Deciphering hierarchical features in the energy landscape of adenylate kinase folding/unfolding

J. Nicholas Taylor, Menahem Pirchi, Gilad Haran, and Tamiki Komatsuzaki

Citation: The Journal of Chemical Physics **148**, 123325 (2018); doi: 10.1063/1.5016487
View online: https://doi.org/10.1063/1.5016487
View Table of Contents: http://aip.scitation.org/toc/jcp/148/12
Published by the American Institute of Physics

Deciphering hierarchical features in the energy landscape of adenylate kinase folding/unfolding

J. Nicholas Taylor,[1] Menahem Pirchi,[2] Gilad Haran,[2] and Tamiki Komatsuzaki[1,3]
[1]*Research Institute for Electronic Science, Hokkaido University, Kita 20 Nishi 10, Kita-Ku, Sapporo 001-0020, Japan*
[2]*Weizmann Institute of Science, Rehovot 76100, Israel*
[3]*Graduate School of Life Science, Hokkaido University, Sapporo 001-0020, Japan*

(Received 20 November 2017; accepted 5 January 2018; published online 24 January 2018)

Hierarchical features of the energy landscape of the folding/unfolding behavior of adenylate kinase, including its dependence on denaturant concentration, are elucidated in terms of single-molecule fluorescence resonance energy transfer (smFRET) measurements in which the proteins are encapsulated in a lipid vesicle. The core in constructing the energy landscape from single-molecule time-series across different denaturant concentrations is the application of rate-distortion theory (RDT), which naturally considers the effects of measurement noise and sampling error, in combination with change-point detection and the quantification of the FRET efficiency-dependent photobleaching behavior. Energy landscapes are constructed as a function of observation time scale, revealing multiple partially folded conformations at small time scales that are situated in a superbasin. As the time scale increases, these denatured states merge into a single basin, demonstrating the coarse-graining of the energy landscape as observation time increases. Because the photobleaching time scale is dependent on the conformational state of the protein, possible nonequilibrium features are discussed, and a statistical test for violation of the detailed balance condition is developed based on the state sequences arising from the RDT framework. *Published by AIP Publishing.* https://doi.org/10.1063/1.5016487

I. INTRODUCTION

How one can decipher the underlying model from noisy, finite single-molecule time-series such as those arising from single-molecule fluorescence resonance energy transfer (smFRET) experiments? Since the inception of single-molecule observations, characterization of underlying states and the determination of the kinetic properties of them, such as interconversion rates and pathways, have been at the forefront in the analysis of single-molecule experiments. Many different approaches[1] have utilized the hidden Markov model (HMM) to enable the inference of molecular mechanisms. The HMM has been applied to photon arrival trajectories[2] and smFRET trajectories that are binned[3] or acquired photon-by-photon,[4] allowing for biomolecular elucidations such as the characterization of folding landscape of adenylate kinase (AK) at different concentrations of the denaturant guanidinium chloride (GdmCl).[5] Other HMM constructions applied to single-molecule time-series include variational Bayes formulations[6,7] and the more recent infinite HMM formulations.[8,9] Alternative approaches to infer the state-space network along time-series include the epsilon machine of computational mechanics, which groups past sequences of states in order to predict future sequences of states, and can account for non-Markovian behavior in the systems.[10,11]

Other approaches to state characterization include those that are data-driven, using statistics and unsupervised learning methods to allow the states to emerge from the data rather than imposing a predetermined model, as with the HMM. For example, the time-series can be divided into segments of uniform length and then clustered to produce a set of local equilibrium states[12,13] along the time-series, from which a representation of the free energy landscape can be constructed. Other methods identify step transitions, or change-points, in the trajectories, and subsequently group the intervals between them to produce states.[14] Finally, a method using the uniform segmentation approach[12] and utilizing the information theoretical rate-distortion theory[15] (RDT) as an unsupervised learning method has incorporated the quantification of empirical and finite sampling errors.[16] This allows variations in error magnitude across multiple trajectories to be considered in extracting a series of the states from single-molecule time-series as well as in performing model selection.[16]

Some experiments investigate a system across multiple experimental conditions, such as those following the folding/unfolding behavior of a protein across varying temperatures or denaturant concentrations. In these situations, it is often desirable that a single, consistent model be extracted across the multiple sets of acquired data. For example, in their smFRET study of the AK folding landscape, Pirchi *et al.*[5] used a combination of change-point detection and a HMM to extract a set of states across five denaturant concentrations. This was accomplished by extracting the states from a single denaturant concentration and imposing them on the trajectories acquired at other concentrations. Because the measurements consist of the same protein (AK) presumably experiencing a similar folding landscape across the conditions, it is a natural

assumption that a consistent set of states should be observed at each condition, perhaps with different residential and transition probabilities.

From a data-driven viewpoint, the states can be allowed to emerge from the data sets, and their consistency assessed as a validation that the states are truly uniform across the conditions. One pitfall in such an approach, however, is the variation of the error magnitude, e.g., those arising from experimental sources like photon counting noise as well as finite sampling effects, across the different experiments. Because the RDT method[16] relies heavily on the quantification of errors, this pitfall can be avoided and the error properties integrated directly into the procedure. The original scheme to extract states from smFRET time-series using RDT requires the segmentation of the time-series with uniform time windows of length τ. Time-dependent segment distributions are calculated, becoming objects to be clustered with RDT clustering.[16] Because the RDT method precisely quantifies errors, all trajectory segments from all data sets can be clustered simultaneously regardless of their origin, thus yielding a set of states that is fit globally across the multiple data sets. Furthermore, this approach provides an internal validation as to whether the hypothesis that the same model occurs across the multiple data sets; if the properties of the state distributions vary wildly from condition to condition, then the single model hypothesis is invalid.

The original construction[12,13,16] defined the extracted set of states to be "local equilibrium states" when the time window τ is long enough to attain local equilibration. The time window τ is interpreted to be a time scale of observation such that relatively small τ captures microstates, intermediate τ captures the unification of the microstates into basins or states on the energy landscape, and relatively large τ captures the unification of basins into superbasins that are comprised by multiple states,[17] thus providing a means to decipher the hierarchical features of the energy landscape. In practice, not only is the identification of an appropriate set of time windows still an open question but segments constructed from uniform segmentation of the time-series can contain transitions, which may lead to nonphysical artifacts in the analysis of the set of states.

Another issue that may arise is the influence of the photostability of the fluorophores. Specifically, if one fluorophore is more photostable than the other in a smFRET experiment, then the photobleaching rate will be dependent on the FRET efficiency and affected by the conformational state that the biomolecule occupies. For example, after excluding the effects of fluorophore photodynamics, e.g., photoblinking due to occupation of the triplet excited state, through extensive filtration of trajectories containing this behavior, Pirchi *et al.*[5] introduced a photobleaching state into the HMM scheme in their smFRET study of AK to account for a less photostable donor fluorophore and faster expected photobleaching from low efficiency states. To properly elucidate state properties, it is thus very crucial to take into account the dependence of photobleaching, or more generally, trajectory termination, on individual conformational states.

To address these issues, we merge the original RDT soft clustering scheme for smFRET analysis with change-point detection[14,18–21] using the intervals between detected change-points as non-uniform time windows τ_i. Because change-points are likely to arise from conformational transitions, compared to the uniform time window approach, non-uniform time intervals τ_i are regarded as being inherent to the acquired data. The benefit of implementing change-point detection is minimizing the number of time segments containing transitions. We also introduce a photobleaching state so that the photobleaching behavior of states can be analyzed. We apply the method to the sets of trajectories obtained in the aforementioned AK folding experiments.[5] A minimum of four states can be extracted globally from these data sets, and we verify that the distributions of all the extracted states with respect to FRET efficiency are consistent across all the experimental conditions. Escape times from the states indicate a time scale separation for transitions among unfolded states and between unfolded and folded states, suggesting that the unfolded states lie in a denatured superbasin consisting of several energy minima. Assessment of the transition network for the detailed balance condition reveals violations that may arise due to the occurrence of photobleaching. Construction of approximate free energy landscapes[22,23] at multiple time scales reveals the hierarchical features of energy landscape of AK to follow the superbasin arrangement suggested by the escape times, with escape kinetics from the unfolded superbasin controlled by three smaller basins, whereas the escape kinetics from the folded state are dominated by one basin.

In Sec. II, we present our methods—change-point detection and global model extraction—in the framework of the RDT method. In Sec. III, we apply it to AK folding/unfolding smFRET data and offer conclusion and discussion in Sec. IV.

II. THEORY AND METHODS

A. Soft clustering combined with change-point detection

Rate-distortion theory (RDT), based on information theory,[15,24] provides a soft clustering algorithm in which the conditional probabilities $p(\boldsymbol{S}|\boldsymbol{g})$ of a set of n states $\boldsymbol{S} = \{S_1,\ldots,S_n\}$ given the observation of the set of N segments $\boldsymbol{g} = \{g_1, \ldots, g_N\}$ are returned through an iterative procedure.[25,26] Schematically, soft clustering combined with change-point detection is represented in Fig. 1. First, a smFRET time-series is decomposed into a series of disjoint subsequences with non-uniform time windows τ_i using a change-point detection algorithm [Fig. 1(a)].[18] For each interval τ_i, a probability mass function $g_i(E)$ is computed. Note that change-point analysis usually involves type I (false positive) and type II (false negative) errors; the former is the probability to assign a change-point although it does not occur, and the latter is the probability of not assigning a change-point that does occur. For instance, a location in which type I error occurs is indicated by a red arrow in Fig. 1(a). Segment distributions $g_i(E)$ are calculated from each of the segments g_i, i.e., intervals between change-point locations, and then are used to calculate pairwise distances among segments as the Kantorovich metric.[27] Pairwise distances in

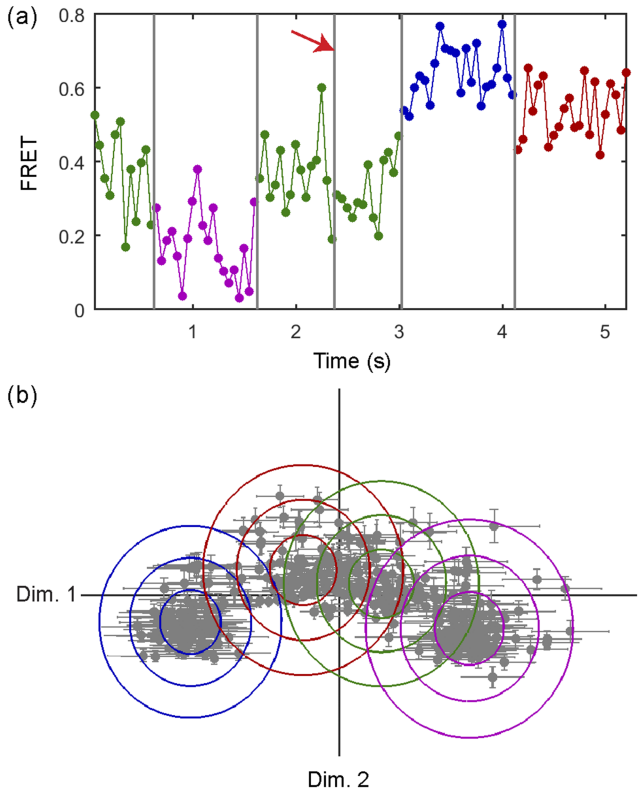

FIG. 1. Schematic representation of soft clustering combined change-point detection. (a) A smFRET time series with non-uniform time windows τ_i. Thin vertical lines indicate the location of detected change points, including one of type I error (indicated by an arrow). (b) A two-dimensional projection of a set of segments of the observable E, $\boldsymbol{g}(E) = \{g_1(E), \ldots, g_i(E), \ldots, g_N(E)\}$, located in a high-dimensional space (Kantorovich metric). Error bars denote finite sampling errors and experimental errors associated with each $g_i(E)$ in the metric space. Concentric circles indicate a set of possible clusters, where some $g_i(E)$ can belong to several clusters simultaneously.

the high-dimensional metric space are then fed into the RDT algorithm. Here we use the Kantorovich metric because it offers superior performance to the other metrics such as the relative entropy and Hellinger distance.[12] Figure 1(b) illustrates a two-dimensional projection of segment distributions $\boldsymbol{g}(E)$ in the high-dimensional space, where errors in measurements are taken into account for each $g_i(E)$. Soft clustering allows some of the $g_i(E)$ to belong to more than single cluster S_k simultaneously with probability $p(S_k|g_i)$. This is advantageous when the error magnitude is comparable to the distances among $g_i(E)$ in comparison to conventional (hard) clustering in which $g_i(E)$ are assigned to only one state. The conditionals $p(S_k|g_i)$ can be interpreted as the certainty to which the segment distribution is assigned to state S_k. Here it should also be noted that the observed type I error in change-point detection, i.e., extraneous change-points, is also improved in the clustering procedure.

The RDT algorithm is the minimization of a functional $\mathcal{F}$ defined by

$$\mathcal{F} = I(\boldsymbol{S}, \boldsymbol{g}) + \beta \langle D \rangle. \tag{1}$$

Here $I(\boldsymbol{S}, \boldsymbol{g})$ is known as the *rate*, i.e., the average amount of information needed to specify states S_k with segments g_i and vice versa. $I(\boldsymbol{S}, \boldsymbol{g})$ is computed as the average mutual information between the set of states $\boldsymbol{S}$ and the set of segments $\boldsymbol{g}$. $\langle D \rangle$ is the mean *distortion* among segments, the average of the pairwise distortions between all pairs of segments within the set of states $\boldsymbol{S}$. The parameter β controls the ratio between the rate and the distortion in the minimization of $\mathcal{F}$.

$I(\boldsymbol{S}, \boldsymbol{g})$ and $\langle D \rangle$ are represented as follows:

$$I(\boldsymbol{S}; \boldsymbol{g}) = \sum_{k=1}^{n} \sum_{i=1}^{N} p(S_k|g_i)\, p(g_i) \log_2 \frac{p(S_k|g_i)}{p(S_k)}, \tag{2}$$

$$\langle D \rangle = \sum_{k=1}^{n} p(S_k) \sum_{i=1}^{N} \sum_{j=1}^{N} p(g_i|S_k)\, p\left(g_j|S_k\right) d_{ij}, \tag{3}$$

where $p(S_k|g_i)$ is the conditional probability discussed above, $p(g_i)$ is the probability of observing segment g_i, and $p(S_k)$ is the marginal or occupation probability of the state S_k. The distance d_{ij} between the segment distributions $g_i(E)$ and $g_j(E)$ is measured with the Kantorovich distance,

$$d_{ij} = \int_0^1 dE \left| \int_{-\infty}^{E} \left(g_i(E') - g_j(E') \right) dE' \right|, \tag{4}$$

which is the area between the cumulative distribution functions of the pair of segments. The probability mass function of the segment g_i is computed as

$$g_i(E') = \frac{1}{N_i} \sum_{t=t_0^i}^{t_0^i+\tau_i} \delta(E(t) - E'), \tag{5}$$

where δ, $E(t)$, t_0^i, and N_i denote Dirac's delta function, FRET efficiency time series, the initial time of the time interval τ_i and $\int_0^1 \sum_{t=t_0^i}^{t_0^i+\tau_i} \delta(E(t) - E')dE'$, respectively.

The minimization of the rate $I(\boldsymbol{S}; \boldsymbol{g})$ with respect to $p(S_k|g_i)$ and the number of clusters n corresponds to compressing the data set into as few clusters as possible, with $p(S_k|g_i)$ distributed across clusters as evenly as possible. For example, in the most compressed case, there exists only a single cluster, i.e., $n = 1$, $I(\boldsymbol{S}; \boldsymbol{g}) = 0$. For a fixed number of clusters n and a variable level of distortion, as is the formulation of the information-based clustering problem,[28] minimizing $I(\boldsymbol{S}; \boldsymbol{g})$ distributes $p(S_k|g_i)$ across the set of n states $\boldsymbol{S}$ as evenly as possible. See, for example, the regions of overlap between adjacent states in Fig. 1(b), in which some segments g_i may belong to more than one state. Minimizing $I(\boldsymbol{S}; \boldsymbol{g})$ distributes these $p(S_k|g_i)$ across more than one state in $\boldsymbol{S}$, increasing state overlap and adjusting for error in the segment distribution $g_i(E)$. State overlap along with error magnitude causes difficulty in uniquely identifying the information of $\boldsymbol{S}$ given g_i, which is compensated with a smaller $I(\boldsymbol{S}; \boldsymbol{g})$. Note that when all $p(S_k|g_i)$ are either 1 or 0, the average mutual information $I(\boldsymbol{S}; \boldsymbol{g})$ is at its maximum. In turn, the minimization of $\langle D \rangle$ corresponds to minimizing the average intra-cluster distance. The smallest value of $\langle D \rangle$ is, in principle, zero when the number of clusters n coincides with that of segments N, i.e., the case in which each segment is in its own cluster (that is, the least compressed case).

Inputs to the RDT algorithm include the number of states n and the value of the parameter β. We perform clustering at several different numbers of states as well as values of β

and must subsequently select an appropriate model. Although there exist many model selection criteria, such as the Akaike information criterion,[29] Bayesian information criterion,[30] and the minimum description length principle,[31] these criteria are based on asymptotic results as the number of samples (the number of segments N in our case) increases, a condition that may not be satisfied in the case of single-molecule experiments. However, the essence of these model selection criteria is the same in all cases; we seek to minimize model complexity, i.e., the number of fitting parameters, while simultaneously maximizing goodness-of-fit. In the case of RDT, the mean distortion is a goodness-of-fit parameter in the sense that a good fitting model will have low distortion and a poor fitting model will have high distortion. Additionally, the mutual information is a model complexity parameter in the sense that a highly complex model needs a larger average rate of information.

In our calculation, we define a "distortion cutoff" by calculating the distortion arising in the best fitting model. In this model, each segment resides in its own cluster, i.e., $n = N$, and any nonzero distortion arises from the presence of errors. For noisy, finite time series acquired in smFRET measurements, we must consider the contributions of errors; errors arise from various sources, e.g., instrumental shot-noise, photophysical sources, and finite sampling error in the construction of the segment distributions. Briefly, the distortion cutoff was evaluated by incorporating both finite sampling error using bootstrap sampling[32] and experimental errors, incorporated in this work by randomly sampling the efficiency at time t, $E(t)$, from the normal distribution $N(E(t),\Delta E(t))$, where the empirical error $\Delta E(t)$ is derived from the observed numbers of acceptor, donor, and background photon counts using a normal error approximation.[16] Models having distortion below this cutoff satisfy the goodness-of-fit criterion because the distortion arising from the model is in the range of distortion arising solely from errors.

To assess model complexity, we note that the rate $I(\boldsymbol{S};\boldsymbol{g})$ provides the average amount of information needed to specify a segment g_i within the set of states $\boldsymbol{S}$. As such it provides a natural way to measure model complexity in that a more complex model will have a larger rate of information. We isolate the subset of models satisfying distortion criterion and compare them via their values of the mutual information $I(\boldsymbol{S};\boldsymbol{g})$. The model having the smallest mutual information while still satisfying the distortion criterion is selected to be the model for further analysis.[16] See Fig. S2 of the supplementary material for an illustration of the procedure.

1. Global modeling across different conditions

As discussed above, a set of conformational states arising from smFRET experiments observing protein folding/unfolding is expected to be consistent across each condition (e.g., denaturant concentrations). In such a case, segments originating from a particular state should be similar across all the conditions, although the error magnitudes of segments from one trajectory or condition may differ from those originating from a different trajectory or condition. Because the RDT method quantifies and incorporates the error directly into the clustering procedure, the extraction of a global set of states across multiple trajectories and conditions can be achieved by compiling the set of all trajectory segments from all different conditions into a single set of segments to be clustered by the RDT algorithm. In this manner, all segments from all conditions are considered simultaneously by the clustering algorithm, allowing for a consistent set of states to be extracted across different conditions without the imposition of any additional parameters or restrictions. After the conditional probabilities of each state given each trajectory segment are returned by the RDT clustering algorithm, the trajectory segments are reorganized into their respective positions within each set of trajectories at each different condition. Conformational state distributions, occupation probabilities, numbers of transitions, and transition probabilities can then be calculated within each denaturant concentration.

B. Realizations of state sequences and the "termination" state

For a given time-series, how can one generate the underlying state sequences and their kinetic quantities? The soft clustering method provides us with the conditional probabilities $p(\boldsymbol{S}|g_i)$ for each segment g_i from which the underlying states can easily be selected at each segment by sampling from the set of states proportionally to $p(\boldsymbol{S}|g_i)$. Kinetic quantities such as transition probabilities will be affected because the segment lengths τ_i are not uniform with the use of change-point detection, so we construct the underlying state sequences for each trajectory with a uniform time interval that is equivalent to the time step of the measurement (e.g., 50 ms) and assign a state to each uniform time step with probability proportional to the conditional probabilities $p(\boldsymbol{S}|g_i)$ for the segment g_i occurring at the time step. Construction of a state sequence for each trajectory in the data set allows for a statistically accurate calculation of the kinetic properties of states, such as transition probabilities and escape times. For example, the most probable state sequences are constructed as follows:

$$S^{(0)}(t_0), S^{(1)}(t_1), \ldots, S^{(i)}(t_i), \ldots, \tag{6}$$

where

$$S^{(i)} = S^{(i)}(t_i) = \underset{S_l}{\operatorname{argmax}}\, p(S_l|g(t_i)). \tag{7}$$

Here $g(t_i)$ denotes the corresponding segment distribution whose change-point interval $[t_{initial},t_{final}]$ includes the time t_i, i.e., $[t_{initial},\ldots,t_i,\ldots,t_{final}]$.

We are not limited to the most probable state sequences, however, as the state sequences can be randomly sampled proportionally to the conditional probabilities $p(\boldsymbol{S}|g(t_i))$. Repeating such a state assignment procedure many times will result in slightly different state sequences owing to the softness of RDT clustering, thereby increasing state sampling statistics and providing errors associated with each of the calculations as well. For this work, we generated 1000 independently sampled sets of state sequences, from which the transition probabilities, occupation probabilities, etc., were estimated as the median values of the calculated quantities. Errors are reported as the 95% confidence interval of the distribution of each quantity as generated from the 1000 sets of state sequences.

Because of the nature of the fluorophore photophysics in the AK experiments, the photobleaching kinetics of a particular state was found to be dependent on its FRET efficiency, possibly yielding a misinterpretation of state properties such as occupation and transition probabilities.[5] To account for these state-dependent photobleaching kinetics, a modified HMM in which photobleaching was modeled by artificial attachment of a segment to the end of each trajectory having a hypothetical FRET efficiency was constructed. These artificial segments constituted a "photobleaching" state from which there is no escape, which is known as an absorbing state in the absorbing Markov chain (AMC) theory.[33]

So that these kinetics may be analyzed from the results returned by the RDT clustering algorithm as well, we further modify the generated state sequences by adding a termination or "photobleaching" state S_{pb} at time step t_{n+1} just after each state sequence ending at t_n,

$$\begin{aligned} &S^{(0)}(t_0), S^{(1)}(t_1), \ldots, S^{(k)}(t_k), \ldots, S^{(n)}(t_n) \\ &\quad \rightarrow S^{(0)}(t_0), S^{(1)}(t_1), \ldots, S^{(k)}(t_k), \ldots, S^{(n)}(t_n), S_{pb}(t_{n+1}). \end{aligned} \tag{8}$$

Here $S^{(k)}$ denotes a state the system visits at time t_k. The termination probabilities arising from any state can be calculated by counting the number of transitions from $S^{(n)}$ to S_{pb}. Modification of the state sequences in this manner allows for the analysis of each state's transitions into this absorbing or "photobleaching" state in a natural way, without modification of the classification algorithm, as was the case with the modified HMM.[5]

C. Probability flow test for detailed balance

Here we develop a hypothesis test to determine whether or not the detailed balance condition is violated for a pair of states. For any pair of states S_i and S_j at equilibrium conditions, the flow of probability between them, J_{ij}, must be zero. We can express this probability flow as the difference between the rate from state S_i to state S_j, $p(S_i,S_j) = \pi_i p_{ij}$, and that from S_j to S_i, $p(S_j,S_i) = \pi_j p_{ji}$ where, for example, π_i and p_{ij} denote the stationary probability of state S_i and conditional probability of transition from state S_i to state S_j, respectively. We may also express this probability flow in terms of the observed numbers of transitions by using frequentist approximations for the rates, e.g., $p(i,j) = N_{ij}/N$ and $p(j,i) = N_{ji}/N$, where N_{ij} (N_{ji}) is the number of S_i to S_j (S_j to S_i) transitions, and N is the total number of observed time steps,

$$J_{ij} = p\left(S_i, S_j\right) - p\left(S_j, S_i\right) = \pi_i p_{ij} - \pi_j p_{ji} = (N_{ij} - N_{ji})/N. \tag{9}$$

We note that in order to satisfy the detailed balance condition in which $J_{ij} = 0$, the number of transitions in one direction must be equivalent to the number of transitions in the reverse direction, i.e., $N_{ij} = N_{ji}$. Because the smFRET trajectories are finite in time, fluctuation in the observed numbers of transitions will occur, thus leading to the situation that N_{ij} and N_{ji} are not equivalent due to this fluctuation. The problem then becomes one of verifying that N_{ij} and N_{ji} are drawn from the same distribution.

Though there are methods that test such hypotheses, such as binomial tests,[12,13] we wish to take into consideration that counting the number of transitions in a stochastic system over a fixed and finite length of time, as we have done for the smFRET system of AK folding, can be viewed as a Poisson process with the numbers of observed transitions thus obeying the Poisson distribution. We note that the quantity $\Delta_{ij} = J_{ij}N$ is the difference of the two Poisson variables N_{ij} and N_{ji} and that the number of time steps N remains constant for each trial. Thus, the difference of the two numbers of transitions, Δ_{ij}, will follow what is known as a Skellam distribution.[34] The probability that $\Delta_{ij} = \Delta$, where Δ is the integer-valued difference between the numbers of transitions, thus takes the following form:

$$p(\Delta_{ij} = \Delta) = \left(\frac{N_{ij}}{N_{ji}}\right)^{\frac{\Delta}{2}} e^{-(N_{ij}+N_{ji})} I_{\Delta}\left(2\sqrt{N_{ij}N_{ji}}\right). \tag{10}$$

Here, $I_{\Delta}(x)$ is a modified Bessel function of the first kind.

Having obtained the probability distribution of the difference in the numbers of transitions between a pair of states in the network, we may now test the null hypothesis that the transition obeys equilibrium properties, i.e., H_0:$\Delta_{ij} = 0$. This is achieved by constructing the Skellam distribution assuming that $N_{ij} = N_{ji} = (N_{ij} + N_{ji})/2$ thus having zero mean and variance $N_{ij} + N_{ji}$, i.e.,

$$p(\Delta|H_0) = e^{-(N_{ij}+N_{ji})} I_{\Delta}\left(N_{ij} + N_{ji}\right). \tag{11}$$

The probability $p(\Delta_{ij}|H_0)$ is then obtained by evaluating Eq. (11) at Δ_{ij}. Because this is a two-tailed hypothesis test, we may reject H_0 if $p\left(|\Delta| > |\Delta_{ij}||H_0\right) \leq \alpha$, where α is a confidence interval, e.g., 5%. Rejection of the null hypothesis indicates the transition is not at equilibrium.

D. Change-point detection

Change-point detection has been applied in many time-series analyses and in many variations.[5,14,18–21] Here we employ a distribution-free method that involves identification of the extrema of the cumulative sum of a modified time-series.[18] After these extrema are identified, a hypothesis test is then performed to determine whether or not the difference between them is large enough to be classified as a change point.

The first step in the procedure is the subtraction of the mean $\bar{E}$ of the original m-step time-series $E(t)$ [Fig. 2(a)], $\bar{E} = \frac{1}{m}\sum_{t=0}^{m} E(t)$, generating the modified time-series as shown in Fig. 2(b). The cumulative sum time-series $E'(t)$, shown in Fig. 2(c), is then calculated from this modified time series. In general, for any time step t, the cumulative sum time series $E'(t)$ is generated as follows:

$$E'(t) = \sum_{t'=0}^{t} \left(E(t') - \bar{E}\right). \tag{12}$$

The extremum of the cumulative sum time-series indicates the most probable location of a change-point. As such, we calculate a test statistic Δ and perform a hypothesis test to determine whether or not the time-series contains a change-point,

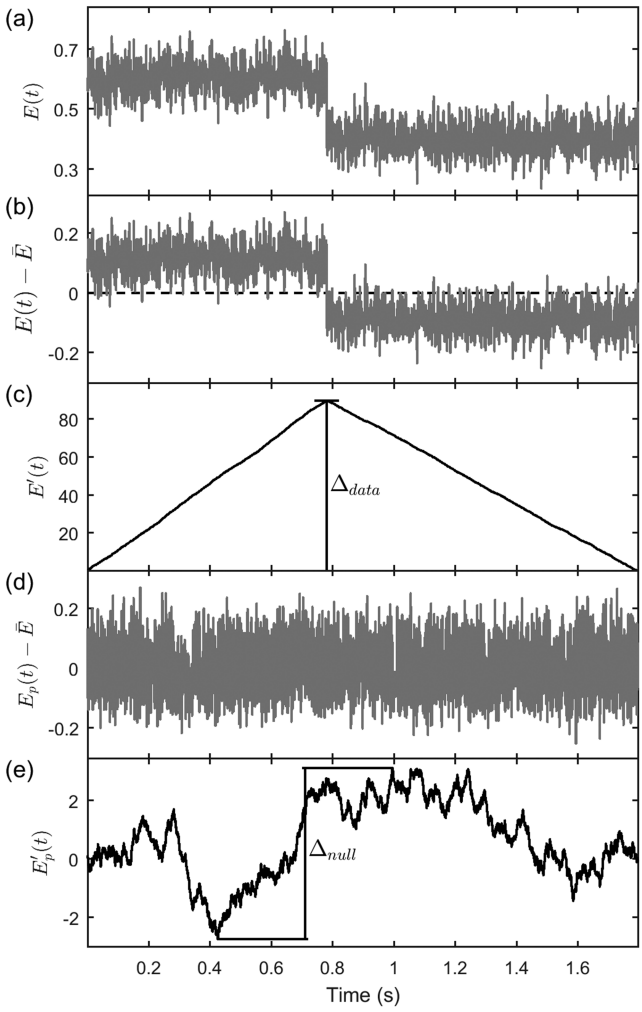

FIG. 2. Illustration of the change-point detection algorithm. (a) The original time-series. (b) The mean-subtracted time-series. (c) The cumulative sum time-series showing the test statistic, Δ_{data}. (d) Random permutation of the mean-subtracted time-series. (e) Cumulative sum time-series of the permutation showing a test statistic under the null hypothesis, Δ_{null}.

$$\Delta = \max_t E'(t) - \min_t E'(t). \tag{13}$$

The null hypothesis H_0 is then assigned to be that the time-series contains no change point, and we must now obtain the sampling distribution for our test statistic under assumption that null hypothesis is true, e.g., $P(\Delta|H_0)$. Because H_0 assumes that there is no change point in the time-series, random permutation of the time series $E(t)$, generating $E_p(t)$, will have little effect on the value of the test statistic Δ if H_0 is true. As such, we construct the sampling distribution we desire, $P(\Delta|H_0)$, through repeated random permutations of the time series $E(t)$ and subsequent calculations of the test statistic Δ from the permuted time-series. This procedure is illustrated in Figs. 2(d) and 2(e). To ensure convergence of $P(\Delta|H_0)$, the permutation is repeated until all points in a discretely binned distribution do not change within a specified tolerance. In practice, for the smFRET data we analyzed, convergence is typically achieved in approximately 1000 permutations.

A p-value for the test statistic from the original time-series, which we term Δ_{data}, is then calculated, i.e., $p = P(\Delta > \Delta_{data}|H_0)$. The resulting p-value is then compared to the input type I error rate α, e.g., 5%. Then, if $p \leq \alpha$, the null hypothesis is rejected, and a change-point is assigned in the original time series at the extremum of the cumulative sum time-series $E'(t)$.

Type II error, i.e., the probability of missing an existing change-point, cannot be explicitly controlled. To prevent missing a large number of change-points, the choice of an extremely small value of α should be avoided. This increases the detection of false change-points (type I error), but many of these falsely detected change-points can be removed by the soft clustering procedure discussed above.

To test for more than one change-point occurring within a single trajectory, a procedure called binary segmentation[20] is used. The original time series is divided into two disjoint time-series at the first identified change-point, and the procedure outlined above is repeated on the two resulting time-series individually. This procedure is then repeated until no further change-points are found within any of the resulting segments of the time-series.

Once the change-points within a trajectory have been acquired, the error in the location of a change-point between two segments is calculated as follows: Under the assumption that any segment between two change-points arises from the same distribution, each of the two adjacent segments are bootstrapped, i.e., randomly sampled with replacement, to generate a bootstrapped pair of segments. Next, each data point within the bootstrapped segments is randomly sampled from its empirical error distribution, producing realizations of the pair of segments that may arise in the uncertainty of finite sampling and empirical errors. Although the error distributions in this work are treated as normal, errors that are not normally distributed are easily incorporated as long as the error distribution is known. After the two segments have been bootstrapped and randomly sampled, the change-point is again identified, most likely resulting in a slightly different location. This entire procedure is then repeated many times (e.g., ~1000) to obtain the distribution of change-point locations between the two segments in question, from which the error in the location is inferred.[21]

III. RESULTS AND DISCUSSION

A. State distributions, occupation probabilities, and state assignments along smFRET trajectories of AK unfolding

The state distributions arising from the four-state model acquired by the method presented in Sec. II are shown in Fig. 3, where the set of states is extracted across the GdmCl concentrations by compiling the set of all trajectory segments from all denaturant concentrations into a single set of segments to be clustered by the RDT algorithm. We note that although the selected model contains fewer than the six states extracted by the previous HMM analysis,[5] because our model selection procedure returns a minimal model that is based on quantifying errors in the measurement there may indeed be more states underlying the data, but errors incurred during the experimental measurement obstruct their observation. Figures 3(a)–3(e) contain the conformational state

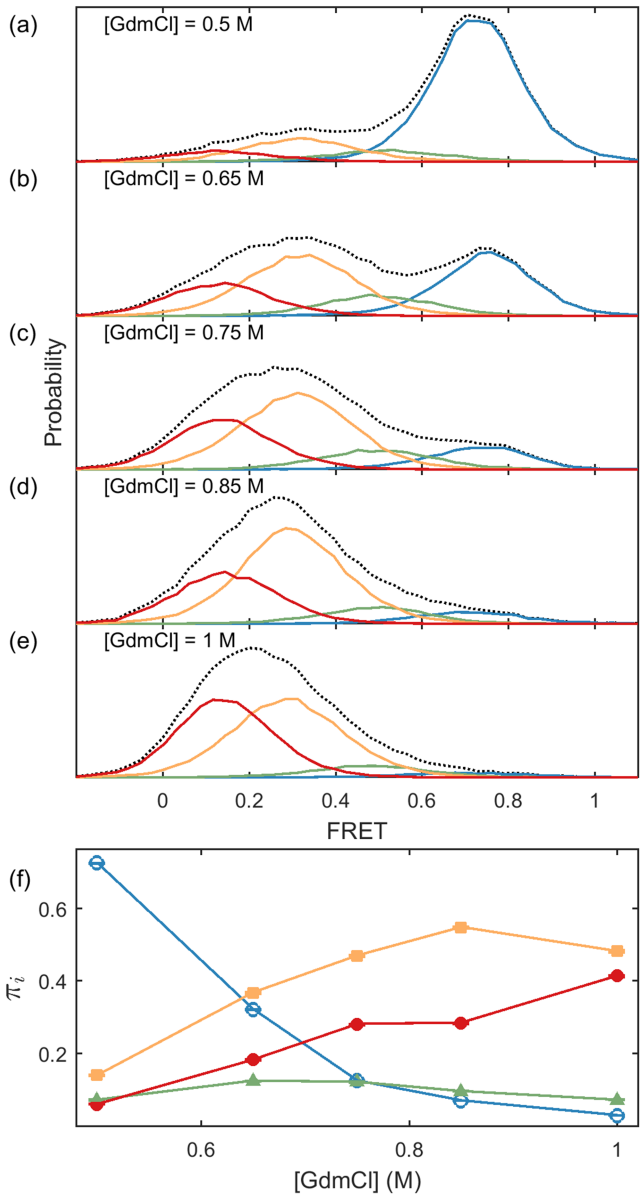

FIG. 3. Globally extracted state distributions and occupation probabilities for all denaturant concentrations. State distributions are shown for (a) [GdmCl] = 0.5M, (b) [GdmCl] = 0.65M, (c) [GdmCl] = 0.75M, (d) [GdmCl] = 0.85M, (e) [GdmCl] = 1M. Dotted lines in (a)–(e) denote experimental distributions. (f) Occupation probabilities π_i of each state vs. GdmCl concentration. States are represented with open circle, triangle, square, and closed circle markers, respectively, in order of descending efficiency.

distributions resulting at the denaturant concentrations 0.5, 0.65, 0.75, 0.85, and 1M, respectively, while Fig. 3(f) follows changes in the state occupation probabilities as a function of GdmCl concentration. To obtain error bars for the occupation probabilities, each 50 ms time step in each trajectory at each denaturant concentration was assigned to a state according to the $p(S_k|g_i)$ at the corresponding segment. Occupation probabilities $p(S_k) = N_k/N$ were then calculated, with N_k being the number of time steps assigned to S_k and N being the total number of time steps at a particular denaturant concentration. This process was repeated 1000 times to obtain a distribution for each of the $p(S_k)$, from which we obtain the error bars as the 95% confidence interval of the distribution. Error bars in Fig. 3(f) are on the order of 1×10^{-3} and are contained within the markers for each state at each concentration.

If the hypothesis that a similar set of states underlies the acquired data at all conditions is valid, then the state distributions must also be similar across the denaturant concentrations. If any of the states show inconsistent distributions across the conditions, then the hypothesis is shown invalid. To check whether this condition is satisfied, we measure the Kantorovich distance $d\left(G_k(E,c);\bar{G}_k(E)\right)$ of each of the state distributions at each concentration,

$$G_k(E;[\mathrm{GdmCl}]) = \sum_{i=1}^{N} p(g_i|S_k)g_i(E), \tag{14}$$

from the mean distribution of each state across all denaturant concentrations,

$$\bar{G}_k(E) = \frac{1}{C}\sum_{c=1}^{C} G_k(E;c), \tag{15}$$

where C is the number of different GdmCl concentrations. To obtain a relative distance, each of the $d\left(G_k(E,c);\bar{G}_k(E)\right)$ is normalized by the average distance $\bar{d}$ among all $\bar{G}_k(E)$. We refer to this quantity in Fig. 4 as "relative state distance." Figure 4 indicates the maximum deviation to be approximately 7.5% relative to the average distance between states, thereby validating that all state distributions are consistent across all denaturant concentrations.

Examination of the state distributions shown in Figs. 3(a)–3(e) visually confirms this consistency across all denaturant concentrations, as all state distributions are visually similar to their counterpart distributions at the other conditions. Furthermore, as shown in Fig. 3(f), the conformational occupation is consistent with the expectation that increased concentrations of denaturant serve to destabilize the folded conformations of the protein. The occupation probability π_i of the most compact form of the protein, having the highest FRET efficiency (shown in blue in Fig. 3), is seen to decrease consistently as the concentration of the denaturant increases. Conversely, the occupation probabilities of the less compact, lower FRET efficiency states (shown in purple) consistently rise with the denaturant concentration, demonstrating the increased occupation of less

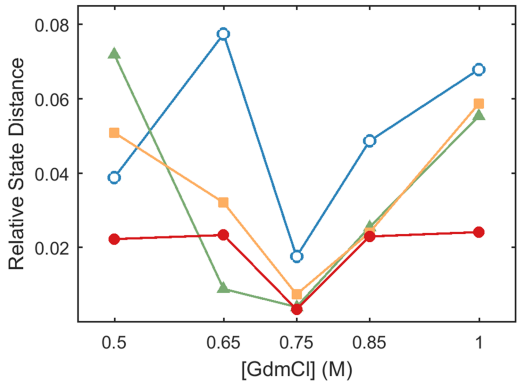

FIG. 4. Kantorovich distances of each state's distribution at each denaturant concentration from its mean distribution relative to the average distance between states. States are again represented with open circle, triangle, square, and closed circle markers in order of descending efficiency.

compact conformational states at increased [GdmCl]. Interestingly, the conformational state shown with green triangles in Fig. 3, occupying the intermediate FRET efficiencies, displays a slight increase in occupation at smaller GdmCl concentrations that is followed by a slight decrease at higher concentrations. This behavior is well outside the 95% error bounds for each occupation probability, suggesting that it is an intermediate conformation along the unfolding pathway, in agreement with similar states found in Ref. 5 using a HMM approach. In general, these results are consistent with the expected conformational behavior of the protein as a function of denaturant concentrations; at low [GdmCl], the most compact forms of the protein are favored and are progressively destabilized as the denaturant concentration increases.

Trajectories shown in Fig. 5 include a representative for each of the five denaturant concentrations. Upper panels of Figs. 5(a)–5(e) show FRET efficiency vs. time, and each point in the trajectory is colored according to its most probable state $S^{(i)}(t_i)$. Lower panels illustrate the conditional probabilities for each state given each segment, $p(S_k|g_i)$, which are returned by the RDT clustering algorithm. Each segment returned from change-point detection is assigned a vertical bar along the time course, and the heights of each colored bar correspond to the magnitude of the conditional probability $p(S_k|g_i)$ for each state S_k and each segment g_i. The colors representing each state are the same as those used in Figs. 3 and 4.

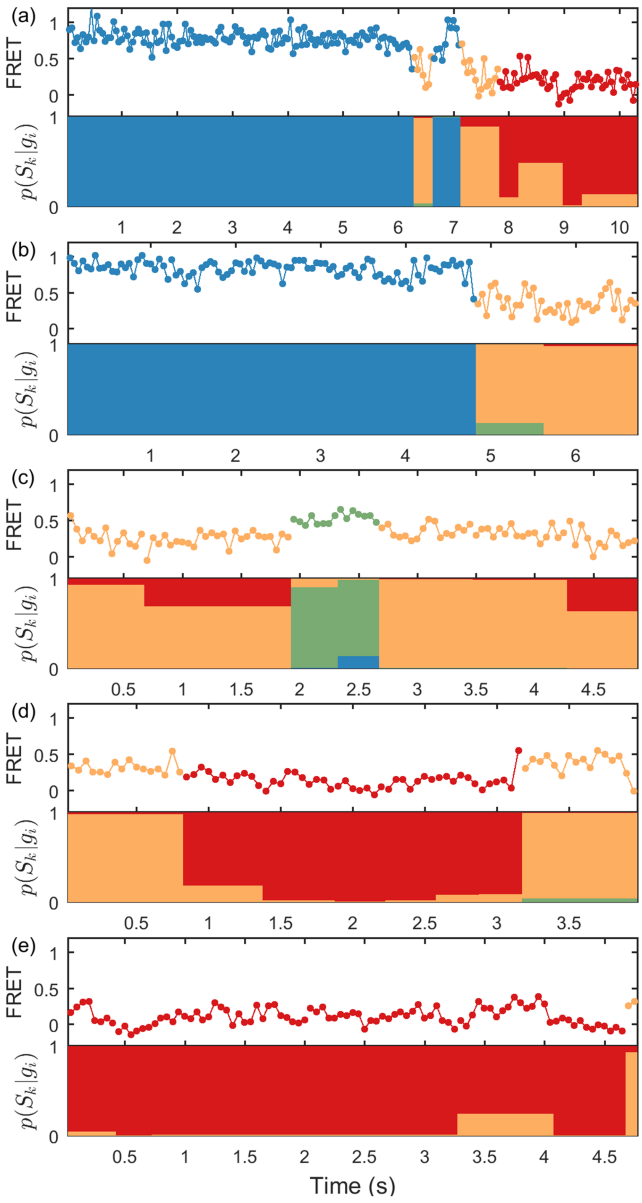

FIG. 5. Representative trajectories and relative state assignments for: (a) [GdmCl] = 0.5 M, (b) [GdmCl] = 0.65 M, (c) [GdmCl] = 0.75 M, (d) [GdmCl] = 0.85 M, (e) [GdmCl] = 1M.

The most compact form of the protein is represented in blue color in Fig. 5 and is characterized by relatively long residence times and high certainties in state assignment, as shown by the large magnitudes of $p(S_k|g_i)$ at segments in which this state is favored. As GdmCl concentration increases, the relative decrease in occupation probability of this state is reflected by its absence in the representative trajectories. The trajectories at lower GdmCl concentrations are also shorter in comparison to those at higher concentrations, which is in agreement with the expectation[5] that termination probability is higher for lower efficiency states.

B. State uncertainties, residence times, and termination probabilities

Figure 6 displays various properties of the states as a function of the GdmCl concentration. Figures 6(a) and 6(b) show the quantification of uncertainty in state assignments at different GdmCl concentrations. For example, as seen in Fig. 5(a) at [GdmCl] = 0.5 M, the highest efficiency state S_1 (indicated with closed blue circles) is most favored at segments occurring on the interval $0 \le t \le\sim 6$ s, i.e., $S^{(i)}(t) = S_1$ according to Eq. (7), with a high degree of certainty, i.e., $p(S_1|g_i) \to 1$. At all but one segment occurring after ~6s, the lower efficiency states S_3 and S_4 (orange squares and closed red circles, respectively) are the favored states, having maximum values of $p(S_k|g_i)$ at those segments. However, as indicated by the lower panel of Fig. 5(a), the magnitudes of the $p(S_k|g_i)$ at these segments are smaller than those of $p(S_1|g_i)$ on the time interval $0 \le t \le \sim 6$ s, implying relative uncertainty in their state assignments. To quantify this uncertainty, we calculate the mean of $p(S_k|g_i)$ when the most favored state at segment g_i is S_k, that is,

$$\hat{p}\,(S_k) := \frac{\sum_{t_i=0} p\left(S'^{(i)}|g\,(t_i)\right)\delta\left(S'^{(i)} - S_k\right)}{\sum_{t_i=0}\delta\left(S'^{(i)} - S_k\right)}. \qquad (16)$$

Here δ indicates the Kronecker delta function. Larger values of $\hat{p}\,(S_k)$, plotted vs. [GdmCl] in Fig. 6(a), imply a higher degree of certainty in state assignment while lower values imply the opposite. Figure 6(b) shows the 95% confidence interval $\Delta\hat{p}(S_k)$ on the distribution of the $p(S_k|g_i)$ when S_k is the most favored state. From Fig. 6(a), we can see that $\hat{p}\,(S_k)$ of the lower efficiency states are fairly constant with changing [GdmCl], as is the 95% confidence interval $\Delta\hat{p}(S_k)$. Conversely, $\hat{p}(S_k)$ for the most compact form decreases with increasing [GdmCl] while the fluctuation in them, $\Delta\hat{p}(S_k)$, increases. This is most likely due to decreased residence time in state S_1, thereby increasing the sampling error in the calculation of the probability distributions of segments.

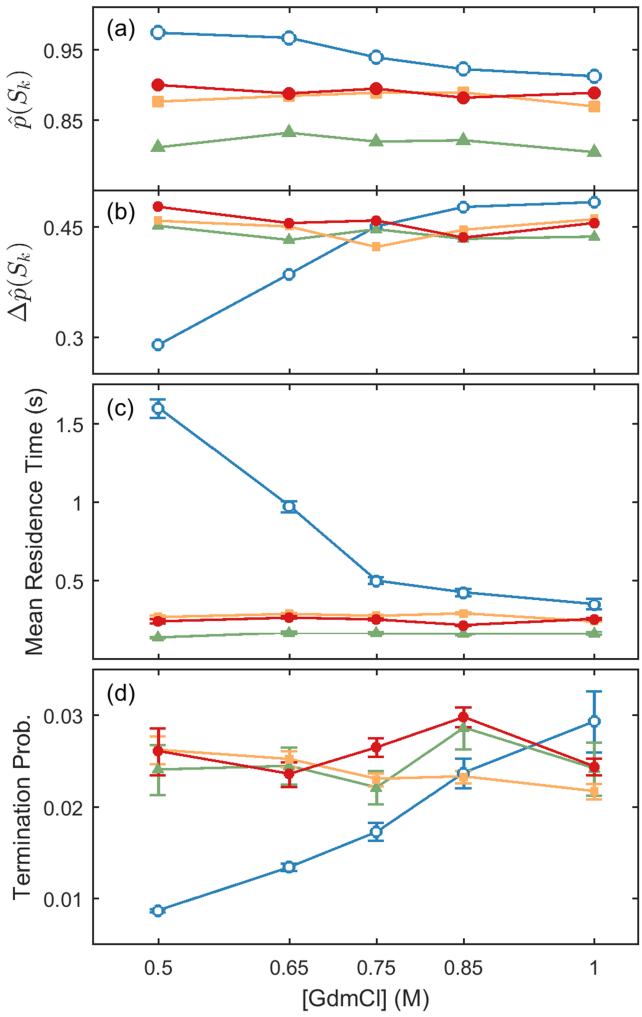

FIG. 6. State properties vs. GdmCl concentration. (a) The mean conditional probabilities calculated from segments in which the state in question is favored. (b) The 95% confidence interval of the conditional probability as calculated from segments in which the state in question is favored. (c) The mean escape time of each state as a function of denaturant concentration. (d) The termination probability of each state as a function of GdmCl concentration. States are represented with open circle, triangle, square, and closed circle markers in order of descending efficiency in all panels (a)–(d).

Figure 6(c) confirms a decrease in the residence time of the most compact conformational state (shown with open blue circles), as it is reduced by a factor of 4.6 from 1.6 s to 0.35 s as the denaturant concentration increases. Interestingly, the less compact forms of the protein do not exhibit the same behavior, with their residence times remaining fast and reasonably constant for all denaturant concentrations. This result suggests that these three less compact states of the protein occupy an unfolded/partially folded "superbasin" in which escape kinetics are dominated by fast internal transitions rather than transitions to the folded state and are relatively unaffected by GdmCl concentration.

Finally, we examine the termination probabilities of each state in Fig. 6(d). In Ref. 5, the authors added a photobleaching state to their HMM formalism to account for the expectation that the photobleaching times of states having lower FRET efficiencies would be faster owing to the donor fluorophore being less photostable than the acceptor fluorophore. In short, because Förster energy transfer generally occurs on a faster time scale than fluorescence, the donor spends less time in an excited state and thus is less likely to be photobleached in a high efficiency, more compact, conformational state than in a lower efficiency, less compact state. The results shown in Fig. 6(d) indicate that this is the case, as the lower efficiency states, S_2 (green triangles), S_3, and S_4 typically have a larger termination probability than the high efficiency state, S_1. The behavior shown in Fig. 6(d) is not so simple, however, in that the termination probabilities depend on GdmCl concentration. This is particularly evident for state S_1 at higher [GdmCl], at which the termination probability increases beyond what is explained by the increased size of the error bar that arises from decreased sampling of state S_1.

We note that although trajectory termination has been modeled here as a transition to a single photobleaching state, termination in a smFRET experiment can occur via multiple mechanisms having distinct rates of transition. For example, photobleaching of the donor fluorophore may be faster on average than photobleaching of the acceptor fluorophore. Another source of trajectory termination may be the premature truncation prior to any photobleaching events due to large shifts in the total fluorescence intensity.[5] To investigate the effects of multiple termination pathways, we use the experimental photon trajectories to identify whether the events occurring at the end of each trajectory are donor photobleaching, acceptor photobleaching, or other termination events, such as the intensity fluctuations described above. See Sec. S5 of the supplementary material for full details of event assignment.

Figure S3 of the supplementary material displays the results of the computation. Figure S3(a) shows the state-dependent fraction of acceptor photobleaching events as a function of [GdmCl], Fig. S3(b) shows that of donor photobleaching, and Fig. S3(c) displays that of other termination events. As shown in Fig. S3 of the supplementary material, while the lower efficiency states show consistent behavior, with acceptor photobleaching events increasing, donor photobleaching events decreasing, and other termination events remaining fairly constant as [GdmCl] increases, the behavior of the highest efficiency state deviates from the others. The fraction of acceptor photobleaching decreases while those of donor photobleaching and other termination events increase. While the present discussion suggests possible reconciliation of the probabilities of photobleaching arising from these various termination events, such reconciliation is nontrivial and is beyond the scope of the present work.

C. Probability flow tests reveal violations of detailed balance

To this point, RDT analysis of experimental smFRET trajectories of the folding and unfolding behavior of adenylate kinase suggests a transition network consisting of at least four conformational states, one state being a compact, folded form of the protein and the other three states being various partially folded and/or unfolded conformations. The escape times of the four states also suggest that the most compact form exhibits relatively slow transitions and high occupation probabilities at low concentrations of the denaturant while the

other three states occupy an unfolded/partially folded superbasin that exhibits fast internal transitions with residence times remaining relatively constant at all GdmCl concentrations. A natural way to visualize such behavior is through the construction of transition disconnectivity graphs (TRDGs),[33,34] which is a projection of the multidimensional free energy landscape onto a 1-D observable coordinate (e.g., smFRET).[16] The construction of a TRDG requires detailed balance to hold for the equilibrium system, so we must verify that the state networks returned have equilibrium properties. We thus examine the numbers of transitions and the probability flow between pairs of states. Figure 7 contains the numbers of single-time step transitions obtained for each of the five GdmCl concentrations. Each entry in the transition matrices contains the median number of transitions N_{ij} between an initial state S_i, along the rows, and a final state S_j, along the columns, which were estimated from 1000 sets of randomly sampled state sequences. Diagonal entries in each transition matrix contain the occupation probabilities $\pi_i = N_i/N$ for each state, with N_i being the number of visits to S_i and N being the total number of time steps, and each is outlined with color corresponding to those used in Figs. 3–5 for each state. Rows and columns of each matrix are arranged in order of descending efficiency from top to bottom and from left to right, respectively. The fifth column of each transition matrix, outlined in black, contains the median number of transitions from each state to the termination state. See Figs. S4 and S5 of the supplementary material for the transition matrices corresponding to the most probable state sequences and the transition probability matrices, respectively.

The flow of probability among states has been called a principal characteristic in the study of nonequilibrium systems[35] and has been proposed as a measure of the degree of nonequilibrium behavior.[36] Here we use the probability flow between each pair of states S_i and S_j, $J_{ij} = \pi_i p_{ij} - \pi_j p_{ji}$, to assess whether or not equilibrium properties are maintained between pairs of states. Here $p_{ij} = N_{ij}/N_i$ represents the transition probability from state S_i to state S_j. The probability flow J_{ij} derives from the detailed balance condition, which states that the net probability flow in an equilibrium system should approach zero. We develop a hypothesis test, described in Sec. II C, to accept or reject the null hypothesis that the probability flow between each pair of states is indeed within error of zero.

We calculate J_{ij} for each pair of states and from each of the 1000 sets of randomly sampled state sequences and then test for violation of the detailed balance condition. Each off-diagonal entry in Fig. 7 is colored according to the color scale shown to the right of the figure, which indicates the fraction of the 1000 randomly sampled sets of state sequences that violate detailed balance for each transition in each network. Transitions that are red-colored are transitions in which detailed balance is violated in most of the randomly sampled state sequences, while those that are white-colored indicate transitions for which equilibrium properties hold for most sets of sequences.

Compared to the corresponding Fig. S4 of the supplementary material, constructed using the most probable state sequences according to Eq. (7), in which many pairs of states violate the probability flow test, Fig. 7 shows that the increase in statistical sampling arising from the use of randomly sampled state sequences results in transition networks that are much closer to satisfying the detailed balance condition. Figure 7 still shows, however, that all five denaturant concentrations have transitions that violate the detailed balance condition to some extent, and the lower the denaturant concentration [GdmCl], the more the number of the pairs of states that violate the test increases. While these violations mostly involve infrequent transitions, such as the highest efficiency state transitioning to the lowest efficiency state, there are others that involve large numbers of transitions, namely, the transition involving the two lowest efficiency states at the 0.85M GdmCl concentration. Because the conformational motion of the protein is expected to be at equilibrium,[5] the extracted conformational state network is expected maintain equilibrium properties. As such, the source of these equilibrium violations will be the primary focus of the remainder of this work.

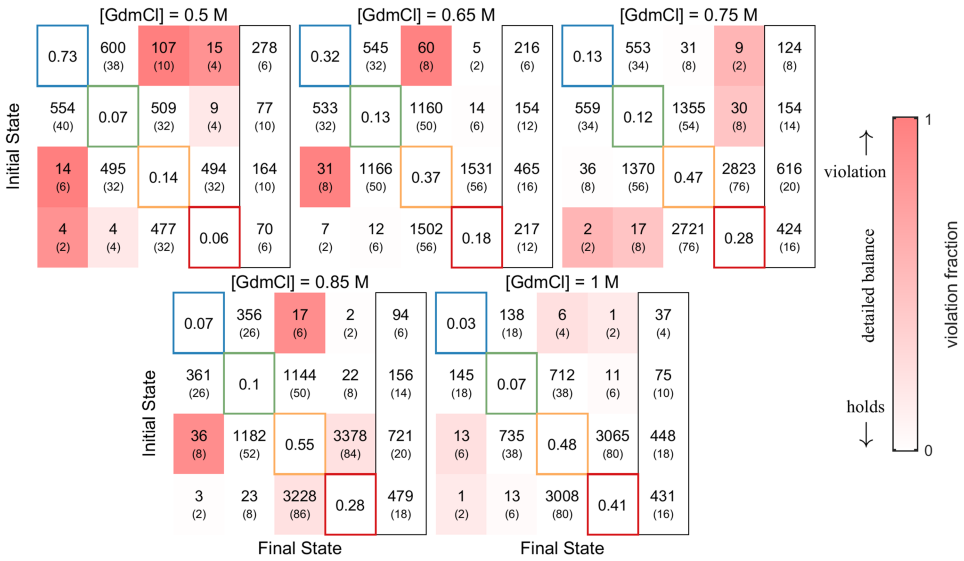

FIG. 7. Median numbers of state-to-state transitions observed in the randomly sampled sequences. 95% confidence intervals for each number of transitions are indicated in parentheses below each entry. The diagonal entries of each matrix contain each state's occupation probability, and each is outlined with the color used in Figs. 3–6 for each state. States are arranged in order of descending mean efficiency from top to bottom and from left to right. Each of the off-diagonal entries is colored according to the color scale to the right. Off-diagonal entries that are white correspond to transitions in which detailed balance holds while those that are red show significant violations of detailed balance. Numbers of transitions of each state to the absorbing, or photobleaching state, are denoted by black outlines.

D. Unbalanced termination rates induce apparent nonequilibrium behavior

Perhaps the most intuitive reason that the system would display nonequilibrium behavior is because the transition probabilities are still changing with time because too few transitions have been observed. However, Fig. S6 of the supplementary material shows that this is not the case. Figure S6 tracks changes in each of the transition probabilities after each time step in the randomly sampled state sequences and shows that while there is fluctuation in the early time steps, the median values of all the transition probabilities have converged to constant values. Because the transition probabilities have been calculated from 1000 sets of state sequences that are randomly generated proportionally to $p(S_k|g_i)$ and because the number of time steps observed in the smFRET trajectories is large, even the median probabilities having very few numbers of transitions have converged.

Having confirmed that the probabilities of transition are sufficiently converged, we posit that another reason for equilibrium violation is trajectory termination. Assuming the behavior of the underlying protein dynamics is Markovian, the termination state constitutes an absorbing state, and a state sequence resulting from state assignment along a smFRET trajectory is an absorbing Markov chain (AMC).[33] As discussed in Sec. S7 of the supplementary material, AMCs are Markov chains that contain at least one absorbing, or inescapable state, along with another (set of) non-absorbing state(s), which is often called the transient set. As implied, the probability of observing a transient state goes to zero in the long-time limit, thus destroying positive-recurrence and violating equilibrium properties.

The smFRET measurement can thus be interpreted as nonequilibrium owing to the destruction of the positive-recurrence of the set of conformational states. This does not necessarily imply that the underlying system is not at equilibrium; rather it implies that the method of observation may induce nonequilibrium behavior. In effect, the smFRET measurement is two interacting systems, the photophysical system being a window through which the underlying biological system is observed. Changes in the biological system may alter the behavior of the photophysical system in the form of different termination rates originating from different conformational states, as we observe above in Fig. 6(d). It is thus of interest to investigate the effects of termination on the equilibrium properties of a smFRET system.

We present this investigation in Sec. S8 of the supplementary material. Section S8 presents two Markov chain Monte Carlo (MCMC) simulations. The first is an AMC simulation in which termination is modeled as an absorbing state, and the second is a coupled Markov chain in which termination is modeled as an external, coupled window consisting of an on state and an off state. A two-state equilibrium system is observed while the on/off system occupies the on state, and observation ceases when the system transitions to the off state. Figure S7 of the supplementary material displays the results of these simulations. Figure S7(a) shows that a 2-state AMC having equivalent termination probabilities maintains properties expected of equilibrium systems, returning equivalent numbers of transitions and input occupation and transition probabilities. The properties of the system shown in Fig. S7(b), in which the termination probabilities are unbalanced, are quite different. The numbers of transitions are reduced and are no longer equivalent, but most notably the occupation probabilities are not equivalent to the inputs, favoring the state with the smaller termination probability. Finally, the coupled MCMC shown in Fig. S7(c) indicates that even if the underlying system is at equilibrium, its behavior may appear to be nonequilibrium when coupled to the external on/off window model of termination.

The behavior of the absorbing and coupled systems observed in Fig. S7 of the supplementary material is consistent with the behavior observed in the experimental smFRET systems of the AK folding landscape presented above. From Fig. 6(d), the largest difference in photobleaching probabilities occurs at the lowest GdmCl concentration, at which the observed termination probability of the highest efficiency state is approximately 2.5 times smaller than those of the lower efficiency states. It is at this GdmCl concentration that we also observe the most significant deviations from equilibrium behavior, as indicated by the observed numbers of transitions and the results of the probability flow tests in Fig. 7. As the GdmCl concentration increases, the termination probability of the highest efficiency state also increases such that at the highest denaturant concentration, all termination probabilities from all conformational states are similar. It is at this highest GdmCl concentration that we observe the transition network that most resembles equilibrium behavior. These results suggest that it is the response of the fluorophores used in the smFRET measurement to conformational motion in the biological system, rather than the conformational motion itself, that leads to the observation of nonequilibrium behavior. At low GdmCl concentrations, large differences in observed termination probabilities contribute to deviation from equilibrium behavior, while at higher GdmCl concentrations, smaller differences in observed termination probabilities contribute to less significant deviations from equilibrium behavior.

E. Transition disconnectivity graphs at different time scales illustrate the free energy landscape of AK folding/unfolding

To examine approximations of the free energy landscapes of AK folding/unfolding on multiple time scales, we display transition disconnectivity graphs at different [GdmCl] and observation time scales. Time-dependent aspects of free energy landscapes were discussed in the traveling salesman problem,[37] Ising spin models,[38] computer simulation of a model protein,[13,39–41] and also pointed out in single-molecule experiments,[42,43] but there exists no systematic elucidation in terms of smFRET time series. To generate the time scale-dependent TRDGs, the photon-by-photon smFRET trajectories were first binned to uniform time steps of 50, 200, and 400 ms. Change-points were then detected in each data set, generating sets of segments to be clustered with RDT as described in Sec. II A. After segments were clustered, model selection was performed, and state sequences were constructed as described in Sec. II B.

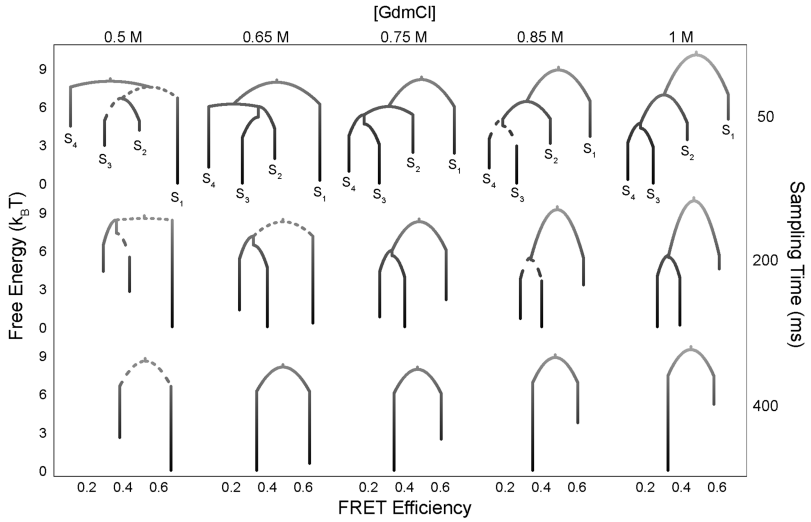

FIG. 8. Free energy landscape approximations of the AK folding landscape are shown as transition disconnectivity graphs at multiple time scales for all GdmCl concentrations. Each TRDG was constructed as described in Sec. S4 of the supplementary material. Each TRDG is positioned horizontally according to [GdmCl] and vertically according to sampling time. The vertical scale is normalized such that minimum free energy of zero corresponds to the most occupied state in each TRDG, and the maximum is the free energy of the set of all barrier energies relative to their corresponding minimum free energy state. The color gradient in each TRDG is scaled according to the maximum of the set of all relative barrier energies in units of k_BT. The position of each state corresponds to the mean efficiency of the state, as calculated from the state distributions returned by the RDT clustering method. Barrier calculations in which the detailed balance condition was not satisfied according to the probability flow test described in Sec. II C appear as dashed lines.

Figure 8 presents TRDGs representing approximate free energy landscapes of AK folding/unfolding on observation time scales of 50, 200, and 400 ms at each of the five GdmCl concentrations. The graphs for each condition are plotted horizontally vs. [GdmCl] and vertically vs. observation time scale. Free energies at the transition barriers and for each node (i.e., conformational state) were calculated as described in detail in Sec. S4 of the supplementary material. Each node is plotted at the mean efficiency of the corresponding state, and the free energy at each node is normalized such that the most occupied state in each system has zero free energy. Each barrier is represented as a curve, with the relative energy at the barrier being positioned at the apex of each curve. We note that curved barriers are used for visual clarity only and do not represent changes in free energy. Each line in the TRDGs is colored according to free energy in units of k_BT, with black color representing those near zero and lighter gray colors representing larger values. All free energies are normalized to the same color scale. Calculated barriers that do not satisfy the detailed balance condition according to the probability flow test across the barrier, as described in Sec. II C, are indicated with dashed lines.

TRDGs are often calculated with methods designed to reduce computational complexity, such as the Ford-Fulkerson method[44] or the Gomory-Hu algorithm.[45] Because our networks only have four states, we instead use a brute force approach, directly calculate the numbers of transitions between every possible pair of subsets in the networks, and then arrange them in descending order to construct the TRDGs. This brute force approach returns barriers that represent energies required to transition from a single state to any of the other states, rather than barriers between pairs of states as in the Ford-Fulkerson approach. This is advantageous in that avoids the assumption that transitions between pairs of states having low barrier heights are fast enough to be at equilibrium in comparison to slower transitions with larger barrier energies. See Sec. S4 of the supplementary material for complete details.

First we note the hierarchical nature of the states as observation time scale increases; moving from the 50 ms time scale to 200 ms reduces the observed number of states from four to three, with the two lowest efficiency states (S_3,S_4) merging to become a single state. With another increase in observation time scale to 400 ms, the intermediate efficiency state (S_2) merges with the already-combined lower efficiency states, producing a free energy landscape with two basins; one basin containing the more folded forms of the protein (S_1) and another containing the more unfolded forms (S_2,S_3,S_4).

The reduced occupation probability and relative destabilization of the folded form of the protein (S_1) is reflected as a function of [GdmCl] at all observation times; as [GdmCl] increases, the relative free energy of the highest efficiency S_1 increases and that of S_2, S_3, and S_4 decreases. Also reflecting this destabilization are the relative barrier heights from S_1 to the unfolded states. At low [GdmCl], the barrier height from S_1 to the open states is relatively large but decreases as [GdmCl] increases, which is indicative of the decrease in its escape time as shown in Fig. 6(c). Furthermore, at the 50 ms observation time, the barrier heights from each of the unfolded forms do not change significantly as a function of [GdmCl], indicating that this is the reason for the escape times from S_2, S_3, and S_4 remain relatively constant. Note that at the 400 ms time scale, in which the three energy basins of the unfolded states are unified, the free energy barrier from the unfolded state increases as an increase of [GdmCl].

Because termination rates differ from state to state and across [GdmCl], some state-to-state transitions do not satisfy the detailed balance condition. For example, in Fig. 7, at 1M [GdmCl], most state-to-state transitions do not violate the detailed balance condition but, at 0.5 M, state-to-state transitions between S_1 and $\{S_3, S_4\}$ do. Thus, we should emphasize that the TRDG representation is an approximation of the underlying free energy landscape of AK.

IV. CONCLUSIONS

Rate-distortion theory is a data-driven, information theoretical method in which state models emerge from the data through segmentation and subsequent clustering of the segments.[16] Soft clustering allows segments to belong to multiple states, a natural treatment in the presence of the photon counting and finite sampling errors of smFRET measurements.

Errors may cause uncertain state assignments, expressed as conditional probabilities of states given segments. Uncertainties are exploited to extract a minimal state model and to construct state sequences that allow the effects of errors to propagate through to any calculated quantities. Because uniform-length segments may contain transitions, we use change-point detection[18] to minimize the number of segments containing transitions.

The method was applied to smFRET trajectories following (un)folding behavior of AK versus denaturant concentration [GdmCl].[5] After change-points were detected, all segments, i.e., regions between change-points, across all [GdmCl] were clustered simultaneously, producing a set of states that is fit globally across multiple data sets. The selected model contained four consistent states in which the most compact form of the protein declined in occupation, while the three less compact forms increased in occupation with increased [GdmCl]. The escape times of the less compact states were relatively independent of [GdmCl] compared to that of the most compact state, suggesting the three less compact states occupy a collective superbasin in which escape times are dominated by fast internal transitions rather than transition to the folded state. Less compact states were expected to have faster termination rates than the higher efficiency states;[5] through the addition of an artificial photobleaching state, we found that in general this is a true assumption, but the existence of multiple routes to trajectory termination plays a major role in the observed termination rate.

We also devised a hypothesis test for the detailed balance condition to assess equilibrium characteristics and found that the degree of violation of detailed balance is not independent of [GdmCl]. Through simulation we showed that an equilibrium system may appear to be an absorbing Markov chain[33] when viewed through an external, coupled window with a finite observation time. Simulation results also indicate that this type of absorbing Markov chain may be modeled as a coupled Markov chain, suggesting that methodologies such as the coupled hidden Markov model[46] may be of use in these situations.

Finally, we applied transition disconnectivity graph methodology[22,23] to construct approximate free energy landscapes of AK folding at multiple time scales. We found merging of the more unfolded states of the protein into an unfolded state superbasin as observation time scale increases. Also we observed the general characteristics of the destabilization of the folded form of the protein as denaturant concentration increases in the forms of increased free energy of the state and decreased barrier heights for transition from the state.

SUPPLEMENTARY MATERIAL

See supplementary material for details involving model selection, construction of TRDGs, a discussion of trajectory termination events, the transition matrices of the most probable sequences, the transition probability matrices, a discussion on the convergence of the transition probabilities, and a discussion and accompanying simulation of absorbing Markov chains.

ACKNOWLEDGMENTS

J.N.T. and T.K. would like to thank Chun-Biu Li, Hiroshi Teramoto, Jason R. Green, and Schuyler B. Nicholson for helpful discussions. This work was supported by Grant-in-Aid for Young Scientists (B) (No. 15K18511) (to J.N.T.), Grant-in-Aid for Scientific Research (B) (No. 17H02940) (to T.K.), Grant-in-Aid for Specially Promoted Research (B) (No. 15KT0055) (to T.K.), Grant-in-Aid for Exploratory Research (No. 16K14703), JSPS (to T.K.), Grant-in-Aid for Scientific Research on Innovative Areas (No. 16H01525), MEXT (to T.K.), and a grant from the Israel Science Foundation (686/14) (to G.H.). The authors also thank the workshop "Macromolecular dynamics: structure, function and diseases," Kavli Institute for Theoretical Physics, Beijing, China, (2014) through which this work was initiated.

[1] M. Tavakoli, J. N. Taylor, C.-B. Li, T. Komatsuzaki, and S. Pressé, *Advances in Chemical Physics* (John Wiley & Sons, Inc., 2017), pp. 205–305.
[2] M. Andrec, R. M. Levy, and D. S. Talaga, J. Phys. Chem. A **107**, 7454 (2003).
[3] S. A. McKinney, C. Joo, and T. Ha, Biophys. J. **91**, 1941 (2006).
[4] M. Pirchi, R. Tsukanov, R. Khamis, T. E. Tomov, Y. Berger, D. C. Khara, H. Volkov, G. Haran, and E. Nir, J. Phys. Chem. B **120**, 13065 (2016).
[5] M. Pirchi, G. Ziv, I. Riven, S. S. Cohen, N. Zohar, Y. Barak, and G. Haran, Nat. Commun. **2**, 493 (2011).
[6] J. E. Bronson, J. Fei, J. M. Hofman, R. L. Gonzalez, Jr., and C. H. Wiggins, Biophys. J. **97**, 3196 (2009).
[7] J.-W. van de Meent, J. E. Bronson, F. Wood, R. L. Gonzalez, and C. H. Wiggins, JMLR Workshop Conf. Proc. **28**, 361 (2013).
[8] I. Sgouralis and S. Pressé, Biophys. J. **112**, 2021 (2017).
[9] I. Sgouralis and S. Pressé, Biophys. J. **112**, 2117 (2017).
[10] C. R. Shalizi and J. P. Crutchfield, J. Stat. Phys. **104**, 817 (2001).
[11] C. J. Ellison, J. R. Mahoney, R. G. James, J. P. Crutchfield, and J. Reichardt, Chaos **21**, 037107 (2011).
[12] A. Baba and T. Komatsuzaki, Proc. Natl. Acad. Sci.U. S. A. **104**, 19297 (2007).
[13] A. Baba and T. Komatsuzaki, Phys. Chem. Chem. Phys. **13**, 1395 (2011).
[14] B. Shuang, D. Cooper, J. N. Taylor, L. Kisley, J. Chen, W. Wang, C. B. Li, T. Komatsuzaki, and C. F. Landes, J. Phys. Chem. Lett. **5**, 3157 (2014).
[15] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[16] J. N. Taylor, C.-B. Li, D. R. Cooper, C. F. Landes, and T. Komatsuzaki, Sci. Rep. **5**, 9174 (2015).
[17] D. J. Wales, *Energy Landscapes* (Cambridge University Press, Cambridge, 2003).
[18] W. A. Taylor, Change-point analysis: A powerful new tool for detecting changes, http://www.variation.com/cpa/tech/changepoint.html, 2000.
[19] D. Montiel, H. Cang, and H. Yang, J. Phys. Chem. B **110**, 19763 (2006).
[20] L. P. Watkins and H. Yang, J. Phys. Chem. B **109**, 617 (2004).
[21] C.-B. Li, H. Ueno, R. Watanabe, H. Noji, and T. Komatsuzaki, Nat. Commun. **6**, 10223 (2015).
[22] S. V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).
[23] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U. S. A. **101**, 14766 (2004).
[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[25] R. E. Blahut, IEEE Trans. Inf. Theory **18**, 460 (1972).
[26] S. Arimoto, IEEE Trans. Inf. Theory **18**, 14 (1972).
[27] L. V. Kantorovich, J. Math. Sci. **133**, 1381 (2006) [Dokl. Akad. Nauk SSSR **37**, 227 (1942)] (in Russian).
[28] N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, Proc. Natl. Acad. Sci. U. S. A. **102**, 18297 (2005).
[29] H. Akaike, IEEE Trans. Autom. Control **19**, 716 (1974).
[30] G. Schwarz, Ann. Stat. **6**, 461 (1978).
[31] J. Rissanen, Automatica **14**, 465 (1978).
[32] B. Efron, Ann. Stat. **7**, 1 (1979).
[33] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, 1st ed. (Springer-Verlag, New York, 1976).
[34] J. G. Skellam, J. R. Stat. Soc. **109**, 296 (1946).
[35] R. K.P. Zia and B. Schmittmann, J. Stat. Mech.: Theory Exp. **2007**, P07012.

[36]T. Platini, Phys. Rev. E **83**, 011119 (2011).
[37]P. Sibani, J. C. Schön, P. Salamon, and J.-O. Andersson, Europhys. Lett. **22**, 479 (1993).
[38]P. Sibani and P. Schriver, Phys. Rev. B **49**, 6667 (1994).
[39]A. Baba and T. Komatsuzaki, *Single Molecule Biophysics* (John Wiley & Sons, Inc., 2011), pp. 299–327.
[40]D. A. Evans and D. J. Wales, J. Chem. Phys. **118**, 3891 (2003).
[41]D. J. Wales and P. Salamon, Proc. Natl. Acad. Sci.U. S. A. **111**, 617 (2014).
[42]I. V. Gopich and A. Szabo, J. Phys. Chem. B **107**, 5058 (2003).
[43]K. Kamagata, T. Kawaguchi, Y. Iwahashi, A. Baba, K. Fujimoto, T. Komatsuzaki, Y. Sambongi, Y. Goto, and S. Takahashi, J. Am. Chem. Soc. **134**, 11525 (2012).
[44]L. R. Ford and D. R. Fulkerson, Can. J. Math. **8**, 399 (1956).
[45]R. E. Gomory and T. C. Hu, J. Soc. Ind. Appl. Math. **9**, 551 (1961).
[46]M. Brand, N. Oliver, and A. Pentland, in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE, 1997), p. 994.