Extracting Kinetics Information from Single-Molecule Fluorescence Resonance Energy Transfer Data Using Hidden Markov Models

Tae-Hee Lee

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802

Received: April 26, 2009; Revised Manuscript Received: June 22, 2009

Hidden Markov models (HMM) have been proposed as a method of analysis for noisy single-molecule fluorescence resonance energy transfer (SM FRET) data. However, there are practical and fundamental limits in applying HMM to SM FRET data due to the short photobleaching lifetimes of fluorophores and the limited time resolution of detection devices. The fast photobleaching fluorophores yield short SM FRET time traces, and the limited detection time resolution generates abnormal FRET values, which result in systematic underestimation of kinetic rates. In this work, a HMM algorithm is implemented to optimize one set of HMM parameters with multiple short SM FRET traces. The FRET efficiency distribution function for the HMM optimization was modified to accommodate the abnormal FRET values resulting from limited detection time resolution. Computer simulations reveal that one set of HMM parameters is optimized successfully using multiple short SM FRET traces and that the degree of the kinetic rate underestimation is reduced by using the proposed modified FRET efficiency distribution. In conclusion, it is demonstrated that HMM can be used to reproducibly analyze short SM FRET time traces.

Introduction

Single-molecule fluorescence resonance energy transfer (SM FRET) is a powerful tool that can probe subpopulation dynamics of complex biological processes[1,2] involving DNA,[3] RNA,[4,5] proteins,[6,7] and macromolecular assemblies.[8] Monitoring dynamic single molecules in real time has generated information previously unavailable with static or bulk methods.[3,7,8] Analyses of SM FRET data rely mostly on simple threshold discrimination. Although threshold discrimination works relatively well on data with a high signal-to-noise ratio (SNR), it suffers large errors and uncertainty with a typical experimental SNR which ranges from 5 to 10.

The Hidden Markov model (HMM) is a finite state machine defined by an observation sequence (O) and a model (λ) comprising a transition matrix defining transition probabilities between states (a) with single exponential lifetimes, emission probabilities (b) of states to map the observations to the hidden events, and the initial state (π).[9] The prerequisites for applying HMM to an experimental system are (i) the system must dwell on finite states, each of which can be observed directly or indirectly with certain errors and (ii) the conditional probability distribution of future states of the system depends only on the current state, that is, transition probabilities between two states can be defined by a single value. On the basis of HMM, one can calculate the probability of a future event with a past observation sequence.[9,10] The probability of obtaining an observation sequence O with a model λ is represented by $P(O|\lambda)$, where $\lambda = \{a, b, \pi\}$. HMM model parameters incorporate all of the information on the kinetics of a system. One can use Viterbi's algorithm to find a hidden sequence of states emitting the observation sequence.[9,11,12] Baum−Welch's iteration method or gradient techniques can be used to find the optimum model parameters for a given observation sequence.[9,10] HMM has been utilized to analyze single ion channel dynamics and motor protein dynamics.[13−15] Although SM FRET data from many enzymatic processes are good targets for HMM, HMM optimization for SM FRET data was implemented only recently with limitations.[16]

There are fundamental and practical limitations when applying HMM to SM FRET signals. First, a longer signal integration time than the event duration yields artifacts in the signal,[17] that is, short lifetime events register lower or higher FRET values than normal, which can be seen as either a different state or noise (Figure 1). Second, the unsynchronized detection to enzyme dynamics also causes artifacts (Figure 1). The first and the last detection frames of a single FRET event include only a partial frame event because the enzyme dynamics is not synchronized to detection frames. Transitions between two states, therefore, generally leave a small population of FRET events between two FRET peaks (Figure 1). Short lifetime events elongate the detected lifetime of a FRET state, and unsynchronized detection shortens it. A formula to fit FRET distribution histograms with these artifacts has already been reported.[17] However, the reported formula yields an analytical solution only in the case of a two-state model. Moreover, the solution takes an unfeasibly long time to be employed in a HMM optimization algorithm, where the probability distribution of a state is typically calculated a million times or more to optimize a reasonable amount of experimental data. Lastly, due to the limited photobleaching lifetimes of conventional dyes, SM FRET traces in many experiments are short fragments, each of which contains only a portion of all possible transitions between states. Therefore, individually optimized HMM parameters per individual trace contain partial information. Recently, it was shown that the average of the logarithm of individual transition matrices can represent the universal transition matrix in some cases.[16] For another instance, computer simulations reveal that a Winsorized mean of the lower 70% of transition matrices can approximate the representative universal transition matrix fairly well in some random cases (data not shown). However, all of the averaging methods yield an unknown level of uncertainty due to the empirically determined weights on individual transition matrices. In order to address these three problems,

10.1021/jp903831z CCC: $40.75
Published on Web 07/24/2009

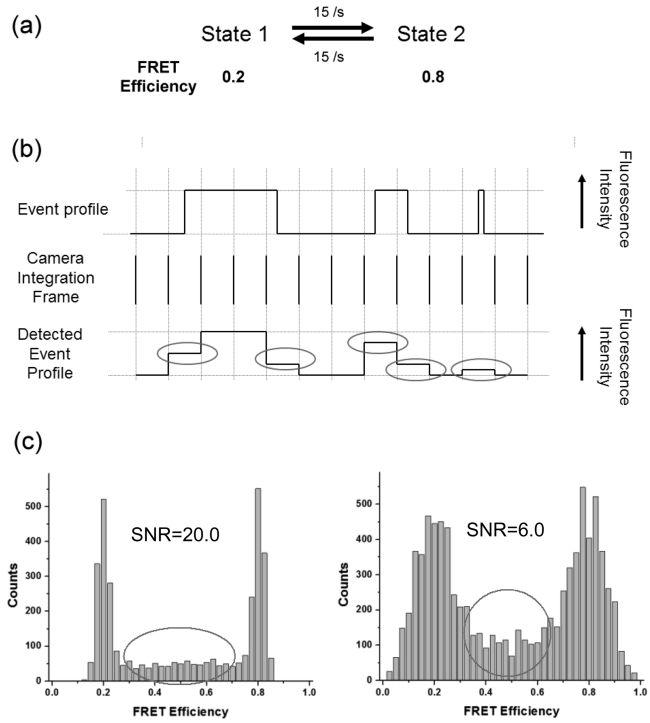


Figure 1. The effect of short lifetime events and the unsynchronized detection on FRET efficiency distribution. (a) Kinetic scheme of the simulated traces. Two sets of SNR (20.0 and 6.0) were simulated. (b) Illustration of real events and detected events showing examples of short lifetime events and the unsynchronized events to the detector time bin. The first and the last detected frames of the first long lifetime event do not have the same fluorescence intensity level as the rest of the frames in the middle. The second event with a lifetime similar to the detection integration time also registers two frames with a lower fluorescence level than normal. The third short lifetime event will register a single frame with a lower fluorescence intensity level than normal. The gray ellipses indicate these abnormal low fluorescence intensities resulting from either the unsynchronized detection or short lifetime events. (c) SM FRET histograms from simulated traces with the kinetic scheme in (a). As SNR becomes higher, randomly scattered FRET efficiency counts between the two FRET peaks become evident (counts in gray ellipses). These counts are due to short lifetime events and the unsynchronized detection.

algorithms of HMM with a modified FRET efficiency distribution and a combined probability of multiple observation sequences were implemented.

Experimental Methods

HMM Model Parameter Optimization. In order to extract the kinetic scheme from SM FRET traces, HMM parameters were optimized with the given set of FRET traces. Baum−Welch's iteration algorithm was used to perform the optimization.[9] A technical problem of underflow in probabilities can be easily fixed with the known rescaling procedure.[9] Equations 1 and 2 are the re-estimation formulas for the transition matrix a and the initial state π. For emission probabilities b, continuous observation densities were used to avoid any artifacts arising from digitizing FRET traces.[9] Observation density distributions of SM FRET traces were assumed to be Gaussian, which is widely used in fitting SM FRET histograms.[18] To consider different background fluorescence intensities and slight shifts in FRET efficiencies due to environmental heterogeneity, multiple Gaussian distributions per state were used. The re-estimation formula for the emission probabilities, then, is given as in eq 3. For the re-estimation formulas of μ_j and σ_j, one can follow the procedure for the maximum likelihood estimation of multivariate mixture observation as reported.[9,19]

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{1}$$

$$\pi_i = \gamma_1(i) \tag{2}$$

where T is the number of time points in the trace, $\sum_{t=1}^{T-1} \xi_t(i,j)$ is the expected number of transitions from state i to state j, and $\sum_{t=1}^{T-1} \gamma_t(i)$ is the expected number of transitions from state i.

$$b_i(O) = \sum_{j=1}^{m} c_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(O-\mu_j)^2}{2\sigma_j^2}\right) \tag{3}$$

where O is the observation, m is the number of Gaussian distributions per state, μ_j is the peak position of the jth Gaussian component of state i, and σ_j is the width of the jth Gaussian distribution of state i. To accommodate the scattered FRET efficiencies between peaks (Figure 1), one more asymmetric Gaussian distribution is added to eq 3. The Gaussian component is approximated to

$$\sum_j \frac{2c_j}{n_j+1} \sum_m \frac{k_m}{k_{\text{total}}} \frac{2}{\sqrt{2\pi}\left(\sigma_j + \left(\frac{|\mu_j-\mu_m|}{3}\right)\right)} \times \left(\exp\left(-\frac{(O-\mu_j)^2}{2\sigma_j^2}\right) \quad \text{or} \quad \exp\left(-\frac{(O-\mu_j)^2}{2\left(\frac{\mu_j-\mu_m}{3}\right)^2}\right)\right) \tag{4}$$

and then, that for the main peak is normalized to

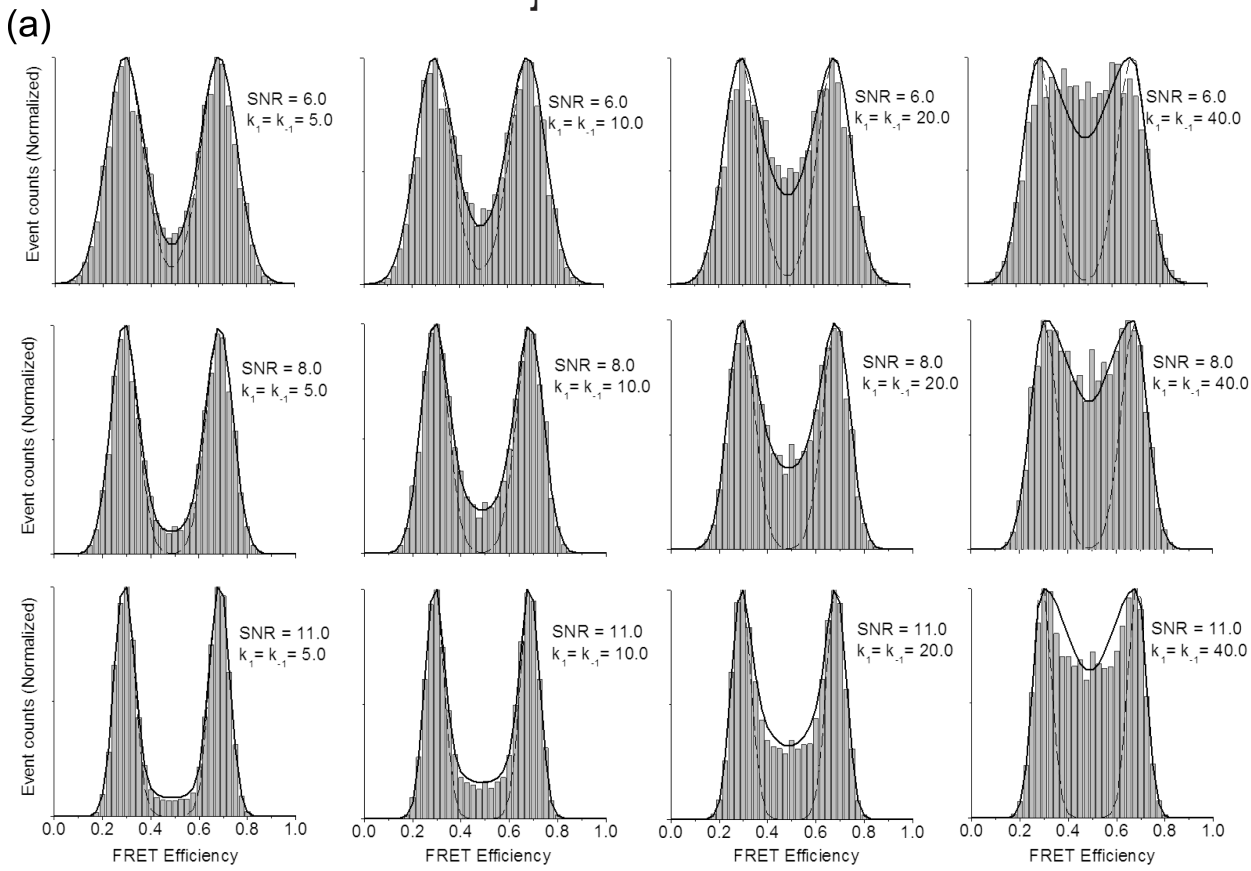
$$\sum_j \frac{(n_j-1)c_j}{n_j+1} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(O-\mu_j)^2}{2\sigma_j^2}\right) \tag{5}$$

where k are the rate constants defining rates out of the state j and n is the biggest integer smaller than the average number of consecutive data points for the state (e.g., average duration of the state in terms of signal frames). The first exponential term in eq 4 is applied when O is not related to state m, while the second term is applied when O falls between states j and m. These two equations are valid only when the state lifetime is equal to or longer than the signal integration time. The denominator 3 in the width of the new Gaussian in eq 4 is chosen to have negligible probability of one FRET state j beyond the other FRET state m (<0.27%) while there is still significant FRET distribution between the FRET peaks. It is confirmed by HMM optimization that a denominator of 3 works best among 2, 3, and 4 (data not shown). The final formula for b is then as follows.

$$b_i(O) = \sum_j \frac{C_j}{n_j + 1}\left[\sum_m \frac{k_m}{k_{\text{total}}} \frac{4}{\sqrt{2\pi\left(\sigma + \left(\frac{|\mu_j - \mu_m|}{3}\right)\right)}} \times \left(\exp\left(-\frac{(O-\mu_j)^2}{2\sigma_j^2}\right) \quad \text{or} \quad \exp\left(-\frac{(O-\mu_j)^2}{2\left(\frac{\mu_j - \mu_m}{3}\right)^2}\right)\right) + \frac{(n_j - 1)}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(O-\mu_j)^2}{2\sigma_j^2}\right)\right] \quad (6)$$

A straight line between the FRET peaks convolved with Gaussian distributions is found to yield less accurate results with a significantly longer optimization time than the asymmetric Gaussian distribution.

In addition to the above modifications in the FRET efficiency distribution, a single transition matrix and a single set of emission probabilities are used to maximize the total probability of individual $P(O|\lambda)$, that is, $\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})$, instead of optimizing $P(O|\lambda)$ of individual traces, where l is the index of individual SM FRET traces of which the total number is n.



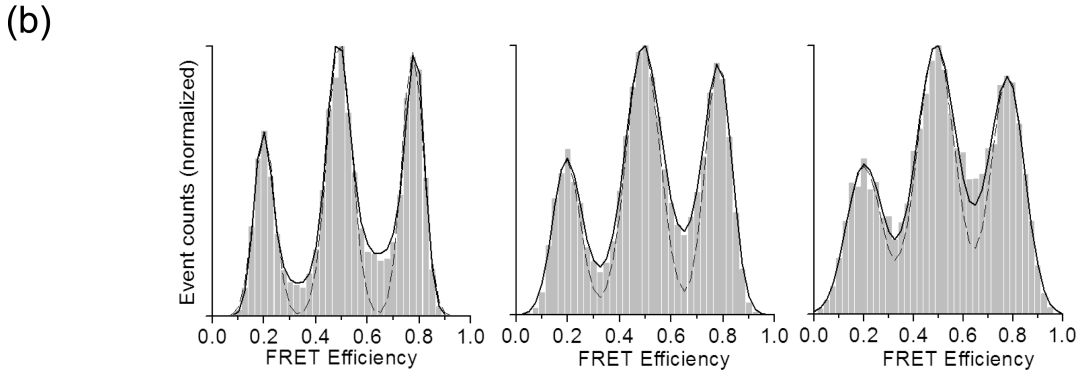


Figure 2. The unmodified (eq 3) and modified (eq 6) Gaussian distributions as an approximated FRET efficiency distribution. (a) The two distribution functions are used to fit histograms constructed from simulated SM FRET traces with a two-state model (solid line, fit using eq 6; dotted line, fit using eq 3). Each panel represents a case of 100 SM FRET traces (500 time points per trace). FRET efficiencies of the two states are 0.3 and 0.7. SNR is controlled by changing the photon emission rate. The kinetic rates between the two states are also varied as labeled on each panel. The signal integration time is 25 ms (observation frame rate = 40/s). The SNR range (6.0, 8.0, and 11.0) is chosen to simulate experimental data with reasonable quality attainable in a laboratory. (b) A three-state model with two different SNR values fit with the two distribution functions (solid line, fit using eq 6; dotted line, fit using eq 3). Each panel represents a case of 100 SM FRET traces of 500 time points. FRET efficiencies for the three states are 0.2, 0.5, and 0.8. Kinetic rates between the three states are $k_1 = 10.0$, $k_{-1} = 5.0$, $k_2 = 10.0$, and $k_{-2} = 12.0$. The SNR for the left panel is 11.0 and 8.0 for the center panel and 6.0 for the right panel.

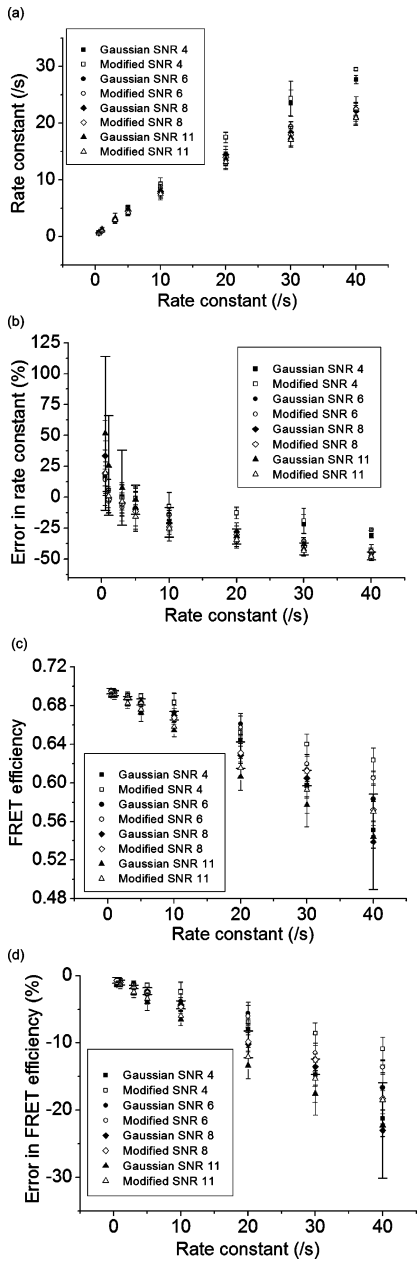


Figure 3. The effect of the rate constant on the HMM optimization performance of the two probability distribution functions (eqs 3 and 6). Five sets of 500 simulated traces with 350 time points per trace were optimized to give one data point with an error bar. The system is composed of two FRET states with FRET efficiencies of 0.3 and 0.7. The rate going from the 0.3 to 0.7 state is fixed at 0.5/s, and the backward rate is varied (0.5, 1.0, 3.0, 5.0, 10.0, 20.0, 30.0, and 40.0/s). The signal integration time is 25 ms (observation frame rate = 40/s). The SNR (i.e., photon emission rate from the fluorophore) was varied in order to examine its effect on the results. A data label starting with "Gaussian" indicates that the optimization is performed with eq 3. "Modified" is used if the optimization was performed with eq 6. (a) Optimized backward rates plotted against the given rate constants. (b) The error in the optimized backward rate constants plotted against the given rate constants. (c) Optimized FRET efficiency of the 0.7 FRET state plotted against the backward rates. (d) Errors in the optimized FRET efficiency of the 0.7 FRET state plotted against the backward rates.

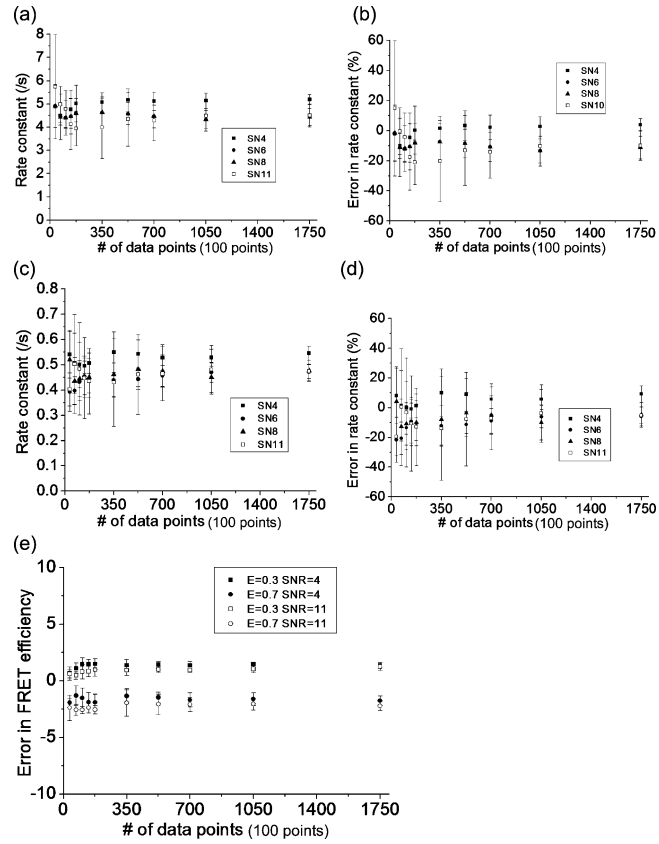


Figure 4. The effect of the amount of data on the performance of HMM optimization. The simulation conditions are the same as those in Figure 3, except that the optimization was performed using only the modified distribution function (eq 6), and the kinetic rates are fixed at 0.5 and 5.0/s, respectively for the forward and the backward rates. The amount of data used in the optimization was varied (10, 20, 30, 40, 50, 100, 150, 200, 300, and 500 traces, which are equivalent to 3500, 7000, 10500, 14000, 17500, 35000, 52500, 70000, 105000, and 175000 data points), and the SNR was also varied (4.0, 6.0, 8.0, and 11.0). "SNR" is further abbreviated to "SN" in (a−d). (a) The optimized backward rates (5.0/s) plotted against the number of data points used in the optimization. (b) The errors in the optimized backward rates plotted against the number of data points. (c) The optimized forward rates (0.5/s) plotted against the number of data points used in the optimization. (d) The errors in the forward rates plotted against the number of data points. (e) Errors in the optimized FRET efficiency levels plotted against the number of data points used in the optimization.

Rabiner's re-estimation formulas for multiple observation sequences are used with unit weighting instead of P^{-1} weighting.[9] It is more logical to use unit weighting for SM FRET data because a mere number of time points in a trace does not necessarily increase the information content of the trace. The number of transitions can better represent the amount of information contained in a trace. Therefore, P^{-1} weighting in cases where many time points are steady instead of dynamic, as in SM FRET, is inappropriate. One optimization of the HMM model parameters generally takes several tens of seconds to several hours depending on the number of Gaussian mixtures and the total length of SM FRET traces, but it rarely exceeds an hour with a practical amount of data and a reasonable number of Gaussian distributions per state (<5) on a Windows system (Microsoft Corp., U.S.A.) with a Pentium 4 processor (Intel Corp., U.S.A.) or on a Linux system with a Pentium D processor (Intel Corp., U.S.A.). The algorithm is implemented in IDL (ITT Industries, Inc., U.S.A.).

SM FRET Trace Simulations. Monte Carlo simulations were carried out to generate SM FRET traces to evaluate the algorithm. The total photon emission rate from a FRET pair of a donor and an acceptor was varied to adjust the Poissonian noise level. FRET dynamics are independent of the photon emission and detection. The time resolution of photon detection

is 1 μs, and the detector integration time is 25 ms, that is, the observation frame rate is 40/s. The photon detection efficiency is assumed to be 100%. Independent system dynamics from the monitoring scheme insures the incorporation of the abnormal FRET values due to the limited detection time resolution (Figure 1).

Results and Discussion

Comparison between a Gaussian Distribution and the Modified Mixed Gaussian Distribution for the HMM Optimization. First, the two FRET efficiency distributions (eqs 3 and 6) were used to fit histograms from simulated FRET traces (Figure 2). The histograms were constructed from 100 traces of 500 data points. The fitting parameters are the width and the amplitudes of the FRET peaks. The probability distribution between the Gaussian peaks is well approximated by eq 6, as clearly seen in Figure 2. Although the fitting is not as good as the reported analytical solution,[17] eq 6 can be used to fit multiple state models, and the computational time is short enough to be employed in a HMM optimization algorithm. It should be noted that as the kinetic rate is higher than half of the observation frame rate, the fitting becomes significantly deviated. Nonetheless, it is clearly shown in Figure 2 that the modified distribution (eq 6) fits the FRET distribution better than Gaussian distributions (eq 3).

Next, the performance of the two distributions in the HMM optimization is evaluated. The number of states and the kinetic scheme of the system are assumed to be known, that is, the size of the transition matrix is set constant, and some transition matrix elements are set to 0 by using a mask matrix. Kinetic rates are the product of the optimized transition matrix and the observation frame rate (=40/s). A set of 2500 SM FRET traces are generated per case (varying SNR and kinetic rates), where one trace contains 350 data points. The system switches between the 0.3 and 0.7 FRET state, and the rate going from the 0.3 to 0.7 state is fixed at 0.5/s while the rate going from 0.7 to 0.3 state is varied. The optimization is carried out with 175000 data points per case (500 traces per optimization). The plotted results in Figure 3 are obtained from five optimizations per data point. The 175000 points of data are chosen to ensure that the difference in the results is likely due to the difference in the probability distribution functions (the effect of the number of data points on the optimization performance follows in a later section). It is shown in Figure 3 that the Gaussian distribution (eq 3) and the modified Gaussian distribution (eq 6) underestimate both the kinetic rate and the FRET efficiency. The most pronounced difference between the two distribution functions is the high uncertainty in the kinetic rates optimized with the unmodified Gaussian distribution in the case of high SNR traces. This abnormally high optimization uncertainty is likely due to the fact that as the peaks get narrower (i.e., as the SNR improves and the rate becomes lower), the probability distribution between the FRET peaks according to eq 3 becomes effectively 0. The lower uncertainty of the modified distribution (<10% in most of the cases) makes it a better choice for the SM FRET data analysis. It is also clear that FRET efficiency is more accurate when the modified distribution (eq 6) is used, although the difference becomes smaller as the kinetic rate decreases, and SNR becomes more realistic (6−8) because the unmodified Gaussian distribution (eq 3) will be accurate enough to model the system under these conditions. The results for the rate 0.5/s and FRET efficiency 0.3 are omitted because the performance is equally good with eqs 3 and 6 within the error of 5% in the kinetic rate and the FRET efficiency.

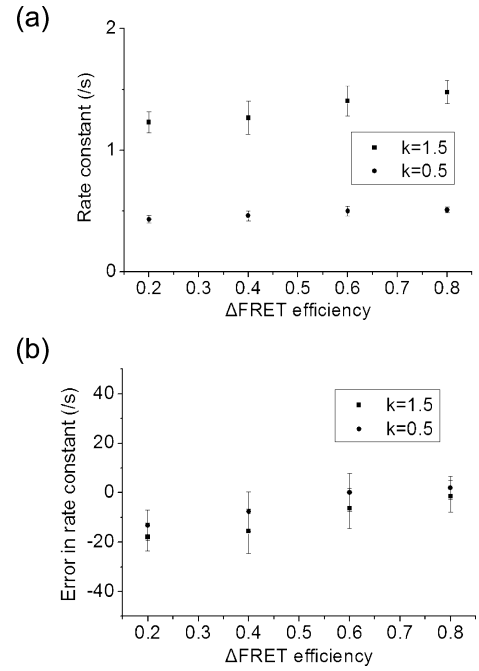


Figure 5. The effect of the ΔFRET efficiency on the HMM optimization performance; 1500 SM FRET traces were simulated with a SNR of 20.0, and the FRET efficiencies were varied (0.4/0.6, 0.3/0.7, 0.2/0.8, and 0.1/0.9 for the two FRET states to simulate ΔFRET of 0.2, 0.4, 0.6, and 0.8, respectively). Rates between the two states are 0.5 and 1.5/s. A high SNR and low rates ensure that any difference in the optimization performance can be attributed to the different ΔFRET. (a) Optimized rate constants plotted against the ΔFRET efficiency. (b) Errors in the optimized rate constant plotted against the ΔFRET efficiency.

Effect of the Number of Data Points on the Performance of the Algorithm. The effect of the number of data points used in the optimization is examined (Figure 4). A set of FRET traces with FRET efficiencies of 0.3 and 0.7 is simulated. The rate going from 0.3 to 0.7 is 0.5/s, and the rate going from 0.7 to 0.3 is 5.0/s. Five optimizations are performed per case. It is shown in Figure 4 that the 3500 data points, which contain 79.5 transitions with the given transition rates, are good enough to yield optimization results with <3% error in FRET efficiency and <21% error in the rates, on average. As the number of the data points increases, the uncertainty in the rates decreases, but the benefit is not sufficient to compensate for the increase in the number of data points after 7000 data points (159 transitions).

Effect of ΔFRET on the Performance of the Algorithm. A set of FRET traces with two states of varying FRET efficiencies, (0.1, 0.9), (0.2, 0.8), (0.3, 0.7), and (0.4, 0.6), is simulated. The rate going from a lower FRET state to a higher FRET state is 0.5/s, and the rate going the other direction is 1.5/s. The optimization is carried out three times on 7000 total data points per case. Figure 5 shows errors in the kinetic rates for different ΔFRET cases. It is clearly shown that the optimization yields more accurate results as ΔFRET increases.

Performance of the Algorithm with Multiple States and Multiple Gaussian Distributions Per State. Thirty SM FRET traces with 350 time points each are simulated to evaluate the algorithm in the optimization with multiple states. Traces follow a given kinetic scheme and rates, as shown in Figure 6a. The SNR is 6.0, and the noise originates solely from Poissonian photon emission statistics. The amount of data simulated per case is about half of what is typically taken to extract kinetics

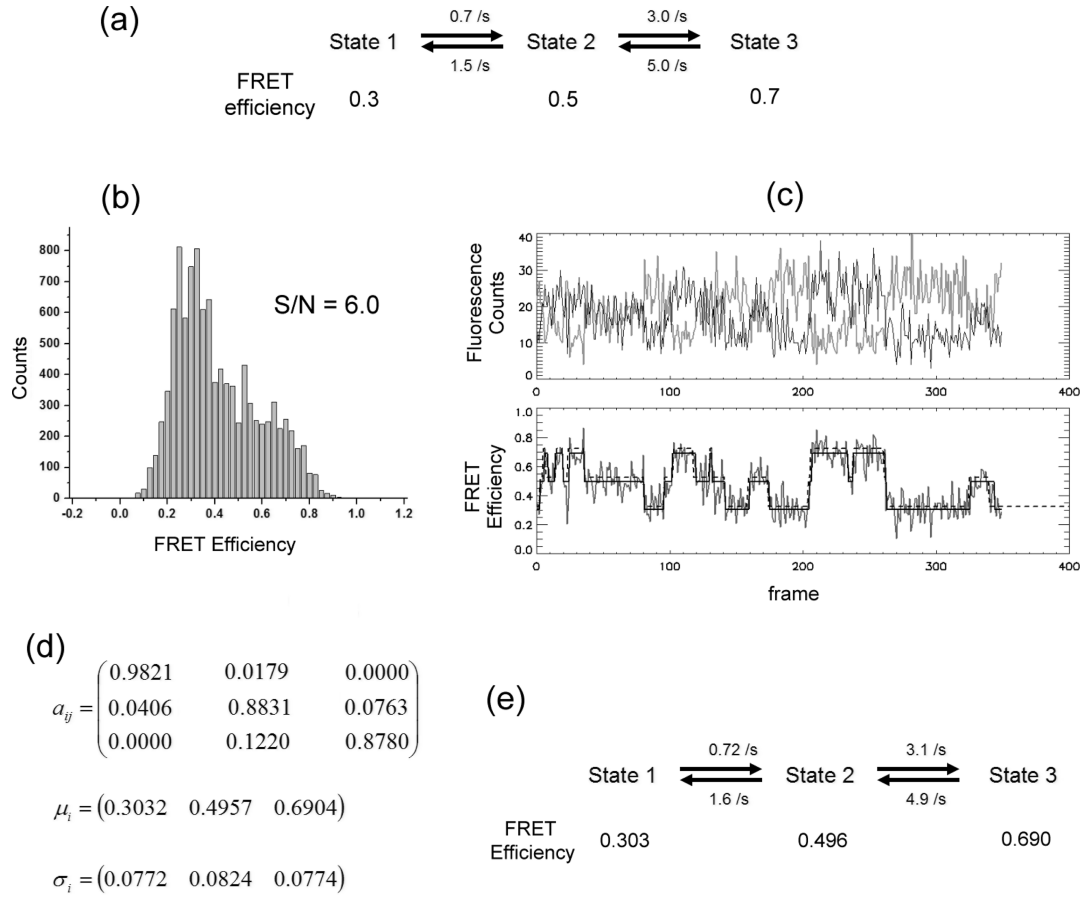
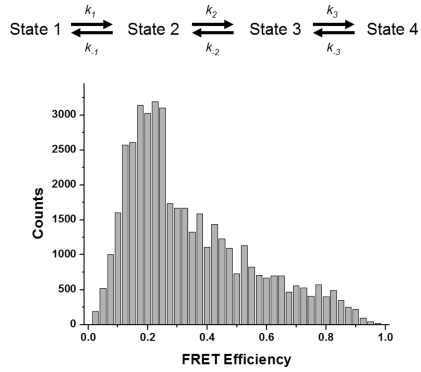


Figure 6. Demonstration of SM FRET data analysis with the proposed HMM algorithm. (a) Kinetic scheme of the three-state system simulated. FRET efficiencies for each state are shown below the state label, and kinetic rates are also shown in the kinetic scheme; 30 SM FRET traces with 350 time points each (about a half of the typical amount of experimental data to extract kinetics information) were simulated. The signal integration time was 25 ms (observation frame rate = 40/s). A photon emission rate of 1440 Hz was used to simulate traces with SNR 6.0. No additional background was added. (b) The histogram of 30 SM FRET traces simulated. (c) An example of simulated SM FRET traces and idealized FRET state transitions by the proposed HMM optimization. The thick gray line in the upper panel is the donor fluorescence count, and the thin black line is the acceptor fluorescence count. The noisy signal in the bottom panel is the calculated FRET efficiency (= acceptor fluorescence intensity/(acceptor fluorescence intensity + donor fluorescence intensity)). The solid straight line is the idealized FRET efficiency trace with the optimized HMM model parameters. The dashed straight line is the hidden state trace. The dashed line is shifted slightly upward to clarify the view. (d) The optimized model parameters of HMM. The transition matrix was restricted to the kinetic scheme as shown in (a), that is, off-diagonal elements were set to 0 by using a mask during the optimization. (e) Kinetic rates and FRET efficiencies calculated from the optimized model parameters in (d). Transition matrix elements multiplied by the detection frame rate (=40/s) yields the corresponding kinetic rates. The μ is the set of FRET efficiencies of each state, and σ is the noise in FRET efficiency and is in good agreement with the calculated FRET efficiency noises for each state according to the Beta distribution, which are 0.075, 0.082, and 0.075 for states 1, 2, and 3, respectively. The slight discrepancy between the estimated and the given FRET efficiency noises of state 1 or 3 is likely due to the approximation of the Beta distribution to a Gaussian distribution.

information (kinetics scheme and kinetic rates) from experiments. Figure 6e shows the optimized kinetic rates and the FRET efficiencies. The highest error in the FRET efficiency is 1.4% for state 3. Errors in the estimated kinetic rates are also low (<6.7%). Overall, it is confirmed that the maximization of $\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})$ yields the optimum model parameters for a system with multiple FRET states.

In experiments, the FRET efficiency of a state can vary slightly from trace to trace due to different background fluorescence levels and other environmental heterogeneity that affects the photophysics of fluorescence labels. To examine how this slight variation in FRET efficiency affects the performance of the algorithm, the optimized model parameters with different numbers of Gaussian distributions per state were compared. The model parameters were optimized for FRET traces with four states. These FRET traces were composed of three sets of slightly varying FRET efficiencies (Figure 7). On the basis of the optimized model parameters, it was revealed that the algorithm does not discriminate slightly varying FRET efficiencies belonging to one state. Instead, it finds the overall average FRET efficiency and standard deviation of the state from all of the FRET traces used in the optimization. Therefore, different background levels and other environmental heterogeneities that cause slight shifts in FRET efficiency do not lower the accuracy of the model parameters optimized with single Gaussian distribution per state.

Deducing the Number of States and the Kinetic Scheme. In previous sections, model parameters were optimized with a known kinetic scheme and the known number of states. In reality, kinetic schemes and the number of states are normally unknown. To deduce the number of states of a system, one can compare $\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})$ optimized with a series of different numbers of states. As the number of states in the optimization increases, $\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})$ will always increase following the power law because it is the product of individual probabilities, each of which is linearly affected by the increase in the number of states. By plotting $\log[\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})]$ with respect to the number of states, it is expected that there will be



Estimated kinetic rates

	k_1	k_{-1}	k_2	k_{-2}	k_3	k_{-3}
Given	0.5	1.0	2.0	4.0	3.0	5.0
1 Gaussian/state	0.54	0.88	1.8	3.5	2.9	4.6
2 Gaussians/state	0.54	0.87	1.8	3.5	2.9	4.6
3 Gaussians/state	0.54	0.87	1.8	3.5	2.9	4.6

Estimated FRET efficiencies

		State1	State2	State3	State4
Given	70 traces	0.2	0.4	0.6	0.8
	70 traces	0.22	0.44	0.64	0.82
	50 traces	0.18	0.37	0.57	0.79
	Average	0.202	0.407	0.607	0.805
Estimated (1 Gaussian/state)		0.207	0.408	0.607	0.795

Figure 7. The effect of variations in FRET efficiencies due to environmental heterogeneity on the accuracy of the optimized HMM model parameters. A four-state system was simulated with the same simulation conditions as those in Figure 6, except the kinetic scheme and the number of traces used. Three different sets of traces were simulated with different sets of FRET efficiencies to simulate FRET efficiencies varied by environmental heterogeneity. The shown FRET efficiency histogram is constructed from 190 simulated traces. Among the 190 traces, 70 traces have one set of FRET efficiencies, another set of 70 traces has a different set of FRET efficiencies, and the other 50 traces have another different set of FRET efficiencies, as shown in the table of "Estimated FRET efficiencies". One, two, or three Gaussian distributions per state were used to optimize the HMM model parameters to see the effect of the variations in FRET efficiencies on the accuracy of the parameters. The transition matrix was restricted to the kinetic scheme (i.e., off-diagonal elements were set to 0 by using a mask during the optimization). Results in the table of "Estimated kinetic rates" show that one Gaussian distribution per state can optimize the model parameters as well as two or more Gaussian distributions per state.

a distinct point where $\Delta\log[\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})]$ abruptly decreases (Figure 8). As shown in Figure 8, the point of abrupt change in $\Delta\log[\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})]$ is the smallest number of states that can model the system and is identified as the number of states of the system. As the noise level of SM FRET traces becomes higher, the residual increase in $\log[\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})]$ past the smallest number of states becomes bigger (Figure 8c). Nevertheless, it is straightforward to determine the number of states. Once the right number of states is identified, the kinetic scheme can be easily deduced from the optimized transition matrix. For instance, if there is no direct transition between two states in FRET traces, the corresponding transition matrix element will be unfeasibly small, as demonstrated in the next section.

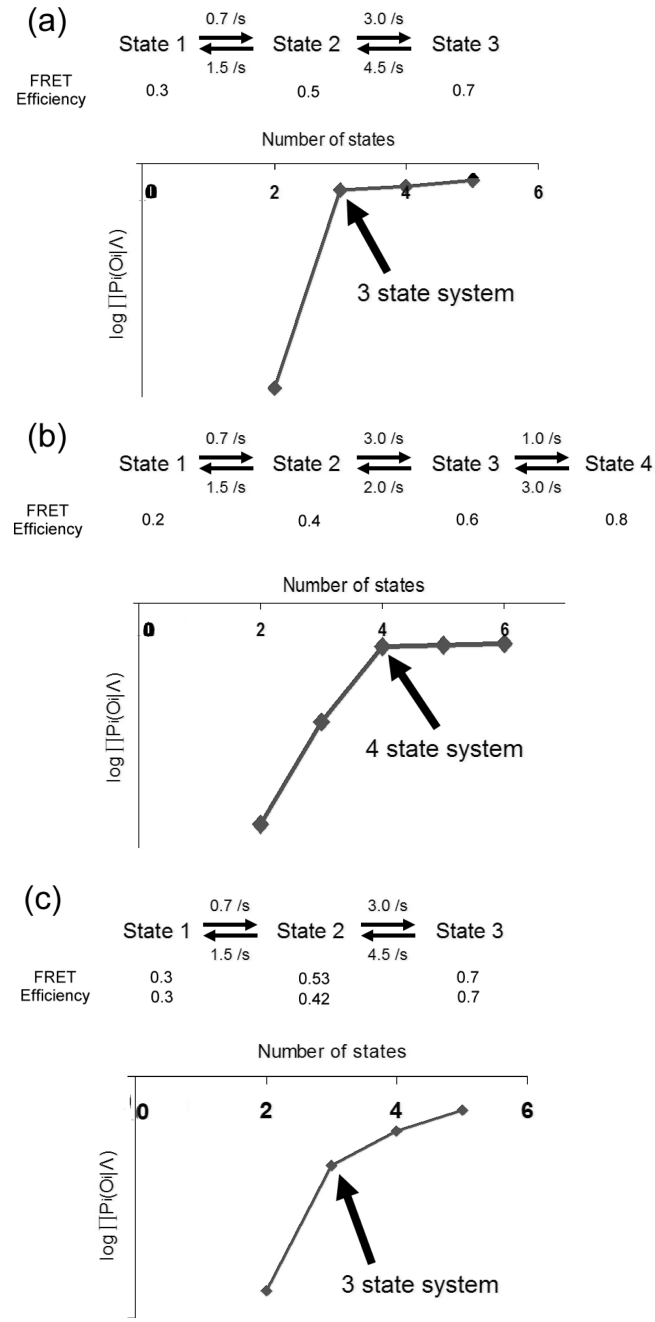


Figure 8. Demonstration of deducing the number of hidden states from SM FRET traces. Two three-state systems and one four-state system were simulated with the given rates and FRET efficiencies. SNR 6.0 was used (noise solely from Poissonian photon emission statistics). Fifty traces each were simulated for (a−c) (350 time points per trace). In (c), two sets of FRET efficiencies (30 and 20 traces) were simulated. Each set of 50 traces was used to optimize the model parameters with three to six states. Each chart in (a−c) shows $\log[\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})]$ plotted against the number of states used in the optimization. In each case of (a−c), there is a distinct number of state where $\Delta\log[\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})]$ drops abruptly, informing the smallest number of required states to model the observation sequence. These points are indicated with arrows. In a noisier data set (c), the $\Delta\log[\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})]$ change is not as abrupt as in cases (a) and (b). Nevertheless, it is straightforward enough to identify the number of states in (c). In the charts, $\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})$ is abbreviated to $\prod P_l(O_l|\Lambda)$.

Demonstration of Extracting Kinetics Information From SM FRET Traces. On the basis of the procedure described above, a process of extracting kinetic information from SM FRET traces is demonstrated in Figure 9. A very noisy set of

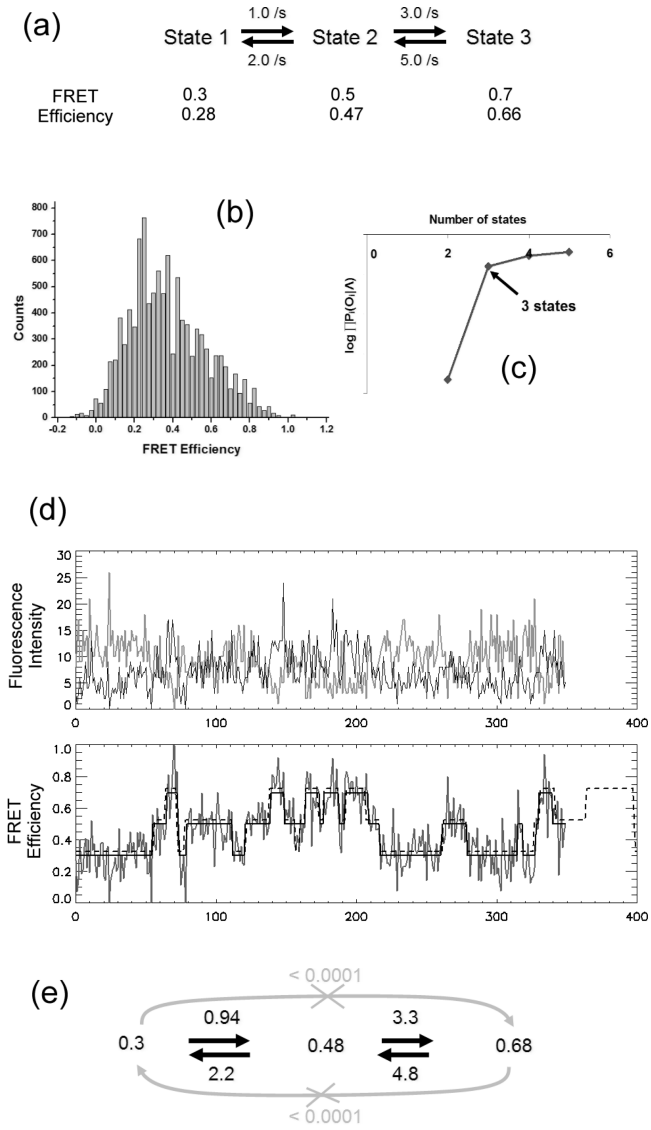


Figure 9. Demonstration of extracting kinetics information from SM FRET traces. A set of 30 SM FRET traces (350 time points each) with SNR 4.0 (noise solely from Poissonian photon emission statistics) was simulated and used to optimize the model parameters of HMM. SNR is set to be worse than a typical experimental SNR in order to test the robustness of the algorithm. (a) Kinetic scheme used in the simulation. (b) FRET efficiency histogram of the simulated traces. (c) $\log[\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})]$ versus the number of states, indicating that the system dwells on three states. (d) An example of SM FRET traces (upper panel, the gray line is the donor fluorescence intensity, and the black line is the acceptor fluorescence intensity), FRET efficiency (the noisy signal in the lower panel), and the idealized FRET efficiency sequence (the solid line over the noisy FRET signal in the lower panel) by the algorithm with three states. The dashed line in the lower panel is the given event sequence shifted slightly upward to clarify the view. (e) Reconstructed kinetic scheme from the optimized transition matrix. Gray lines show null transitions found by the algorithm as explained in the text.

data (SNR calculated from Poissonian photon emission statistics = 4.0) from a three-state system was simulated. Two sets of FRET efficiencies were used to simulate two different sets of data taken in two different environments. First, the maximum $\prod_{l=1}^{n} P_l(O_l|\{a,b,\pi_l\})$ is calculated with the optimized model parameters with two to five states and a single Gaussian distribution per state. As shown in Figure 9c, it is clear that the system dwells on three states. An example of idealized FRET traces from optimum model parameters with three states is shown in Figure 9d. From the simulation, it was found that states 1 and 3 are not connected to each other since the transition matrix elements are too small (<0.0001/s) to be real, that is, on the basis of the length of the longest trace (= 8.75 s), the slowest possible transition rates between states should not be much lower than 1/8.75 = 0.11/s. Estimated kinetic rates are in good agreement with the given rates within 10% error.

Conclusions

Using HMM, a systematic way of extracting kinetics information from noisy SM FRET data is demonstrated. There are three distinct sources of noise in SM FRET signal, (i) Poissonian noise from photon emission statistics, (ii) noise from an environment such as background fluorescence, stray light, and noise in detection devices, and (iii) short lifetime events and the unsynchronized detection. It is demonstrated that the errors from the first two sources can be suppressed by using the proposed algorithm. The third source of noise, however, is unavoidable, although HMM with the proposed modified FRET distribution can reduce the error. Nevertheless, thanks to the reasonably high precision of the proposed method, HMM optimization results can be used to report the kinetic rates of a SM FRET system when the report accompanies the information on the level of error due to the limited detection time resolution.

Acknowledgment. This work was supported by the NIH Pathway to Independence Award (GM079960), Searle Scholar Award, and the Camillie and Henry Dreyfus New Faculty Award.

References and Notes

(1) Ha, T.; Enderle, T.; Ogletree, D. F.; Chemla, D. S.; Selvin, P. R.; Weiss, S. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 6264.
(2) Weiss, S. *Nat. Struct. Biol.* **2000**, *7*, 724.
(3) McKinney, S. A.; Declais, A.-C.; Lilley, D. M. J.; Ha, T. *Nat. Struct. Mol. Biol.* **2003**, *10*, 93.
(4) Ha, T.; Zhuang, X.; Kim, H. D.; Orr, J. W.; Williamson, J. R.; Chu, S. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9077.
(5) Zhuang, X.; Bartley, L. E.; Babcock, H. P.; Russell, R.; Ha, T.; Herschlag, D.; Chu, S. *Science* **2000**, *288*, 2048.
(6) Ha, T.; Rasnik, I.; Cheng, W.; Babcock, H. P.; Gauss, G. H.; Lohman, T. M.; Chu, S. *Nature* **2002**, *419*, 638.
(7) Myong, S.; Rasnik, I.; Joo, C.; Lohman, T. M.; Ha, T. *Nature* **2005**, *437*, 1321.
(8) Lee, T.-H.; Blanchard, S. C.; Kim, H. D.; Puglisi, J. D.; Chu, S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 13661.
(9) Rabiner, L. R. *Proc. IEEE* **1989**, *77*, 257.
(10) Baum, L. E.; Petrie, T. *Ann. Math. Stat.* **1966**, *37*, 1554.
(11) Viterbi, A. J. *IEEE Trans. Inf. Theory* **1967**, *IT-13*, 260.
(12) Forney, G. D. *Proc. IEEE* **1973**, *61*, 268.
(13) Qin, F.; Auerbach, A.; Sachs, F. *Biophys. J.* **2000**, *79*, 1928.
(14) Qin, F.; Auerbach, A.; Sachs, F. *Biophys. J.* **2000**, *79*, 1915.
(15) Smith, D. A.; Steffen, W.; Simmons, R. M.; Sleep, J. *Biophys. J.* **2001**, *81*, 2795.
(16) McKinney, S. A.; Joo, C.; Ha, T. *Biophys. J.* **2006**, *91*, 1941.
(17) Gopich, I. V.; Szabo, A. *J. Phys. Chem. B* **2007**, *111*, 12925.
(18) Dahan, M.; Deniz, A. A.; Ha, T. J.; Chemla, D. S.; Schultz, P. G.; Weiss, S. *Chem. Phys.* **1999**, *247*, 85.
(19) Baum, L. E.; Petrie, T.; Soules, G.; Weiss, N. *Ann. Math. Stat.* **1970**, *41*, 164.

JP903831Z