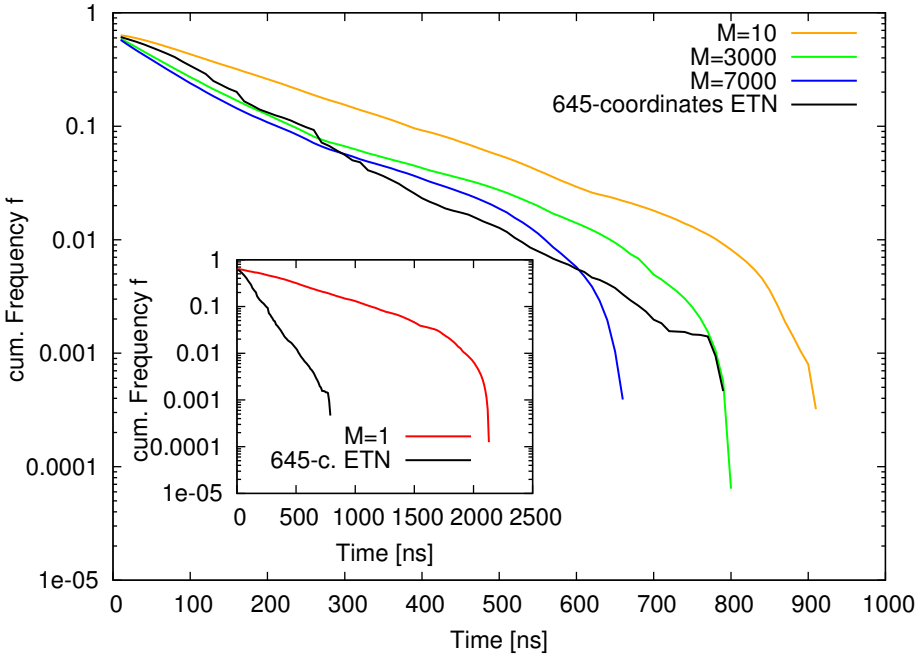

Figure S7: Comparison of folding kinetics for different representatives of the native basin. The plot shows the cumulative distribution of first passage times to the native node, $f(t) = \int_t^\infty p(\tau)d\tau$, where p is the probability distribution of the first passage time. All snapshots were used to calculate $f(t)$. The plot for the 645-coordinates ETN (black lines) is shown as a reference [9]. The inset shows the much slower folding of the FESST-ETN without native node merging (M=1), which is a consequence of the small weight of the most populated node (Fig. S6).

1.4 Comparison of FESST performance for suboptimal intramolecular distances monitored

It is useful to evaluate the performance of FESST for a mediocre and bad separation of the native state peak from the rest of the basins. For this purpose the C_β-C_β distance between residues 1 and 13 (Fig. S8, left) and residues 8 and 18 (Fig. S8, right) are examined.

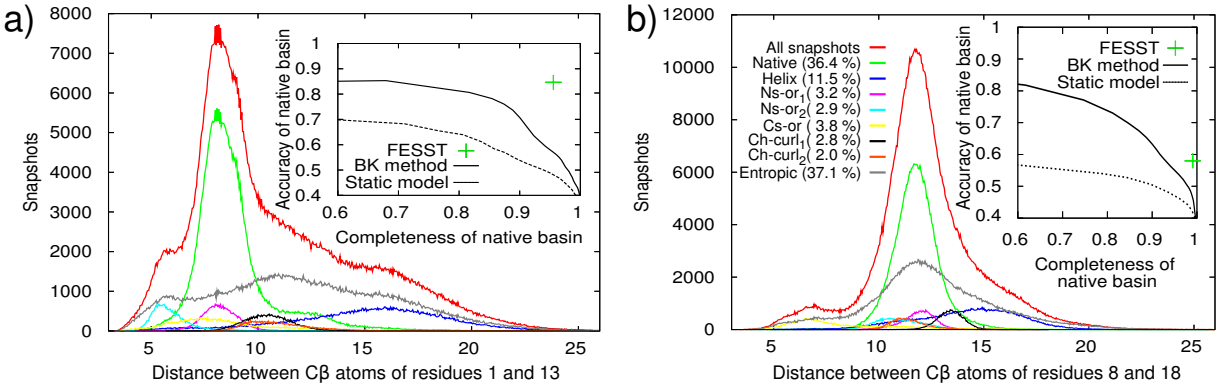

Figure S8: Distribution of inter-residue distances in different free-energy basins as identified using all 645 coordinates of Beta3s. The insets show the accuracy and completeness of the native state detection of FESST (cf. Fig. S3) compared to the BK procedure [10] and static model (described in the main text).

1.5 Choice of the window size for removal of basin overlap

Multiple free-energy basins may overlap on the cFEP because the short-time distribution of the distance is degenerate. To split the basin $\tilde{B}_2$ determined by FESST-cFEP as the first non-native basin (Fig. 2d), the long-time distribution of the distance for each time window $[t_2, t_2 + T]$ (the time bin t_2 belongs to $\tilde{B}_2$) is compared with the distribution of the distance in the entire native basin. The individual sub-basins correspond to individual peaks of the histogram of the comparison metric for a large range of window sizes T (Fig. S9). A value of T=40 ns is used in the main text.

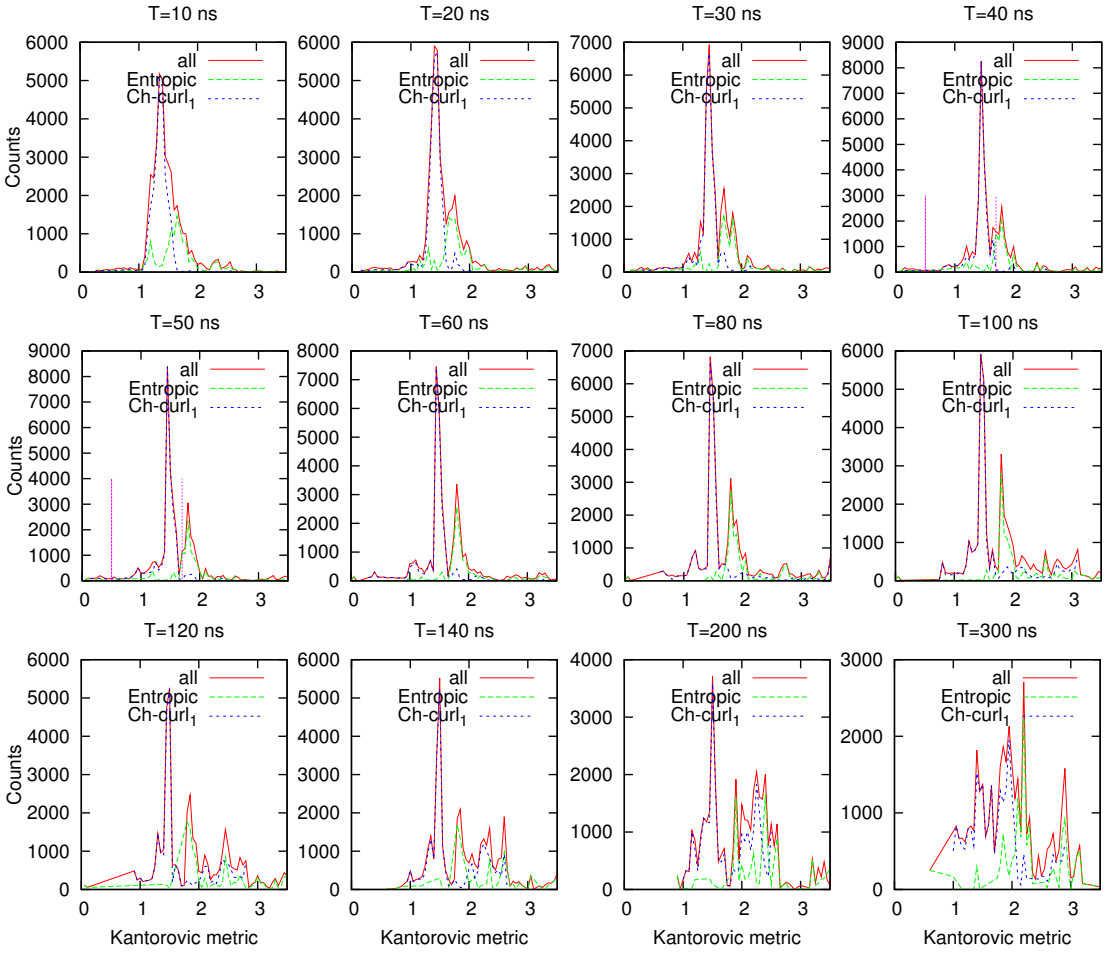

Figure S9: Effect of different window lengths in basin overlap removal. Compared is the Kantorovich metric distribution (area between cumulative histograms) between the long-time distance distributions in windows of varying length T around MD snapshots in $\tilde{B}_2$ and the native distance distribution, i.e., all MD snapshots in B_1 (see Fig. 2 for definition of $B_1, \tilde{B}_2$). This plot illustrates that the ranges of Kantorovich metric are different for the two different subbasins in $\tilde{B}_2$, i.e., Ch-curl$_1$ and Entropic.

1.6 Computational costs

Coarse-graining is the computational bottleneck, and the time it requires depends on the parameters used. Using a variant of the leader algorithm that preserves local kinetics (see main text) and the Kolmogorov-Smirnov test, coarse-graining of 10^6 time windows (of the distance between C_β-atoms of residues 4 and 16 with a window size of N = 100 and a cutoff parameter $\zeta = 0.3$) takes 6 hours on a recent XEON CPU with 2.33 GHz clock frequency. The CPU time depends on the acceptance cutoff. A cutoff of $\zeta = 0.38$ reduces the running time to 4.5 hours. The cFEP calculation takes only five to ten minutes. Very small memory requirements are needed for both procedures. Note that the determination of multiple free energy basins requires only one coarse-graining, but multiple cFEP calculations.

2 FESST application to an emulated FRET signal

To emulate a FRET experiment, the MD-generated time series of the distance between residues 4 and 16 in Beta3s is used together with a Markov state model to generate the photon time series (Fig. 5). The photophysical states considered are DA (both donor and acceptor in ground state), D^*A (donor in excited state, acceptor in ground state), DA^* and D^*A^*. The transition probabilities are approximated by the product of the transition rate and the time step (chosen to be $dt = 0.2$ ps, i.e., 100 observations along the MD saving interval of 20 ps). Finer time steps changed the photon counting results only marginally. For the intrinsic relaxation rates of donor and acceptor, $k_A = k_D = 2500\frac{1}{2\,\mathrm{ns}}$ is used. Direct excitation of the acceptor is set to 5% of the donor excitation rate. The Förster rate $k_F(r) = k_D \left(\frac{R_0}{r}\right)^6$ is calculated from the instantaneous distance r between the C_β-atoms of residues 4 and 16. The Förster radius is $R_0 = 12$ Å, which is the smallest radius that separates the distributions of FRET efficiencies of native and non-native conformations best. This separation is important, because there is an anticorrelation of the score of the native basin and the overlap of the distributions

in native and non-native state (data not shown).

The photon time series from the emulated FRET experiment is divided into bins of size S and the FRET efficiency $E_{\mathrm{FRET}} = \frac{n_{\mathrm{A}}}{n_{\mathrm{A}}+n_{\mathrm{D}}}$ is calculated for each bin (Fig. 5.C), where n_{A} and n_{D} are the number of acceptor and donor photons in the considered bin, respectively. The effect of the number of photons per bin is studied by the variation of the excitation rate (Fig. 6). For comparability with experiments, we report the number of photons emitted during the folding time. To improve the statistics at low emission rates, the photon time series is split into bins of S=0.4 ns, i.e., 20 MD snapshots. To improve sampling, detailed balance is imposed on the equilibrium transition network (ETN), i.e. the weight of each link is set to the average number of transitions in either direction. The score of the native state detection increases with the average emission rate, and reaches a plateau of 85% accuracy and 88% completeness (cf. Fig. S3 for definition) at an emission rate of about 5000 photons per folding time (Fig. 6), compared to 96% accuracy and 95% completeness obtained by applying FESST to the distance time series. The binning time S has little influence on the FESST performance (Fig. S10), whereas the length of the window T (parameter in FESST coarse-graining) shows a significant influence at low excitation rates (Fig. S11).

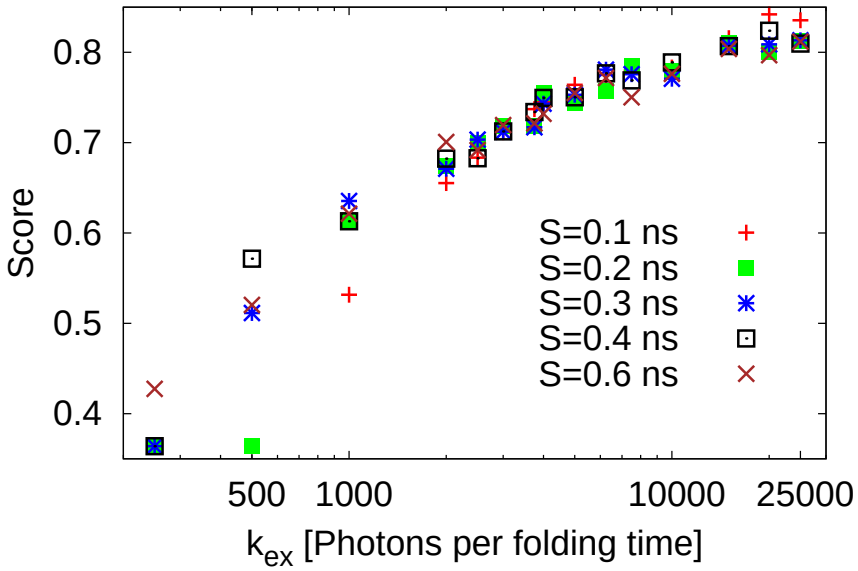

Figure S10: Robustness of the native basin detection in emulated FRET experiments upon change of binning time S. Note that the scores are calculated on a snapshot-wise basin assignment as for Fig. 6. For each excitation rate k_{ex}, only the highest score (tested are window sizes of $2, 4, 6, 8, 10, 20$, and 40 ns) is shown. The excitation rate k_{ex} is expressed as the number of photons emitted per folding time, which is 100 ns for Beta3s at 330 K (Fig. 6).

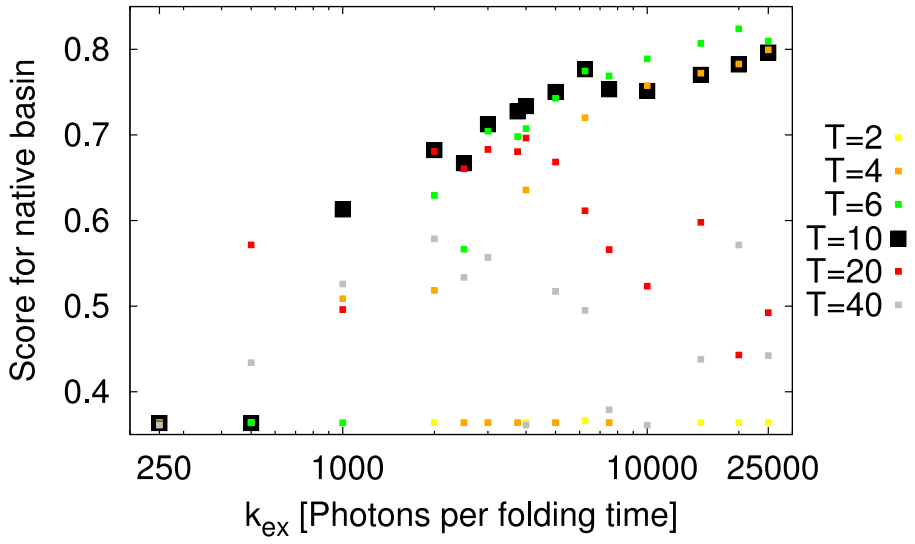

Figure S11: Effect of different window sizes T [ns] on FESST performance in emulated FRET experiments with 0.4 ns bins. The identification of the native basin is robust with respect to the choice of the window size in the range 6 ns $\leq$ T $\leq$ 10 ns. The same setup as for the data shown in Fig. 6 and described in the SI is used. As in Fig. 6, the excitation rate k_{ex} is expressed as the number of photons emitted during the folding time, which is 100 ns for Beta3s at 330 K [11].

3 One-dimensional two-state system: Resolution limit of FESST

It is useful to investigate the resolution limit of FESST using a simple model (Fig. S12). The time evolution of the monitored signal is given by Langevin dynamics of a particle in a one-dimensional potential. To model a two-state system, the potential is switched with a constant rate between two harmonic wells (Fig. S12 a). The sequence of emitted photons is determined by a Gillespie-type simulation [12]. The time series of FRET efficiencies is derived from the binned photon sequence and analyzed by FESST as described for Beta3s with detailed balance imposed. Remarkably, the accuracy of FESST is always higher than the static model even for very small separations of the minima (Fig. S12 b,c,d). FESST can discriminate the minima based on the curvature alone, albeit with a relatively large number of detected photons required (cf. Fig. S12 and Fig. S13). Although the simple model does not reflect the complexity of a multidimensional system, the present

results indicate that the reliable operation of FESST requires the detection of 100 to 1000 photons during the residence time in one free-energy state. In real single-molecule FRET experiments, 100 to 1000 photons can be detected in about one to ten milliseconds. Accordingly, we expect FESST to be a suitable approach for determining the properties of free energy surfaces of molecules that exhibit dynamics in this range or slower, thus covering a large part of the biologically important time scales [13, 14, 15, 16].

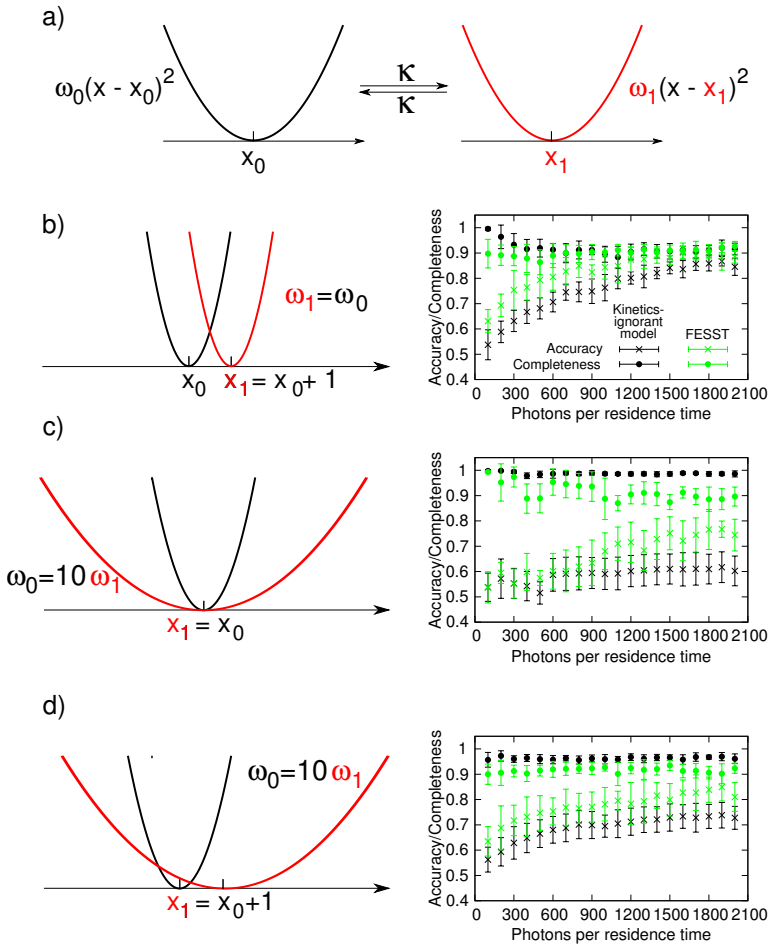

Figure S12: Resolution limits of FESST examined with a one-dimensional two-state model. (a) Schematic illustration of the model: Dynamics of a massive particle in a harmonic potential that switches from one shape and position to the other with rate κ. The position time series is then transformed to a time series of FRET efficiencies as for Beta3s (the equilibrium position is defined as $x_0 = 10$ A.U. (arbitrary units), the Förster radius as $R_0 = 12$ A.U., the curvature of potential 0 as $\omega_0 = 100$ A.U., and the interchange rate is assumed to be $\kappa = 0.005$ A.U.). (b-d, left) Illustration of the potential's shape and position. (b-d, right) Accuracy and completeness of the basin with index 0 as detected by FESST and the best static model in ten independent simulations for the potential's setup on the left. The parameters for FESST are optimized using the height of the unfolding barrier determined self-consistently (cf. Fig. 2c). For the static model, the basin is formed by all bins with a FRET efficiency lower than a given cutoff. To make the most stringent comparison, the detection quality of the static model is maximized by finetuning both the length of the binning interval and the cutoff value based on the knowledge of the solution. The shapes and positions of the potentials used to investigate the different scenaria are given in the individual panels: (b) examines the effect of a shift, (c) the effect of a broadened potential with identical equilibrium position, and for (d) both curvatu./Paper/leader-single-molecule/vorschlag-FRET/re and equilibrium position are changed. For a minor shift $x_1 - x_2 = 1$ A.U. of the two minima of the potential (panels (c) and (d)), FESST yields significantly more accurate solutions with only slightly lower completeness than the static model for as few as 200 photons per residence time. An increased amount of about 1000 photons per residence time is required if the equilibrium positions match and the curvatures differ by a factor of ten (panel c).

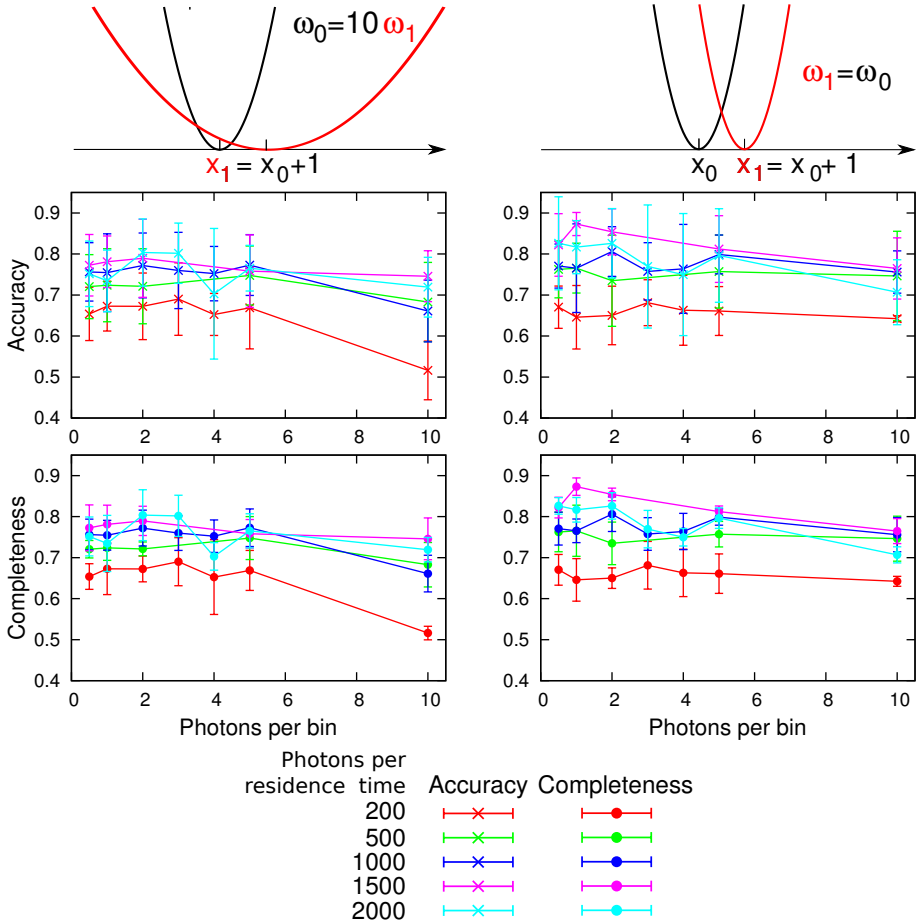

Figure S13: Dependence of FESST performance in the one-dimensional two-state model on the number of photons per FRET bin. The accuracy and completeness of the basin with index 0 (cf. Fig. S12) as detected by FESST depend mainly on the number of photons emitted during the residence time, and only weakly on the explicit number of photons per bin.

4 FESST application to real experimental data (single-molecule FRET on λ-repressor)

4.1 Expression, purification, and labeling of λ-repressor

A plasmid (pET-47b(+)-vector, Novagen) containing the custom-synthesized and codon-optimized sequence coding for the monomeric N-terminal fragment of λ-repressor including an amino terminal hexahistinie tag was purchased from Celtek Genes (Nashville, USA). Threonine 8 and lysine 70 were replaced by cysteine residues for fluorophore labeling. The final amino acid sequence was GPSLCQEQLEDARRLKAIYEKKKNELGLSQESVADKMGMGQSGVGALFNGINALNAYNAALLAKILCVSVEEFSPSIAREIR. The protein was expressed in E.coli BL21 cells at 37°C in LB medium containing kanamycin and 1mM IPTG for induction. The resulting inclusion bodies were harvested by centrifugation after cell lysis with a French pressure cell. Resolubilized protein was subjected to immobilized metal ion affinity chromatography (IMAC; HisTrap H, GE Healthcare) at pH 8. The single peak eluting in the imidazole gradient was collected. The N-terminal His-Tag was cleaved with HRV 3C protease. Uncleaved λ-repressor and protease were separated using IMAC and gel filtration (Superdex 75, GE Healthcare). Labeling was performed essentially as described previously [17]. Purified protein was reacted first with Alexa Fluor 488 maleimide (Invitrogen) at substochiometric concentrations according to the supplier's recommendations. The resulting products were separated using anion exchange chromatography (MonoQ 5/50 GL, GE Healthcare). Singly labeled protein was then concentrated by ultrafiltration (Centricon, Millipore) and reacted with an excess of Alexa Fluor 594 maleimide. Monomeric λ-repressor with one donor and one acceptor dye was purified by anion exchange and size exclusion chromatography. All steps were carried out at high concentrations of denaturant. Correct labeling was confirmed by electrospray ionization mass spectroscopy.

4.2 Single molecule spectroscopy

Measurements were carried out in a MicroTime200 (PicoQuant, Berlin, Germany) with continuous wave laser excitation at 488 nm essentially as described previously [17, 18]. In measurements for the FESST analysis, protein was refolded and diluted from a stock in 8 M guanidinium chloride into a buffer containing 50 mM sodium phosphate buffer pH 7, 0.01% Tween, 150 mM beta-mercaptoethanol, 10 mM cysteamine to a final concentration of guanidinium chloride of 0.68 M and 16 pM of protein. The sample was measured in a temperature-controlled cell [18] with a temperature of 12°C at the position of the confocal volume. The laser power was adjusted to 600 μW. Nanosecond correlation measurements were performed at a protein concentration of 1 nM with a laser power of 30 μW and analyzed as described previously [19].

4.3 Treatment of photon time series with bursts

As freely diffusing molecules are observed in our FRET-experiments, the photon time series is structured in bursts. This signal has to be converted into a time series of nodes by coarse-graining with FESST (Fig. 5.C, S14).

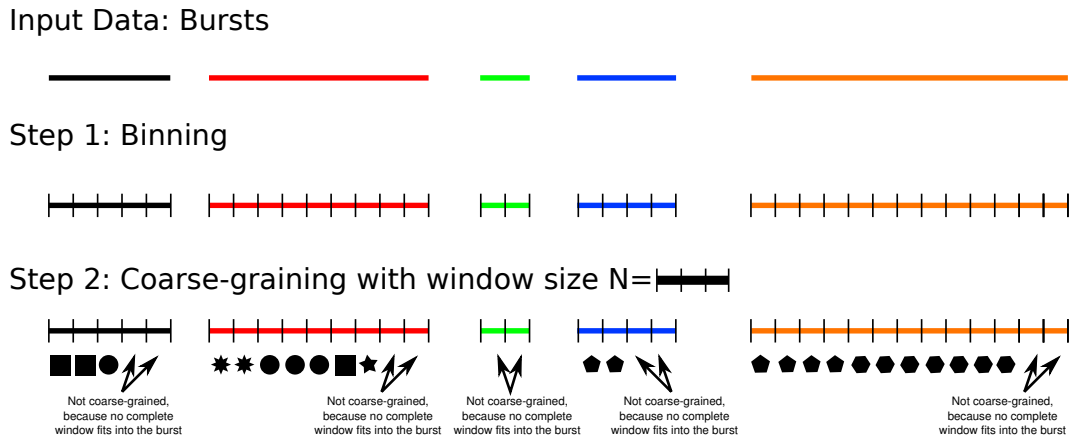

Figure S14: FESST coarse-graining for photon time series from individual bursts. As for a continuous photon time series, the data are first binned (Step 1). In the second step, bins are assigned to a node if their short-time window fits completely in the burst.

4.4 Robustness of FESST upon variation of coarse-graining parameters

This analysis shows that the determination of the folded and unfolded populations/basins is robust for $25 \leq N \leq 30$ (Fig. S15, left) and $0.3 \leq \zeta \leq 0.4$ (Fig. S15, right).

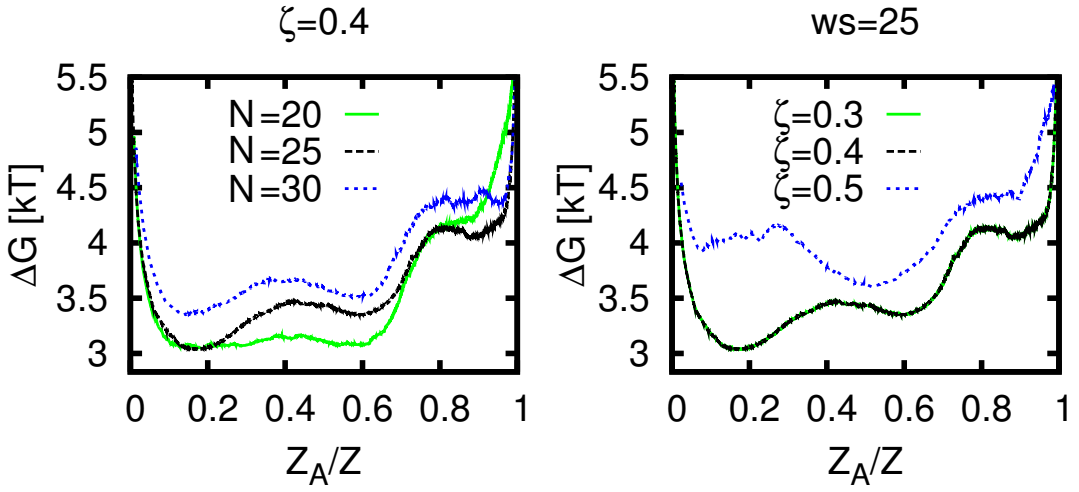

Figure S15: Differences in cut-based free-energy profiles (cFEP) upon changes of the FESST coarse-graining parameters. The input signal consists of a photon time series measured in a FRET experiment of freely diffusing λ-repressor fragments. Bursts are characterised by at most $\delta t = 70\mu$s between two successive photons. FESST is applied to the time series of FRET efficiencies calculated for 0.1 ms bins. The highest barrier resulted for a window size N = 25 bins (which corresponds to a time $\tau = 2.5$ ms) and an acceptance cutoff $\zeta = 0.4$. Note that the $\zeta = 0.3$ (green) and $\zeta = 0.4$ (black) curves overlap fully (right panel), and thus the former is not visible.

4.5 cFEP with the folded state as a reference

It is interesting to compare the cFEP obtained using as a reference the most populated node, which is a representative of the unfolded state (Fig. 7b), with the cFEP from the most populated node in the folded state (Fig. S16). The two cFEPs are consistent. In particular, the (un)folding barrier height and the two main basins are similar.

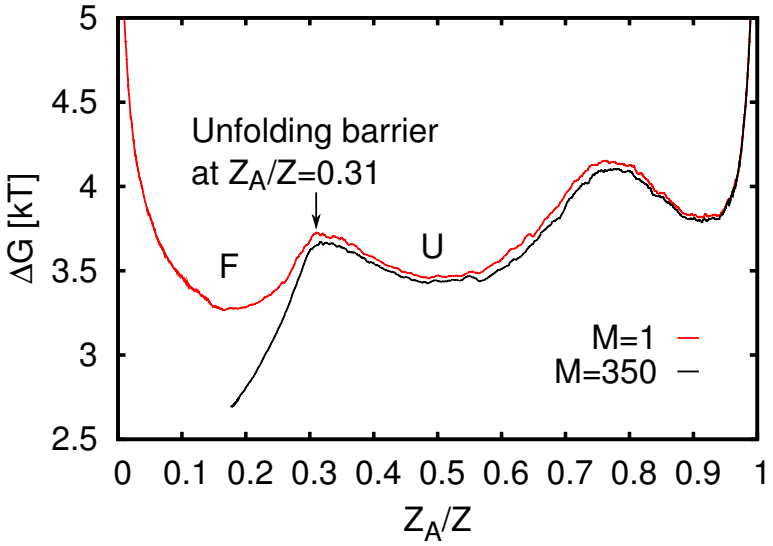

Figure S16: Cut-based free-energy profile (cFEP) from the folded state. The location of the barrier for unfolding is indicated (black arrow). A significantly higher unfolding barrier (black curve) is obtained by merging the M=350 most populated nodes in the folded basin as determined with M=1 (red curve).

4.6 Imposing detailed balance on the ETN of lambda-repressor

For the lambda-repressor, detailed balance (DB) is not imposed. The cFEPs for the ETNs with or without DB differ as shown in S17. The cFEP for the ETN

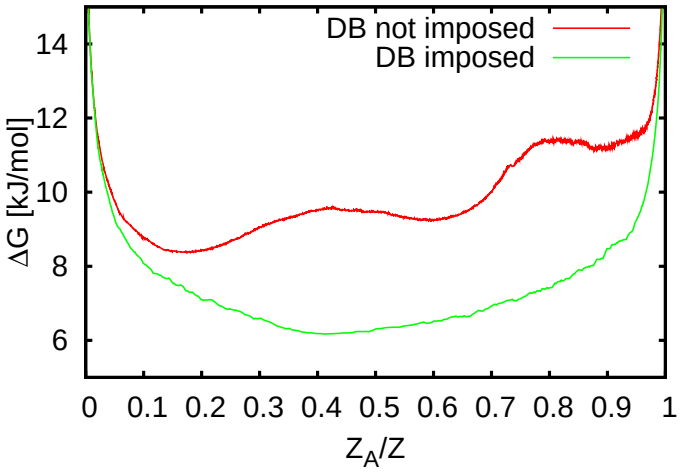

Figure S17: Comparison of the cut-based free-energy profile (cFEP) from the most populated state with and without detailed balance (DB) imposed. The cFEP from the ETN with DB imposed has no barrier in contrast to the cFEP from the unmodified ETN.

with detailed balance imposed shows no barriers. The reason for this observation are spurious transitions observed only in individual bursts. In the ETN, these transitions show up as chains visited only once and only in one direction (a chain is a sequence of nodes $A_1, \ldots, A_n$, where A_i is linked exclusively to A_{i+1} for $i = 1, \ldots, n-1$). This hypothesis is confirmed by comparing the cFEP of the ETN without DB to the cFEP of ETN (cf. Fig. S18) in which on all links but those belonging to a chain visited just once are symmetrised (the weight of the link is set to the average number of transitions in either direction).

As a second test, all links in a unidirectional chain are removed (blue cFEP in S18). The position of the first barrier is equal for all three profiles. The barrier for the cFEP from the ETN with links removed (blue curve in S18) is higher, because spurious transitions are removed. These unidirectional chains are present, because actual states of the system are split into multiple nodes due to shot noise.

As illustrated above, imposing detailed balance on a system such as the lambda-

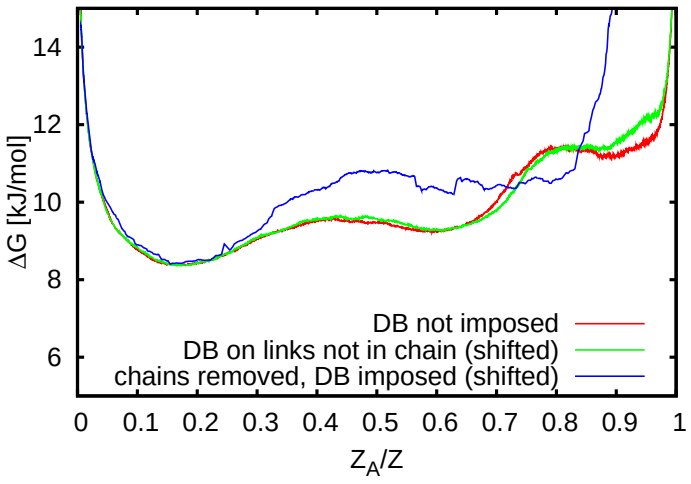

Figure S18: Assessment of the effect of unidirectional chains on the cut-based free-energy profile (cFEP) from the most populated state when detailed balance (DB) is imposed. All links apart from those in a unidirectional chain can be symmetrised (i.e. the weight of the link is set to the average number of transitions in either direction) without any change on the cFEP. Removing the chains from the cFEP leaves the position of the barrier invariant. The height of the barrier increases, because spurious transitions are removed. Two of the three cFEPs were shifted along the y-axis to bring the bottom of the first basin on the left to the same reference value.

repressor leads to a different ETN. A random walker on a unidirectional chain walks from start to end with shortest number of steps. If a chain is symmetrised, the random walker diffuses for artificially long times along the symmetrised chain and might even return to its beginning, which is a prediction in total contradiction to the experimental results. However, if detailed balance is not imposed, the values for the flux between the large nodes (number of transitions between them) are correct.

References

[1] De Alba E, Santoro J, Rico M, Jiménez MA (1999) De novo design of a monomeric three-stranded antiparallel β-sheet. *Protein Science* 8:854–865.

[2] Ferrara P, Caflisch A (2000) Folding simulations of a three-stranded antiparallel β-sheet peptide. *Proc. Natl. Acad. Sci. USA.* 97:10780–10785.

[3] Brooks BR, et al. (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.

[4] Brooks BR, et al. (2009) CHARMM: The biomolecular simulation program. *J. Comput. Chem.* 30:1545–1614.

[5] Ferrara P, Apostolakis J, Caflisch A (2002) Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics* 46: 24–33.

[6] Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. *J. Chem. Phys.* 105:1902–1921.

[7] Cavalli A, Ferrara P, Caflisch A (2002) Weak temperature dependence of the free energy surface and folding pathways of structured peptides. *Proteins: Structure, Function, and Bioinformatics* 47: 305–314.

[8] Krivov SV, Karplus M (2006) One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. *J. Phys. Chem. B* 110:12689–12698.

[9] Krivov SV, Muff S, Caflisch A, Karplus M (2008) One-Dimensional Barrier Preserving Free-Energy Projections of a beta-sheet Miniprotein: New Insights into the Folding Process. *J. Phys. Chem. B* 112:8701–8714.

[10] Baba A, Komatsuzaki T (2007) Construction of effective free energy landscape from single-molecule time series. *Proc. Natl. Acad. Sci. USA.* 104:19297–19302.

[11] Muff S, Caflisch A (2008) Kinetic analysis of molecular dynamics simulationsa reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. *Proteins: Structure, Function, and Bioinformatics* 70: 1185–1195.

[12] Gopich IV, Szabo A (2009) Decoding the Pattern of Photon Colors in Single-Molecule FRET. *J. Phys. Chem. B* 113:10965–10973.

[13] Joo C, et al. (2006) Real-Time Observation of RecA Filament Dynamics with Single Monomer Resolution. *Cell* 126:515–527.

[14] Joo C, Balci H, Ishitsuka Y, Buranachai C, Ha T (2008) Advances in Single-Molecule Fluorescence Methods for Molecular Biology. *Annual Review of Biochemistry* 77:51–76.

[15] Borgia A, Williams PM, Clarke J (2008) Single-Molecule Studies of Protein Folding. *Annual Review of Biochemistry* 77:101–125.

[16] Chung HS, Louis JM, Eaton WA (2009) Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc. Natl. Acad. Sci. USA.* 106:11837–11844.

[17] Hoffmann A, Kane A, Nettels D, Hertzog DE, Baumgärtel P, Lengefeld J, Reichardt G, Horsley DA, Seckler R, Bakajin O, Schuler B (2007) Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proc. Natl. Acad. Sci. USA.* 104:105–110.

[18] Nettels D, Müller-Späth S, Küster F, Hofmann H, Haenni D, Rüegger S, Reymond L, Hoffmann A, Kubelka J, Heinz B, Gast K, Best RB, Schuler B (2009) Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc. Natl. Acad. Sci. USA.* 106:20740-20745.

[19] Nettels D, Gopich IV, Hoffmann A, Schuler B (2007) Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci. USA.* 104:2655–2660.