Supporting Information

Fast Step Transition and State Identification (STaSI) for Discrete Single-Molecule Data Analysis

Bo Shuang, [†] *David Cooper,* [†] *J. Nick Taylor,*[‡] *Lydia Kisley,* [†] *Jixin Chen,* [†] *Wenxiao Wang,*[§] *Chun Biu Li,*[‡] *Tamiki Komatsuzaki,*[‡] *Christy F. Landes*[*, †,§]

† Department of Chemistry, Rice University, MS 60, Houston, Texas 77251-1892, United States

‡ Molecule & Life Nonlinear Sciences Laboratory, Research Institute for Electronic Science, Hokkaido University, Sapporo 001-0020, Japan

§ Department of Electrical and Computer Engineering, Rice Quantum Institute, Rice University, MS 60, Houston, Texas 77251-1892, United States

AUTHOR INFORMATION

Corresponding Author

*Phone: 713-348-4232. E-mail: cflandes@rice.edu.

Contents

1. Derivation of function *G* in the MDL expression for piecewise constant signals:

The function G in eq 3, defined[1] below, is derived for piecewise constant signals as follows:

$$G = \frac{k}{2}\ln\frac{N}{2\pi} + \ln\int_{Vol} d^{k+N_{\mathrm{tp}}}\theta\sqrt{\det\overleftrightarrow{I}(\theta)} \tag{S1}$$

where k is the number of states and N_{tp} is the number of transitions (both the value of each state and the position of each transition are unknown parameters to be determined, represented by θ altogether); N is the total number of data points; Vol accounts for the entire parameter space; $I(\theta)$ is the $(k + N_{\mathrm{tp}}) \times (k + N_{\mathrm{tp}})$ Fisher information matrix of all the parameters of the k states and N_{tp} transitions averaged over the entire data set; and det $I(\theta)$ is the determinant of the matrix $I(\theta)$. The first term in the right hand side accounts for the penalty of using more states, which is similar to Bayesian information criterion. The second term measures the internal complexity of the fitting model. The matrix element $I_{ij}(\theta)$ is defined as[1]

$$\overleftrightarrow{I}_{ij}(\theta) \stackrel{\text{def}}{=} \int p(y|\theta)^{-1}\,\partial_i p(y|\theta)\partial_j p(y|\theta)dy \tag{S2}$$

where $p(y|\theta)$ is the probability density function for y conditional on the value of parameters θ, and ∂_i is the partial deviation with respect to i^{th} parameter θ_i. In general one assumes the functional form of $p(y|\theta)$ by referring the physical rule behind experiments of interest. In this paper, the Fisher information matrix is derived under the assumption of a Gaussian distributed noise model (considering the central limit theorem) and the detailed derivation can be found in Hanson and Fu's work.[1] In this work, each matrix element is:

$$\overleftrightarrow{I}_{ij}(\theta) = \begin{cases} \frac{1}{N\sigma^2} n_i\delta_{ij}, & i,j \in [1,k] \\ \frac{1}{N\sigma^2} T_j^2\delta_{ij}, & i,j \in [k+1, k+N_{\mathrm{tp}}] \\ 0, & \text{otherwise} \end{cases} \tag{S3}$$

Here σ is the noise level estimated as overall noise level independent of each state value, δ_{ij} is Kronecker delta, and T_j is the difference of the fitting values before and after the transition position j. The parameter space for each parameter θ_i is:

$$Vol = \begin{cases} y_{\max} - y_{\min}, & \text{for } i \in [1, k] \\ N, & \text{for } i \in [k+1, k+N_{\mathrm{tp}}] \end{cases} \quad \text{(S4)}$$

where $y_{max(min)}$ means the maximum (minimum) value among the entire values of y. We use V to represent ($y_{max} - y_{min}$) to be consistent with eq 3. So, the second term of G in the right hand side of eq S1 is:

$$\ln \int_{Vol} d^{k+N_{\mathrm{tp}}}\theta \sqrt{\det \overleftrightarrow{I}(\theta)} = \ln \int_V d^k\theta \sqrt{\left(\frac{1}{N\sigma^2}\right)^k} \sqrt{\prod_{i=1}^k n_i} + \ln \int_N d^{N_{\mathrm{tp}}}\theta \sqrt{\left(\frac{1}{N\sigma^2}\right)^{N_{\mathrm{tp}}}} \sqrt{\prod_{j=1}^{N_{\mathrm{tp}}} T_j^2}$$

$$= -\frac{k+N_{\mathrm{tp}}}{2}\ln(N\sigma^2) + \sum_{i=1}^k \ln\left(V\sqrt{n_i}\right) + \sum_{j=1}^{N_{\mathrm{tp}}} \ln\left(N\sqrt{T_j^2}\right)$$

$$= -\frac{k+N_{\mathrm{tp}}}{2}\ln(N\sigma^2) + k\ln V + N_{\mathrm{tp}}\ln N + \frac{1}{2}\left(\sum_{i=1}^k \ln n_i + \sum_{j=1}^{N_{\mathrm{tp}}} \ln T_j^2\right) \quad \text{(S5)}$$

Therefore:

$$G = \frac{k}{2}\ln\frac{N}{2\pi} - \frac{k+N_{\mathrm{tp}}}{2}\ln(N\sigma^2) + k\ln V + N_{\mathrm{tp}}\ln N + \frac{1}{2}\left(\sum_{i=1}^k \ln n_i + \sum_{j=1}^{N_{\mathrm{tp}}} \ln T_j^2\right)$$

$$= \frac{k}{2}\ln\frac{1}{2\pi} + k\ln\frac{V}{\sigma} + \frac{N_{\mathrm{tp}}}{2}\ln N + \frac{1}{2}\left(\sum_{i=1}^k \ln n_i + \sum_{j=1}^{N_{\mathrm{tp}}} \ln\frac{T_j^2}{\sigma^2}\right) \quad \text{(S6)}$$

2. The Student's *t* test to detect step transitions:

The Student's *t* test with unequal sample size and global noise level is applied to each single trace iteratively to detect all of the step transitions. The Student's *t* test calculates the absolute

difference of the averaged intensity before and after point t_i over their combined uncertainty, as shown in eq S7, the one-tailed *t* test statistic:

$$R(t_i) = \frac{|I_2(t_{i+1},t_N) - I_1(t_1,t_i)|}{\sigma\sqrt{\frac{1}{i}+\frac{1}{N-i}}} \quad \text{(S7)}$$

where σ is defined as the noise level of the entire trace, calculated as $1/2^{0.5}$ of the absolute value of the first order Haar wavelet of the signal at 68.2% of the cumulative distribution (see section 4 in online methods for more details); $I_1(t_i, t_i)$ is the averaged intensity from t_1 to t_i, and the uncertainty of I_1 is $\sigma/i^{0.5}$; similarly, $I_2(t_{i+1}, t_N)$ is the averaged intensity from t_{i+1} to t_N and the uncertainty of I_2 is $\sigma/(N-i)^{0.5}$ (N is the total number of the data points), and their combined uncertainty is $\sigma(1/i+1/(N-1))^{0.5}$. During each iteration, the point having the highest *t* test value larger than a universal threshold (3.174, corresponding to t-distribution value with 99.8% confidence for 100 variables) is considered a transition point. This threshold value keeps the false positive rate smaller than 90% and also decreases the false negative rate for detection of short-lived transitions (important for noisy data with fast dynamics). From the identified transitions, the trace is broken down into multiple segments (Figure 1a). A *segment* is defined as a successive section of data points between two step transitions. For example, the second segment in the two segments plot in Figure 1a is broken down into two segments as highlighted by the black arrow. This process is repeated on every new segment and terminated when no further transitions are identified (the example in Figure 1a terminated at five segments). To capture fast but rare transitions, we introduce another parameter *counter* to force the iteration process to continue several more steps (specified by *counter*, default is 3) even though the termination condition is met. More details are included in section 7.

3. State grouping algorithm:

After all of the transition points have been identified, the final segments are assigned to different states based on their average values and a greedy algorithm is applied in each iteration to find the two most similar states and group them into a single state.[2] To do this, the log likelihood merit (M in eq S8) of each two-segment pairing in the remaining n states is calculated:

$$M(i,j) = L_{n-1}(y_t, I_1, \dots, I_{i,j}, \dots, I_{n-1}) - L_n(y_t, I_1, \dots, I_i, \dots, I_j, \dots, I_n)$$

$$= \ln \prod_{t\in i,j} \exp\left(-\frac{(y_t - I_{i,j})^2}{\sigma^2}\right) - \ln \prod_{t\in i} \exp\left(-\frac{(y_t - I_i)^2}{\sigma^2}\right) - \ln \prod_{t\in j} \exp\left(-\frac{(y_t - I_j)^2}{\sigma^2}\right)$$

$$\propto (m_i + m_j) * {I_{i,j}}^2 - (m_i * I_i^2 + m_j * I_j^2) \qquad \text{(S8)}$$

This equation shows the log likelihood merit of clustering two states i and j. In the equation, L_{n-1} and L_n are the log likelihood estimation for n-1 and n states respectively; y_t represents the measured value at time t; m_i and m_j are the number of data points for states i and j, and similarly, I_i is the averaged value for state i, and $I_{i,j}$ is the averaged value if states i and j are clustered together. The two states corresponding to the maximum $M(i,j)$ are then clustered into a single state. This greedy strategy is applied iteratively until only one state remains. This process is illustrated in Figure 1b. In the five states plot, all of the five states have different average FRET efficiencies. In each iteration, two states are clustered into a single state, as highlighted by the black arrows in Figure 1b. At the conclusion of this process, the best grouping strategy for every possible number of states has been calculated and all that is left is to decide which number of states best describes the data using the MDL equation for compressed sensing.

4. Calculating the noise level using the Haar wavelet transform:

For a piecewise constant signal, the average signal changes after each transition. The noise level of a piecewise constant signal therefore cannot be directly calculated as the standard deviation of the signal. The Haar wavelet transform of the signal at the lowest scale (or highest frequency, which is basically taking the differences of adjacent signal values), which we call $w1$ hereafter, mostly captures the fluctuations of the noise due to the memoryless property of the noise (with the assumption that number of transitions is much smaller than the number of data points). The standard deviation of the noise can be estimated as $1/2^{0.5}$ of the standard deviation of $w1$. To avoid the biases of large transitions, we use the cumulative distribution of the absolute value of $w1$. The noise level then corresponds to 68.2% of the cumulative distribution, which is the standard deviation for a cumulative Gaussian distribution.

In StaSI, the noise level is used both in step detection in eq S7 and state determination in the MDL equation. Our algorithm is designed to capture all of the transitions that may be relatively small. For systems where only relatively large transitions are important (such as systems that can be described by binary states), using a reasonably large noise level that is comparable to these transitions will identify the steps.

5. Comparison between the L1 norm and the L2 norm:

Example traces under several conditions are shown in Figure S1-3. For a smaller noise level (0.03) and long mean-lifetime (2.5 s), both the L1 norm and the L2 norm identifies the correct number of states analyzing raw data (Figure S1a,b). When analyzing binned data, STaSI using the L1 norm identifies the correct number of states, but nine redundant states are identified when using the L2 norm (seven of the small populations are indicated by black arrows, the other two are very close to the real states at 0.5 and 0.35) (Figure S1c,d). For data with a relatively large noise level (0.12) and short mean lifetime of the states (0.25 s), STaSI using the L1 norm identifies the correct number of states when analyzing both the raw data and binned data, but STaSI using the L2 norm identifies two redundant states when analyzing raw data and identifies more than 20 states when analyzing binned data (Figure S2). As an extreme example, Figure S3 shows the analysis of raw data with high noise level (0.12) and very short mean lifetime of the states (0.025 s). Both the L1 norm and the L2 norm fail, but the state distribution determined using the L1 norm is much closer to the true distribution.

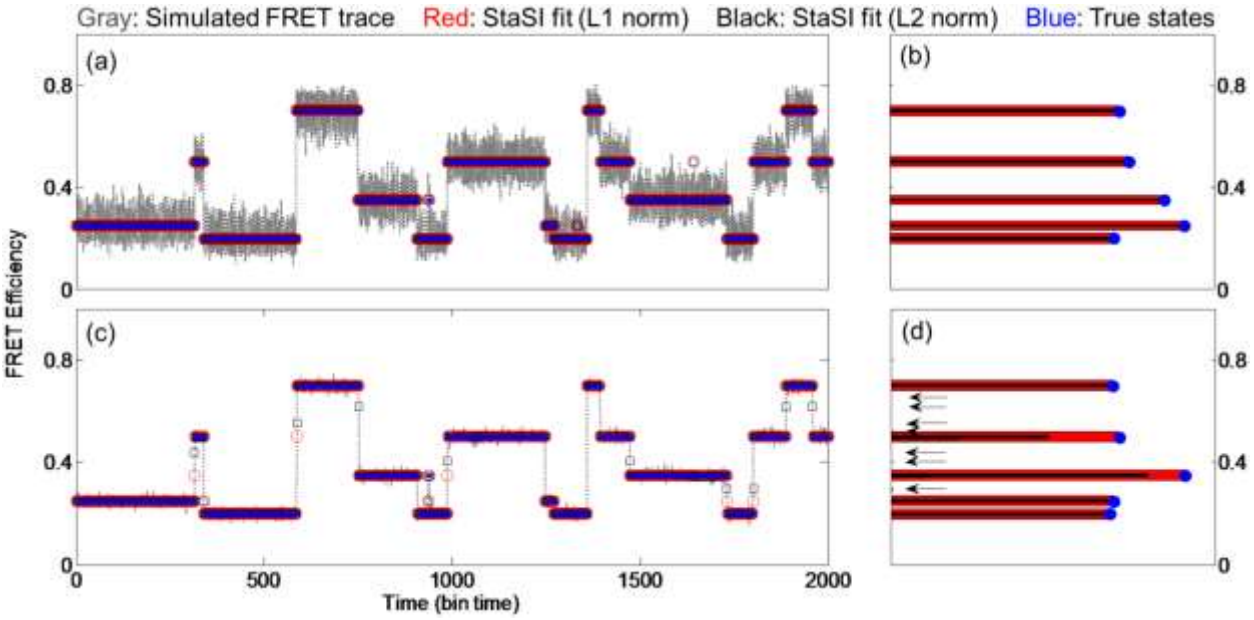

Figure S1. Typical example comparison between the L1 norm and the L2 norm for data with noise level = 0.03 and mean lifetime of the states = 2.5 s. (a) Simulated raw data analyzed by STaSI using the L1 norm and the L2 norm. (b) The corresponding histograms of the analyzed states using the L1 norm and the L2 norm, and the histogram of the true states of raw data. (c) Simulated 10 ms binned data analyzed by STaSI using the L1 norm and the L2 norm. (d) The corresponding histograms of the analyzed states using the L1 norm and the L2 norm, and the histogram of the true states of binned data. The black arrows indicate redundant states with very small populations due to binning.

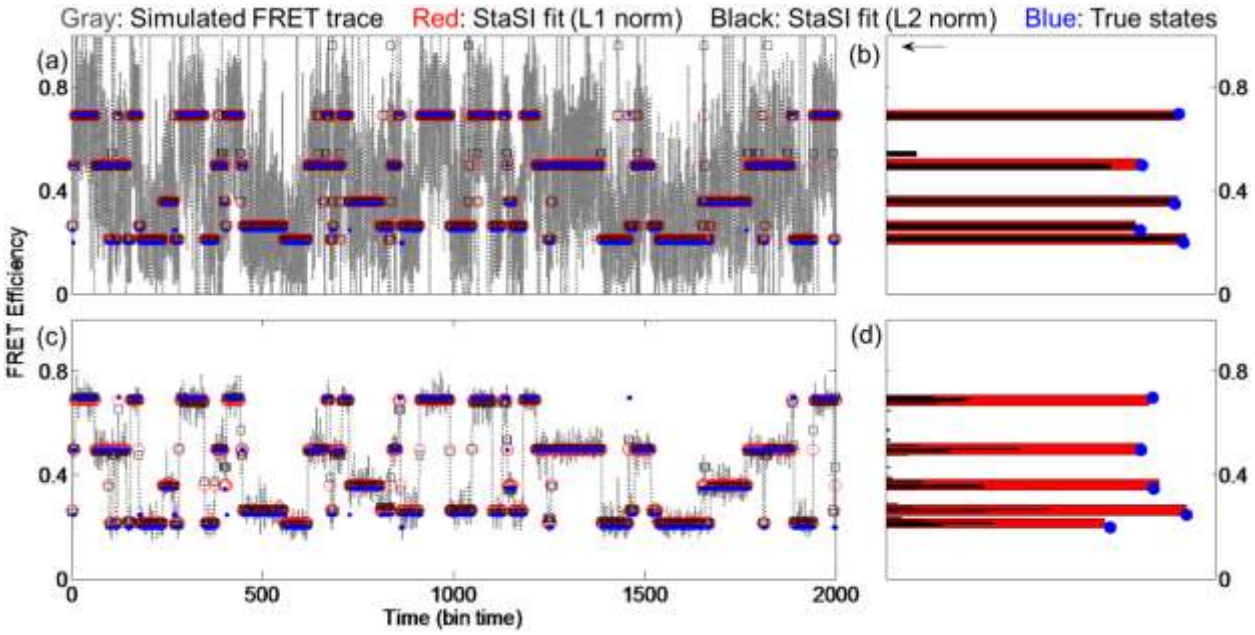

Figure S2. Typical example comparison between the L1 norm and the L2 norm under noise level = 0.12 and mean lifetime of the states = 0.25 s. (a) Simulated raw data analyzed by STaSI using the L1 norm and the L2 norm. (b) The corresponding histograms of the analyzed states using the L1 norm and the L2 norm, and the histogram of the true states of raw data. The black arrow indicates a redundant state with a very small population due to noise. (c) Simulated 10 ms binned data analyzed by STaSI using the L1 norm and the L2 norm. (d) The corresponding histograms of the analyzed states using the L1 norm and the L2 norm, and the histogram of the true states of binned data.

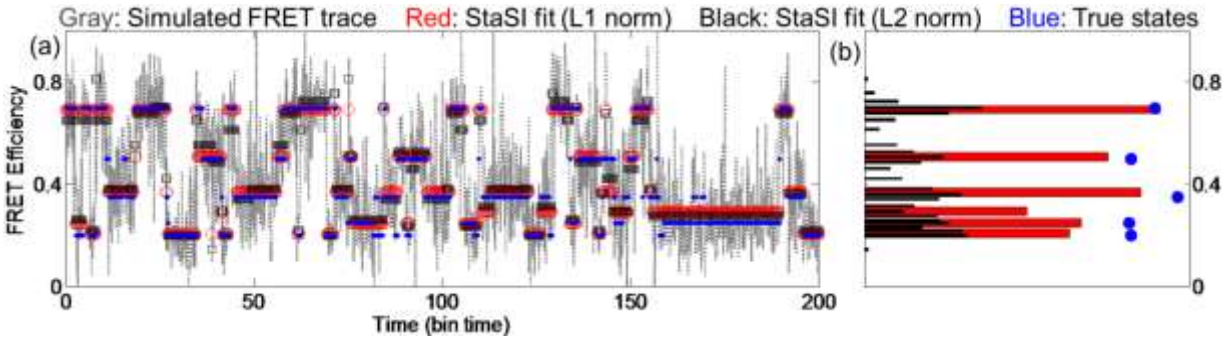

Figure S3. Typical example comparison between the L1 norm and the L2 norm under noise level = 0.12 and mean-lifetime = 0.025 s. (a) Simulated raw data analyzed by STaSI using the L1 norm and the L2 norm. (b) The corresponding histograms of the analyzed states using the L1 norm and the L2 norm, and the histogram of the true states of raw data.

6. Speed comparison:

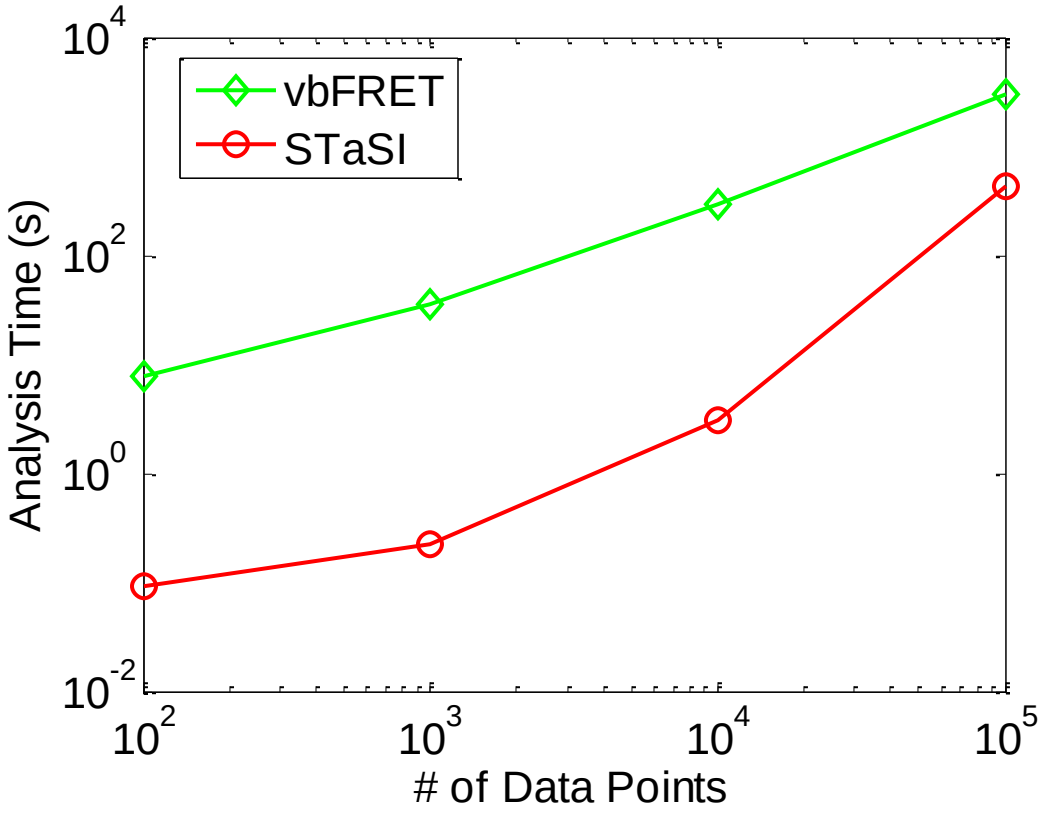

Figure S4. Speed comparison between STaSI and vbFRET.[3] For a FRET trace with the number of data points smaller than 10^5. STaSI is more than ten times faster than vbFRET within this range of data length. The complexity of the Student's *t* test (step 1) is O(Nlog(N)), the complexity of the grouping algorithm (step 2) is O(N^2), and the complexity of the *MDL* calculation is O(N). Therefore, the overall complexity of STaSI is O(N^2). The test is conducted using MATLAB R2013a on a personal computer with an Intel i7, 3.40 GHz processor. For vbFRET, the number of possible states is set to be one to ten, and the fitting attempt per trace is set to be ten.

7. Identifying short-lived state segments:

In the step transition identification process, the detection of the transitions terminates when no further transitions are detected in any remaining segment. To identify short-lived state segments that may be overlooked, each terminated segment (with more than one data point) is broken down into two segments separated by the data point in the middle, and the *t* test is applied to both segments to search for the transitions. A short-lived state segment in a relative smaller segment is easily detected by the *t* test. This process is iterated multiple times in each segment before the final termination. The number of iterations is controlled by a parameter called *counter* in our function. The effect of using this parameter is depicted in Figure S5. Usually, *counter* = 3 to 5 is good enough to identify short-lived state segments 100 times shorter than the mother segment (Figure S5). The default value of *counter* is 3 in this work.

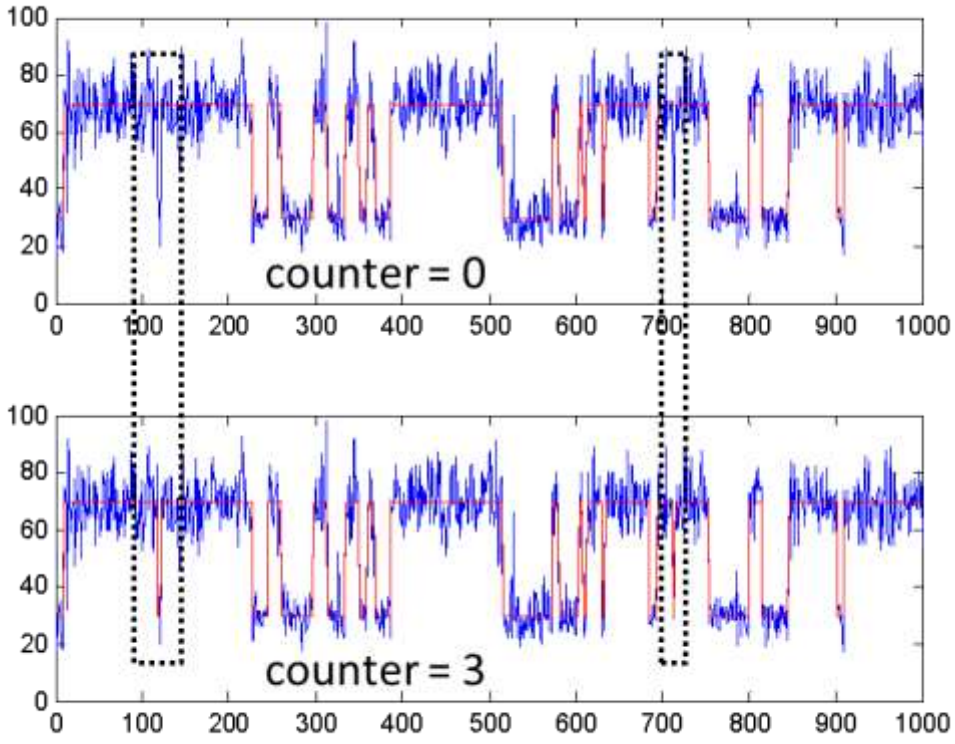

Figure S5. The effect of using *counter* to capture the short-lived transitions. Both figures use the same simulated piecewise constant signal with Poisson noise. The simulated signal is shown in blue and the fit is shown in red. As highlighted by the dashed black boxes, using *counter* = 3 is more likely to capture short-lived state segments compared to using *counter* = 0.

8. FRET trajectory simulation:

The FRET trajectories are simulated with a Monte Carlo method using several parameters. The key parameters include the number of states, FRET efficiency of the states, transition rate between each two states, noise levels, total simulation time, simulation step time, bin time. The simulation is carried out in two steps as following.

Major parameters for the simulation are listed in Figure S6, including the number of states *n*, the transition rate constant k_{ij}, and the FRET values. The molecule is randomly assigned to have an initial state. At each sampling time dt (1 ms), the transition probability is:

$$P_i = 1 - e^{\sum_{j \neq i} k_{ij} dt}$$

If the molecule is confirmed to change states at a certain step, the transition probability to each state is set to be proportional to the corresponding rate constant. After the simulation, each state is assigned to the corresponding pre-set FRET efficiency.

Photon counts in the donor and acceptor channels are calculated based on the FRET efficiency at that time, with total photon counts set to be 20 in 1 ms. For each simulated step, noise is added to both the donor acceptor channels. The noise follows Gaussian distribution and the standard deviation of the Gaussian distribution is proportional to the photon counts for both channels. The slope of this proportional relation can be controlled to generate traces with different noise levels. For binned data, the two channels are averaged at every 10-ms time window separately, and the final FRET value at each binned time step is calculated from the binned photon counts of the two channels.

$$FRET = \frac{I_{acceptor}}{I_{acceptor} + I_{donor}}$$

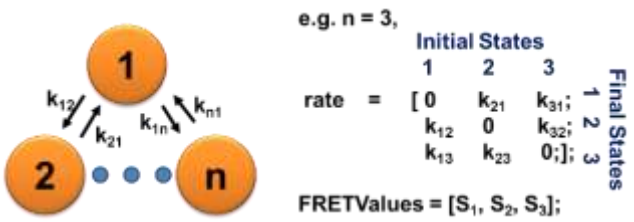

Figure S6. Input parameters for the Monte Carlo simulation. Assuming transitions happen between any two of the n states with transition rate constants $k_{ij} > 0$ (s^{-1}). FRET values S_i are assigned to the i^{th} state after simulation.

REFERENCES

(1) Hanson, A. J.; Fu, P. C.-W. Application of MDL to Selected Families of Models. In *Advances in Minimum Description Length: Theory and Applications (Neural Information Processing)*; Grünwald, P. D., Myung, I. J., Pitt, M. A., Eds.; MIT Press: Cambridge, MA, 2005.

(2) Watkins, L. P.; Yang, H. Detection of intensity change points in time-resolved single-molecule measurements. *J. Phys. Chem. B* **2005**, *109*, 617-628.

(3) Bronson, J. E.; Fei, J.; Hofman, J. M.; Gonzalez, R. L., Jr.; Wiggins, C. H. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* **2009**, *97*, 3196-3205.