The Journal of Physical Chemistry Letters

Letter

pubs.acs.org/JPCL

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,[‡] and Christy F. Landes*[,†,§]

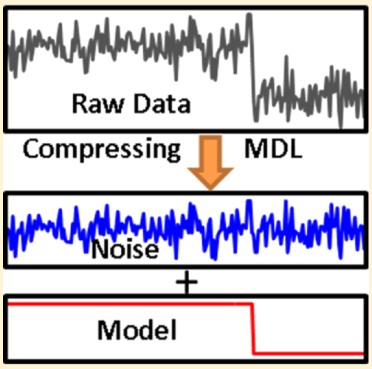
[†]Department of Chemistry, Rice University, MS 60, Houston, Texas 77251-1892, United States

[‡]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

Supporting Information

ABSTRACT: We introduce a step transition and state identification (STaSI) method for piecewise constant single-molecule data with a newly derived minimum description length equation as the objective function. We detect the step transitions using the Student's *t* test and group the segments into states by hierarchical clustering. The optimum number of states is determined based on the minimum description length equation. This method provides comprehensive, objective analysis of multiple traces requiring few user inputs about the underlying physical models and is faster and more precise in determining the number of states than established and cutting-edge methods for single-molecule data analysis. Perhaps most importantly, the method does not require either time-tagged photon counting or photon counting in general and thus can be applied to a broad range of experimental setups and analytes.



SECTION: Biophysical Chemistry and Biomolecules

Single-molecule analysis often involves a compromise when our desire to quantify space/time heterogeneity is challenged by innately low signal-to-noise conditions. Only when these counterbalancing principles are optimized can we acquire access to equilibrium and nonequilibrium details that are unobtainable by ensemble methods. Single-molecule Förster resonance energy transfer (smFRET) measurements explore conformations and dynamics of biomolecules unresolvable in the ensemble state distribution.[1−5] In smFRET experiments, single molecules visit different structural or conformation states and generate piecewise constant signals.[1,2] Identifying states and step transitions between states is important to understand the stationary state distribution of the system and the dynamics among different states and to make testable mechanistic predictions. However, it is often challenging to identify states and transitions due to noise sources during the measurements.

Established state determination methods for smFRET data are designed to extract the heterogeneity of the system buried within the mitigating fluctuations due to noise and include the Watkins and Yang change-point method,[6,7] hidden Markov model-based FRET time trajectory analysis program (HaMMy)[8−10] combined with wavelet denoising,[11,12] and variational Bayesian inference for smFRET time series (vbFRET).[13] The Watkins and Yang change-point method uses few user inputs but is designed for continuous photon-by-photon traces[6] and thus is not practical for binned data. Although collecting time-tagged photon-by-photon data is similar, and in most cases preferable to collecting binned photon data, time-tagged collection systems require more complicated and expensive pulsed excitation sources and hardware to resolve photon arrival times on single-photon counting detectors. In addition, for many other detectors used in single-molecule experiments, the collection frequencies required for single-photon collection are often faster than their temporal resolution. Thus, continuous wave excitation sources and binned photon data collection are widely used experimental simplifications of the more accurate, but expensive, time-tagged methods. Several of the most widely used single-molecule data processing algorithms (e.g., HaMMy[8] and vbFRET[13]) were specifically designed to analyze binned data because of its ubiquity and relative ease of acquisition. Both HaMMy and vbFRET assume that the data can be represented as a hidden Markov chain. HaMMy requires the user to decide the optimum number of states,[8] which is a challenge if a priori knowledge of the underlying states is unavailable. vbFRET automatically determines the optimum number of states based on maximum evidence inference,[13] but

Received: July 10, 2014
Accepted: August 28, 2014
Published: August 28, 2014

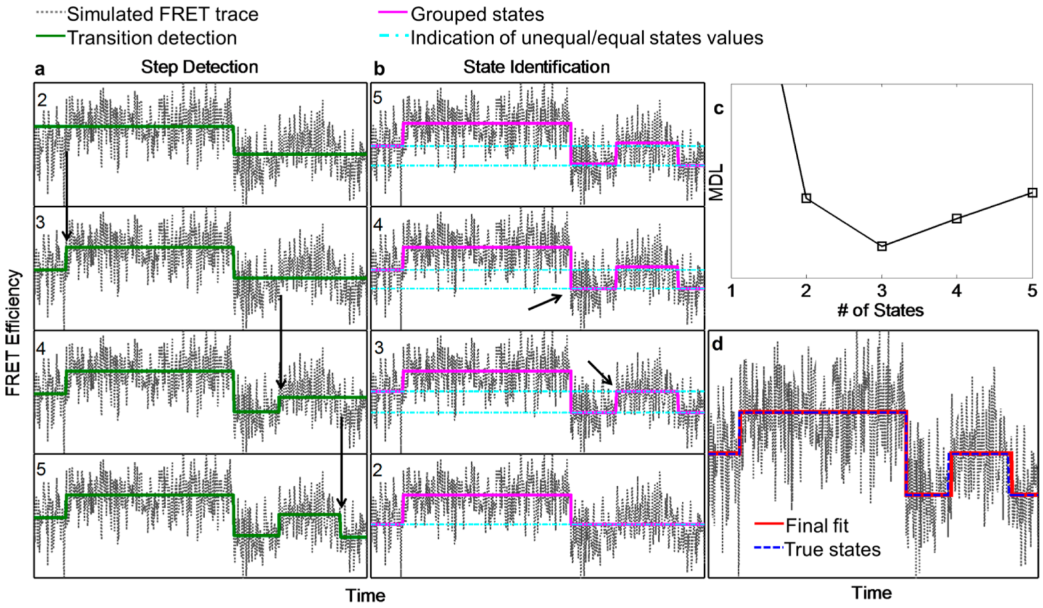


Figure 1. Demonstration of STaSI using a simulated three-state FRET efficiency trace with added Gaussian noise. (a) Recursive process to detect step transitions using the Student's t test. The step transition identified in each recursion is highlighted by the black arrows. The number of segments is indicated in the upper-left corners. (b) The iterative method to group the identified segments into states begins from the final result of step detection process and continues until only a single state remains. The merged segments from five to four states and from four to three states are highlighted by the black arrows. The number of assumed states is indicated in the upper-left corners. (c) The calculated MDL value for each state set. Clearly, the three-state set is the optimum number of states, with the global minimum MDL value. (d) The determined three-state fit (red) compared with true states (blue).

for noisy data (i.e., noise levels larger than the separation of states) or data with fast dynamics (i.e., with mean lifetimes within an order of magnitude larger than the sampling time) the method identifies redundant states due to noise- or binning-induced artifacts (Figure 3). Thus, the optimum solution for state determination remains an open question, especially for binned data.

All of these methods assume that smFRET data are generated by dynamics among several FRET states. This state distribution is usually sparse (meaning that the FRET states can be represented by several delta functions), even though experimental smFRET efficiency traces usually have a broad distribution due to noise. In this work, we introduce a step transition and state identification (STaSI) method to analyze smFRET data and recover the underlying sparse state distribution. STaSI is particularly designed for smFRET data, but, in principle, STaSI is useful for any piecewise constant signals. STaSI applies an equation we have derived for piecewise constant signals based on the minimum description length (MDL) principle[14,15] as the objective function

$$\text{minimize MDL} \tag{1}$$

where MDL = $F + G$, and F measures the goodness of fit using the L^1 norm and G measures the complexity of the fitting model. Compared with other information criteria, the MDL principle accounts for the detailed parameter complexity of the model[14,15]

$$F = \frac{\sum_{i=1}^{N} |y(t_i) - y_{\text{fit}}(t_i)|}{2\sigma} \tag{2}$$

$$G = \frac{k}{2}\ln\frac{1}{2\pi} + k\ln\frac{V}{\sigma} + \frac{N_{\text{tp}}}{2}\ln N + \frac{1}{2}\left(\sum_{k}^{i=1}\ln n_i + \sum_{N_{\text{tp}}}^{j=1}\ln\frac{T_j^2}{\sigma^2}\right) \tag{3}$$

where σ is the overall noise level; $y(t_i)$ and $y_{\text{fit}}(t_i)$ are the real data and fit value at time t_i, respectively; N is the total number of data points of the trace; k is the number of states; N_{tp} is the number of transition positions; V is the domain size (= $y_{\max} - y_{\min}$) for all y; n_i is the number of data points assigned to state i, and T_j is the difference of the fitting values before and after the transition position j. Here we derived G for smFRET data to consider the sparseness of the states and the transitions among these states; the full derivation is provided in the Supporting Information. MDL reaches a minimum when the increase in the complexity of the model (G) using an additional state equals the decrease in the fitting error (F) as measured by the L^1 norm. Overall, the MDL equation accounts for the balance between simplicity and accuracy and guarantees the minimized solution to be the sparsest approximation.[16,17]

The solution domain for the MDL objective function is first reduced by searching for the optimum solution for each number of states. The Student's t test with unequal sample size and global noise level is applied to detect all of the step transitions[18] and breaks down the trace into multiple segments. The recursive process in Figure 1a applies the Student's t test on each segment until no further transition points are found. Similar to the change-point method,[6,7] we then group these segments recursively up to one state to find the best grouping strategy for every possible number of states. In each grouping iteration, the most similar two states are grouped into one state (Figure 1b). More details and the related equations on step transition and state grouping are explained in the Supporting

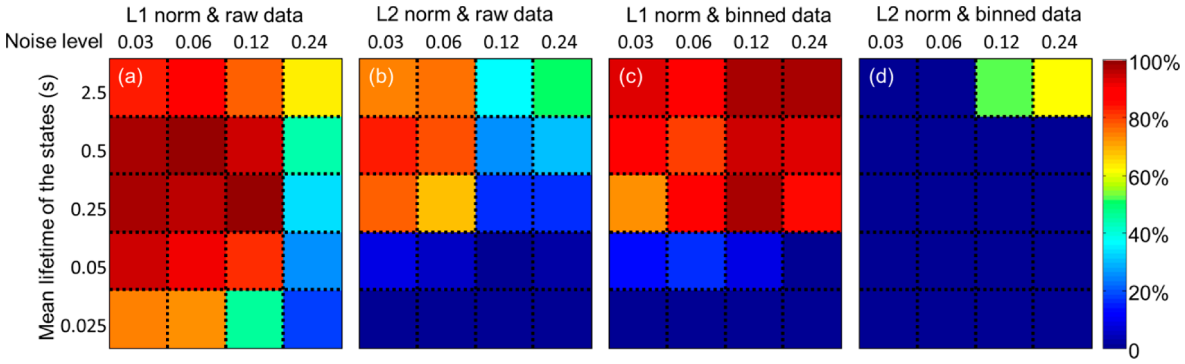


Figure 2. Comparison between the L^1 norm and the L^2 norm for data with different noise levels and mean lifetime of the states. The horizontal axis labels the four different noise levels, and the vertical axis labels the five different mean lifetimes of the states. The simulation uses five FRET states: 0.2, 0.25, 0.35, 0.5, and 0.7; a sampling time of 1 ms; and a binning time of 10 ms. Under each condition, 100 simulations are repeated. The different colors represent the success rates of correctly identifying the number of states. (a) Using the L^1 norm analyzing raw data. (b) Using the L^2 norm analyzing raw data. (c) Using the L^1 norm analyzing binned data. (d) Using the L^2 norm analyzing binned data.

Information. The MDL value of the best solution under each number of states is calculated (Figure 1c), and the number of states corresponding to the global minimum MDL value is considered the optimum fitting model (Figure 1d). In short, this strategy first calculates the best fitting for different number of states, and the minimum MDL determines the optimum number of states. While we do not have undeniable proof that the final analysis reaches the global minimum MDL value, our performance tests (Figure 2 and 3) provide strong evidence of such. Moreover, this preselection scenario dramatically reduces the complexity of the algorithm from a computationally impossible nondeterministic polynomial-time hard (NP-hard)[19] classification to one in which the computational time scales with N^2 (Supporting Information Figure S4).

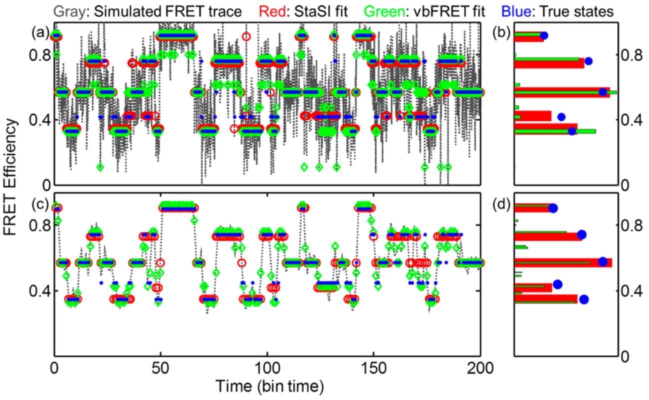


Figure 3. Performance of STaSI using simulated five FRET states traces with fast dynamics. Only the first 200 (out of about 15 000) bin time (corresponding to 2000 sampling time for raw data in panel a) data points are shown for illustration. (a) Simulated raw data analyzed by STaSI and vbFRET. (b) Corresponding histograms of the STaSI fit, vbFRET fit, and the true states for raw data. (c) Simulated ten-point binned data analyzed by STaSI and vbFRET. (d) Corresponding histograms of the STaSI fit, vbFRET fit, and the true states for binned data.

Using the L^1 norm to measure F, the goodness of fit, is important to find the sparsest approximation of the real solution[16,17] and is robust to high noise levels, non-Gaussian noise, and binning artifacts,[20] as shown in Figure 2. While F is usually measured by the L^2 norm (squared error),[15] our simulations show that under the typical noisy conditions of single-molecule measurements the L^2 norm generates many redundant states (Figure 2 and Supporting Information Figures S1−S3). Using the L^1 norm, STaSI identifies the correct number of states successfully (success rate >70%) for noise level smaller than 0.12 and mean lifetime of the states longer than 0.05 s (Figure 2a). Using the L^2 norm, STaSI only finds the correct number of states when the noise level is smaller than 0.06 and the mean-lifetime is longer than 0.25 s (Figure 2b). For the L^1 norm, binning improves the success rate of finding the correct number of states for noisy data with mean-lifetimes longer than 0.25 s (Figure 2c). For the L^2 norm, STaSI fails to find the correct number of states using binned data (Figure 2d). Overall, by using the L^1 norm, STaSI can successfully recover the state distribution under broad noise and mean-lifetime conditions. Similar results of using the L^2 and L^1 norm have been reported in other applications.[20] Therefore, for our desired application to single-molecule data, we use the L^1 norm to quantify F.

The STaSI method can correctly analyze noisy smFRET data containing fast dynamics (i.e., when the interstate transition time is within ∼10 × the collection bin time). The signal-to-noise ratio can usually be improved through binning, but in the presence of fast dynamics, binning introduces artifact states in between these real states and limits the temporal resolution of single-molecule FRET. In Figure 3, STaSI identifies all of the states for both the noisy raw data (with 12% state assignment error, 8% state distribution error, and 0.008 absolute efficiency deviation) and the binned data (with 20% state assignment error, 9% state distribution error, and 0.013 absolute efficiency deviation). In comparison, vbFRET identifies four false states and fails to identify one state due to the influence of the noise for the raw data (Figure 3a,b). For binned data, vbFRET identifies six artifact states in between the real states (Figure 3c,d). This test demonstrates that STaSI provides a solution to interpret noisy data with fast dynamics, improving the experimental temporal resolution. The relatively large error for binned data is due to the presence of fast dynamics where multiple states are averaged in a single bin time. The binned data are preferred for data with relatively slower dynamics and a large noise level (Figure 2c). This improved performance of STaSI is mainly due to the MDL equation we derived for the piecewise constant signals.

Because STaSI is parameter-free in terms of both the analysis and the collection method, the analysis is not limited to smFRET data. STaSI can be directly applied with other piecewise constant signals such as those that occur in imaging,[21] optical tweezers,[22] or scanning probe analyses. Methods based on the hidden Markov chain like HaMMy and vbFRET can be extended to apply the MDL principle to search for the optimum number of states. Tuning the noise level parameter in the Student's t test can make the method more sensitive to smaller transitions or allow it to only capture relatively large transitions. Overall, STaSI is a good example of applying different information theory techniques for robust single-molecule data analysis.

In summary, we have designed STaSI to analyze the states and interstate step transitions of piecewise constant signals. STaSI combines the Student's t test and a new derivation of the MDL equation to optimize the analysis for piecewise constant signals. This method fills the gap of change-point detection with discontinuous binned data, especially in the single-molecule field, and improves the state determination for noisy data or data with fast dynamics. STaSI saves effort and time to identify the transition positions manually and decreases user biases when analyzing complicated data. In the future, we plan to apply this algorithm to other situations such as single-molecule instantaneous displacements in heterogeneous environments,[23] engineered surface association and dissociation,[21] and aggregation of conjugated polymers.[24] The performance of STaSI under different assumptions will be explored in detail using simulated FRET traces under different models from molecular dynamic levels.[25]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Derivation of function G in the MDL expression for piecewise constant signals, the Student's t test to detect step transitions, states grouping algorithm, calculating the noise level using the Haar wavelet transform, comparison between the L^1 norm and the L^2 norm, speed comparison, identifying short-lived state segments, and FRET trajectory simulation. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

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

Present Address

∥J.C.: Department of Chemistry and Biochemistry, Ohio University, Athens, OH, 45701–2979, USA.

Notes

The authors declare no competing financial interests.

■ ACKNOWLEDGMENTS

We thank the National Science Foundation (Grants CBET-1134417 and CHE-1151647), the Welch Foundation (Grant C-1787), and the National Institutes of Health (Grant GM94246-01A1) for support of this work. L.K. acknowledges the National Science Foundation Graduate Research Fellowship (Grant 0940902). Special thanks also go to L. J. Tauzin, C. P. Byers, B. Hoener, and H. Jaurich for their thoughts and discussions as well as S. Link and his research group for their feedback on this project and Prof. K. F. Kelly and Y. Li for their helpful discussion. T.K. thanks Grant-in-Aid for Scientific Research (B), JSPS, Grant-in-Aid for Exploratory Research, JSPS, a Grant-in-Aid for Scientific Research on Innovative Areas "Spying minority in biological phenomena (No.3306)", MEXT, and HFSP for supports.

■ REFERENCES

(1) Roy, R.; Hohng, S.; Ha, T. A Practical Guide to Single-Molecule FRET. *Nat. Methods* **2008**, *5*, 507–516.

(2) Zhao, Y.; Terry, D.; Shi, L.; Weinstein, H.; Blanchard, S. C.; Javitch, J. A. Single-Molecule Dynamics of Gating in a Neurotransmitter Transporter Homologue. *Nature* **2010**, *465*, 188–193.

(3) Landes, C. F.; Rambhadran, A.; Taylor, J. N.; Salatan, F.; Jayaraman, V. Structural Landscape of Isolated Agonist-Binding Domains from Single AMPA Receptors. *Nat. Chem. Biol.* **2011**, *7*, 168–173.

(4) Weiss, S. Measuring Conformational Dynamics of Biomolecules by Single Molecule Fluorescence Spectroscopy. *Nat. Struct. Mol. Biol.* **2000**, *7*, 724–729.

(5) Sakon, J. J.; Weninger, K. R. Detecting the Conformation of Individual Proteins in Live Cells. *Nat. Methods* **2010**, *7*, 203–205.

(6) Watkins, L. P.; Yang, H. Detection of Intensity Change Points in Time-Resolved Single-Molecule Measurements. *J. Phys. Chem. B* **2005**, *109*, 617–628.

(7) Terentyeva, T. G.; Engelkamp, H.; Rowan, A. E.; Komatsuzaki, T.; Hofkens, J.; Li, C. B.; Blank, K. Dynamic Disorder in Single Enzyme Experiments: Facts and Artifacts. *ACS Nano* **2012**, *6*, 346–354.

(8) McKinney, S. A.; Joo, C.; Ha, T. Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. *Biophys. J.* **2006**, *91*, 1941–1951.

(9) Andrec, M.; Levy, R. M.; Talaga, D. S. Direct Determination of Kinetic Rates from Single-Molecule Photon Arrival Trajectories Using Hidden Markov Models. *J. Phys. Chem. A* **2003**, *107*, 7454–7464.

(10) Keller, B. G.; Kobitski, A.; Jaschke, A.; Nienhaus, G. U.; Noe, F. Complex RNA Folding Kinetics Revealed by Single-Molecule FRET and Hidden Markov Models. *J. Am. Chem. Soc.* **2014**, *136*, 4534–4543.

(11) Taylor, J. N.; Makarov, D. E.; Landes, C. F. Denoising Single-Molecule FRET Trajectories with Wavelets and Bayesian Inference. *Biophys. J.* **2010**, *98*, 164–173.

(12) Taylor, J. N.; Landes, C. F. Improved Resolution of Complex Single-Molecule FRET Systems via Wavelet Shrinkage. *J. Phys. Chem. B* **2011**, *115*, 1105–1114.

(13) 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.

(14) Rissanen, J. A Universal Prior for Integers and Estimation by Minimum Description Length. *Ann. Stat.* **1983**, *11*, 416–431.

(15) 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.

(16) Candès, E. J.; Romberg, J.; Tao, T. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Trans. Inf. Theory* **2006**, *52*, 489–509.

(17) Candès, E. J.; Tao, T. Decoding by Linear Programming. *IEEE Trans. Inf. Theory* **2005**, *51*, 4203–4215.

(18) Carter, N. J.; Cross, R. A. Mechanics of the Kinesin Step. *Nature* **2005**, *435*, 308–312.

(19) Johnson, D. S. A Catalog of Complexity Classes. In *Handbook of Theoretical Computer Science: Vol. A: Algorithms and Complexity*; van Leeuwen, J., Ed.; MIT Press: Cambridge, MA, 1990.

(20) Wagner, A.; Wright, J.; Ganesh, A.; Zhou, Z.; Mobahi, H.; Ma, Y. Toward a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 372–386.

(21) Kisley, L.; Chen, J.; Mansur, A. P.; Shuang, B.; Kourentzi, K.; Poongavanam, M. V.; Chen, W. H.; Dhamane, S.; Willson, R. C.; Landes, C. F. Unified Superresolution Experiments and Stochastic Theory Provide Mechanistic Insight into Protein Ion-Exchange

Adsorptive Separations. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 2075−2080.

(22) Kerssemakers, J. W.; Munteanu, E. L.; Laan, L.; Noetzel, T. L.; Janson, M. E.; Dogterom, M. Assembly Dynamics of Microtubules at Molecular Resolution. *Nature* **2006**, *442*, 709−712.

(23) Elliott, L. C.; Barhoum, M.; Harris, J. M.; Bohn, P. W. Trajectory Analysis of Single Molecules Exhibiting Non-Brownian Motion. *Phys. Chem. Chem. Phys.* **2011**, *13*, 4326−4334.

(24) Eisele, D. M.; Knoester, J.; Kirstein, S.; Rabe, J. P.; Vanden Bout, D. A. Uniform Exciton Fluorescence from Individual Molecular Nanotubes Immobilized on Solid Substrates. *Nat. Nanotechnol.* **2009**, *4*, 658−663.

(25) Haas, K. R.; Yang, H.; Chu, J. W. Expectation-Maximization of the Potential of Mean Force and Diffusion Coefficient in Langevin Dynamics from Single Molecule FRET Data Photon by Photon. *J. Phys. Chem. B* **2013**, *117*, 15591−15605.