THE JOURNAL OF PHYSICAL CHEMISTRY B LETTERS

2009, *113,* 13886–13890
Published on Web 09/28/2009

Hidden Markov Analysis of Short Single Molecule Intensity Trajectories

Soonkyo Jung and Robert M. Dickson*

School of Chemistry and Biochemistry and Petit Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, Georgia 30332-0400

Received: July 23, 2009; Revised Manuscript Received: September 16, 2009

Photon trajectories from single molecule experiments can report on biomolecule structural changes and motions. Hidden Markov models (HMM) facilitate extraction of the sequence of hidden states from noisy data through construction of probabilistic models. Typically, the true number of states is determined by the Bayesian information criteria (BIC); however, constraints resulting from short data sets and Poisson-distributed photons in radiative processes like fluorescence can limit successful application of goodness-of-fit statistics. For single molecule intensity trajectories, additional information criteria such as peak localization error (LE) and chi-square probabilities can incorporate theoretical constraints on experimental data while modifying normal HMM. Chi-square minimization also serves as a stopping point of the iteration in which the system parameters are trained. Peak LE enables exclusion of overfitted and overlapped states. These constraints and criteria are tested against BIC on simulated single molecule trajectories to best identify the true number of emissive levels in any sequence.

Introduction

Originally developed for speech recognition,[1] HMM enables reconstruction of the state sequence from observables when the true state is hidden. Iterative state reconstructions refine the three parameters needed to describe time series with well-defined transition probabilities linking all states: the transition, emission, and initiation matrices. The transition matrix elements are probabilities that one state changes to a different state in the subsequent step. The emission matrix links the observable to hidden states that, in fluorescence intensity trajectories, correspond to the different observed intensity levels. The initiation matrix gives the probability of each state at the first step. The objective of HMM analysis is to find the most likely system parameters that describe the system. The Baum−Welch algorithm[1,2] is frequently used to find the answer by training initial system parameters based on the observed sequence. After training is finished, the state sequence is reconstructed by the Viterbi algorithm.[3] Since the most probable state sequence can be extracted from complicated and noisy data using trained system parameters, HMM has been utilized in many fields such as particle tracking,[4,5] single ionic current measurement,[6] dwell time analysis,[7] FRET analysis,[8,9] and simulation of single molecule fluorescence and kinetics.[10,11] Although incompatible with state sequence reconstructions, non-Markovian memory effects due to thermal fluctuation[12] can be quantified through correlation analysis.[13]

Proper state sequence reconstruction, however, requires determination of the true number of states, making evaluation of the most likely system dimension a key problem. Several information criteria have been developed to best estimate the true number of states, including Akaike (AIC),[14] Hannan−Quinn (HQC),[15] and Bayesian (BIC).[8,16−18] These criteria modify maximum likelihood estimates with penalizing terms based on the Laplace−Bayesian approximation of the likelihood and prior probabilities in the limit of large sample size.[19] Although direct, model-independent information theoretical approaches[18] may also work well, especially when a kinetic model is inapplicable, even these model-independent methods rely on these same information criteria that may not consistently find the simplest and most likely model for short data sets. Further, for kinetic models, HMM has enjoyed wide applicability and success,[9,10,20,21] suggesting that any information criteria improving accuracy especially for short data sets may be of great utility. While BIC has been shown to provide a rigorous upper limit on the true number of states for an infinitely long sequence,[19] many important experimental data sets are far too short to satisfy this requirement. Therefore, in order to improve the performance of modeling for Poissonian emitters, we introduced chi-square minimization, chi-square probability-based goodness-of-fit,[22] and localization error (LE) to improve fitting of short trajectories or those with many states. LE, for example, has been used to estimate the precision of single particle images[23,24] by enabling verification of the resolvability of two objects. In this paper, data sets of varying length are generated by pseudorandomly generated transition and emission matrices exhibiting experimentally relevant, Poisson-distributed noise. We fit the data sets by Baum−Welch and modified algorithms, using several types of criteria to determine the "true" number of states and compare to known values of simulated data sets. Effects of trajectory length and robustness relative to choice of initial conditions indicate that the chi-square probability and LE are crucial to proper state reconstructions.

Methods

Simulating typical single molecule fluorescence trajectories, 10000 data sets of varying length with pseudorandomly varying

* To whom correspondence should be addressed. E-mail: dickson@chemistry.gatech.edu.

10.1021/jp907019p CCC: $40.75

system parameters were generated in Matlab software (R2008a, Mathworks). Emission levels were randomly chosen. To match appropriately binned experimental data, however, the transition matrix deviated from truly random values in that staying in the same state for the subsequent step was defined to be more probable than transition to any other individual state. The probability of staying in a given state could, however, be much less than 0.5. Obviously, the sum of each transition matrix row was constrained to unity, but the sum of each column varied widely due to the pseudorandom transition probabilities, yielding a wide range of state populations and transition probabilities in the simulated data traces. The emission levels (i.e., simulated intensity range) varied from 0 to 150 counts per bin. Emission matrices had levels of random mean value (fixed for each trajectory), subject to Poissonian sampling noise as determined by the total number of data points and state weight in each trajectory. The number of hidden states was varied from two to six, with 2000 unique data sets for each number of states. For simulated data having a given number of states, the number of data points per trajectory was increased by 200 for every 80 data sets. Thus, the largest data sets have 5000 data points. In this notation, two states corresponds, for example, to one bright level and one background, or "off" level. Only one emission level per state was allowed.

Baum−Welch training was performed by the standard programs included in Matlab. New algorithms resulted from the modification of the Baum−Welch algorithm using chi-square-minimized Poissonian fits and chi-square probability-determined relative weighting to determine the number of hidden states. The determined "correct" number of hidden states was extracted from the quality of the fits as follows. For each given number of states, after Baum−Welch training, the trained emission matrix is fit to a Poisson distribution, and its initial weight is determined by the Viterbi algorithm. The emission level histogram is reconstructed by summing the properly weighted Poissonian emission levels, each of which is fit to the best Poisson emission level through chi-square minimization. The chi-square is calculated between the real histogram (simulated or experimental data) and the reconstructed histogram. The Poissonian-constrained emission matrix and transition matrix are used as the initial system parameters for the next step. The iteration is continued until the global chi-square is minimized for a given number of states.

Although each level is fit with chi-square minimization to a Poisson distribution of intensities, two different criteria are used to determine the overall goodness-of-fit between the actual data and the reconstructed data—BIC and chi-square probability. BIC is calculated by

$$\mathrm{BIC} = 2\ \text{log-likelihood} - d \ln N_p \tag{1}$$

in which $d = S^2 + S - 1$ is the number of parameters, N_p is the number of data points, and S is the number of states. The chi-square between the observed value X_i and expected value μ_i is defined by eq 2.

$$\chi_n^{\,2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \tag{2}$$

where n is the number of observables and σ_i is the standard deviation associated with the uncertainty in X_i.[22] Conversely, a chi-square variable, t, is governed by the following probability distribution function.

$$\text{probability} = \frac{t^{n/2-1} e^{-t/2}}{2^{n/2}\Gamma^{(n/2)}}, \quad t = \chi_n^{\,2} \tag{3}$$

In this paper, X_i is the probability of observable i in the experimental emission matrix, and μ_i is that in the Poissonian-constrained emission matrix. For a given number of states, the training is stopped when the chi-square value between the experimental and reconstructed histograms is minimized. Once a minimum for a given number of states is obtained, the number of states is changed by one and the global chi-square is again minimized. Since the degrees of freedom are related to the number of states, we used the integrated area of the chi-square distribution at the calculated chi-square value instead of the value itself. In this process, X_i is the number of occurrences of observable i in the real histogram and μ_i is that in the Poissonian-reconstructed histogram. This provides a method for comparing the goodness-of-fit for different numbers of hidden states.[25]

Especially for short data sets, the above process often overfits the number of states, so a penalizing term giving a statistical measure of level distinguishability was incorporated by checking the overlaps of all emission matrix curves by LE. LE is defined by[23,24]

$$\langle (\Delta x)^2 \rangle = \frac{\sigma^2}{N} \tag{4}$$

In optical localization experiments, Δx is the LE, σ is the standard deviation of the point-spread function, and N is the number of collected photons. N and σ were replaced by the integrated area of the emission matrix of an individual state (its weight, or the number of observations in that state) and the range of data values, respectively. When Δx is smaller than the difference between the mean values of two curves, overlapping curves were deleted, and the fitting process is performed again using the remaining curves as the initial parameters. Typically, the fitting regenerates the overfitted results, with overlapping distributions that are indistinguishable by localization error. Therefore, we terminated the iterations when the overlap appears. The previous best-fit number of states is then used as the best global fit. We denote the new algorithms modified by Poisson fit as PB when BIC is used to determine the dimension and PC when chi-square probability is used. In the case that LE is applied, the algorithms are represented as PBL and PCL, respectively.

Results and Discussion

Although quite fast, the traditional Baum−Welch algorithm simultaneously requires a great deal of data for adequate training and is prone to trapping in local minima and overfitting.[26,27] Figure 1A illustrates a typical result of getting trapped in local minima. During the training iteration, the log-likelihood monotonously increases, as shown in Figure 1D (open circles). However, due to the lack of physical constraints and robust stopping point of the iterations, the resulting emission matrix curve is noisy and has indistinctly defined levels.

For time correlated single photon counting data, emission intensities should be Poisson-distributed. Constraining the Baum−Welch algorithm to be physically reasonable demonstrates the advantage of the Poisson-modified Baum−Welch

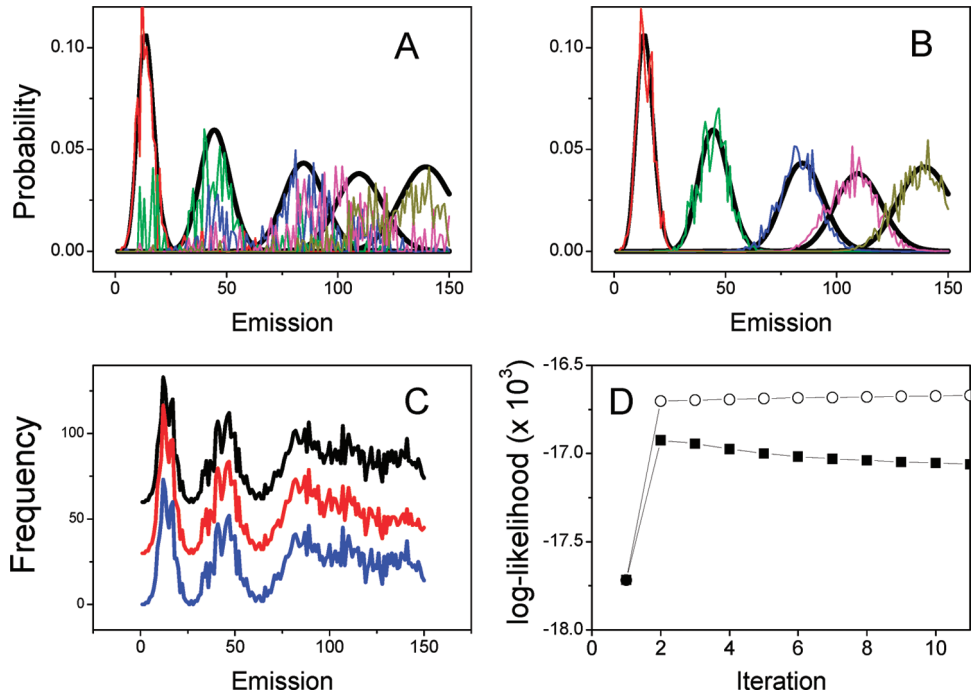

Figure 1. Comparison of traditional and Poisson-modified Baum−Welch algorithms. Emission matrices were trained by Baum−Welch (A) and Poisson-modified Baum−Welch (B) algorithms. Histogram (C) of a data set which has 3600 data points and 5 hidden states (top) and reconstructed histograms from the Baum−Welch (middle) and Poisson-modified Baum−Welch (bottom) algorithms. Histograms from reconstructed data are vertically offset from the simulated intensity histograms for clarity. (D) Log-likelihood from Baum−Welch (circles, top) and Poisson-modified Baum−Welch (squares, bottom) algorithms of the same data set during the training iteration.

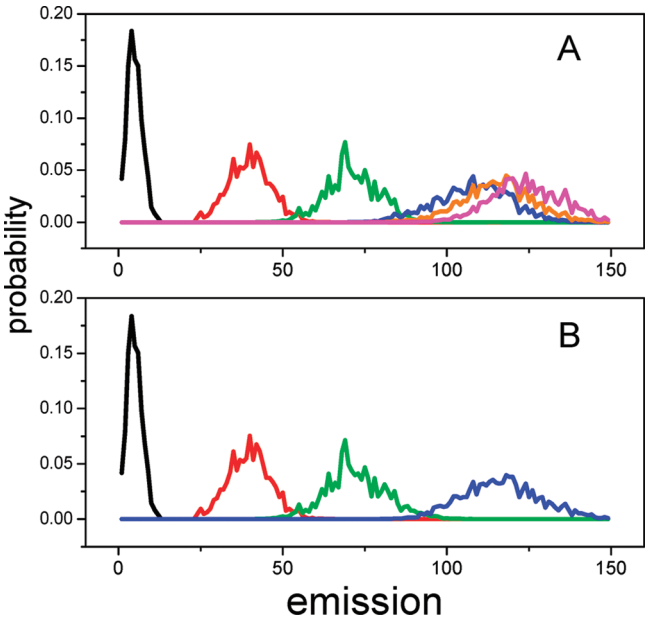

Figure 2. Modification by localization error (LE). Three curves at around 120 counts/bin in part A are consolidated into one curve in part B.

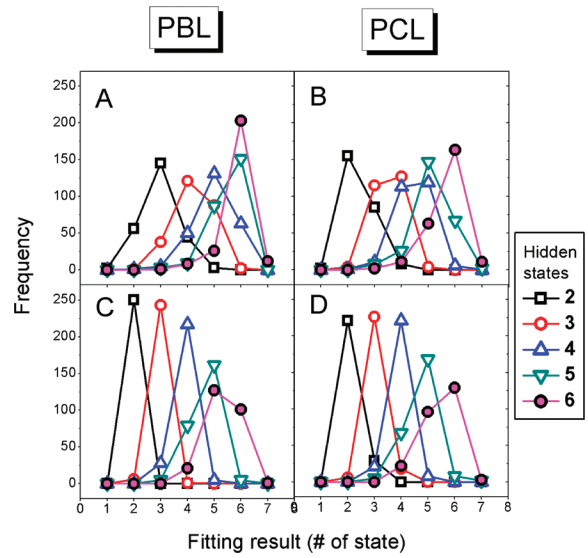

Figure 3. Histograms of fitting results calculated by PBL (A, C) and PCL (B, D) algorithms. Each data set was simulated to have 2, 3, 4, 5, or 6 hidden states. The number of data points was varied from 200 to 5000. The initial parameters for the top (A, B) and bottom panels (C, D) were acquired automatically and manually, respectively.

algorithm (Figure 1). A simulated data set with 3800 data points and 5 hidden states was generated on the basis of an emission matrix consisting of 5 Poisson distributions. Compared to the emission matrix estimated by the Baum−Welch algorithm, the shape and peak position of the emission matrix fit by the Poisson-modified Baum−Welch algorithm was much closer to the original emission matrix (shown in black solid lines) that was used in generating the data set. The overall emission level histogram of the simulated data set and that of the Baum−Welch and Poisson-constrained Baum−Welch reconstructions are nearly identical, but the Poisson constraints significantly improve the individual intensity distributions while simultaneously providing a clear maximum in the likelihood as a stopping point (Figure 1D). Such Poisson constraints on the emission levels enable more robust and more physically meaningful fits with better-defined stopping points.

Even with intensity levels constrained to be Poisson-distributed, however, the maximum likelihood method often overfits the data. For example, Figure 2A shows three curves with slightly different mean values being fit to a single emissive

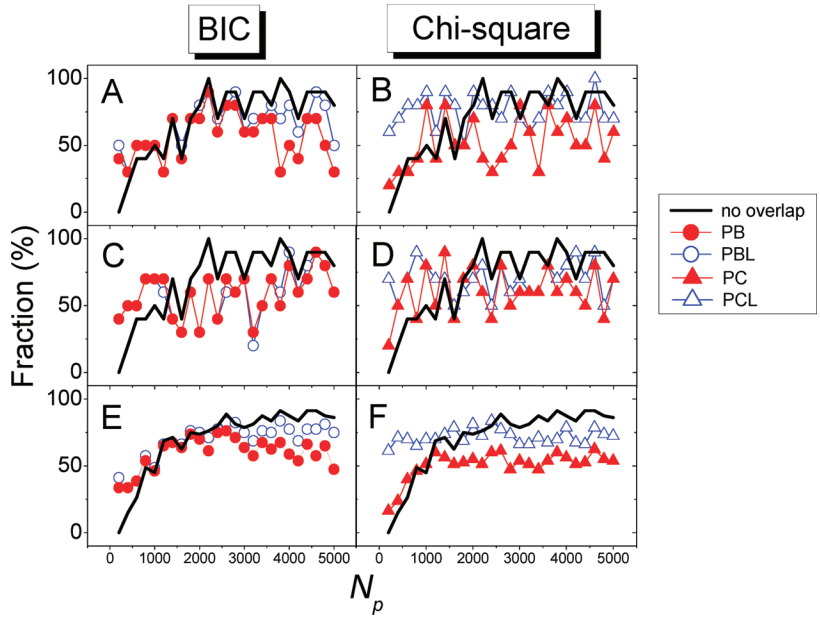

Figure 4. The effect of N_p on the inherent number of states and accuracies of the four algorithms in analyzing six-state data sets. 1250 data sets (A−D) and 10000 data sets (E, F) were generated based on six-state emission matrices. Initial parameters were defined by an automatic peak finding algorithm (A, B, E, F) or manually (C, D). Solid lines show the proportion of six-state data sets that have no overlap of curves in emission matrices. The accuracies were calculated by PB, PBL, PC, and PCL algorithms; see legend.

level near 120. The number of data points in each distribution, however, is quite small, suggesting that the three curves may not be significantly different but instead are consistent with a single distribution. Rather than eliminating overlaps visually, LE was introduced as a statistic to tie the actual weight of each state to the precision with which the distribution center can be determined. Using the standard deviation for the appropriate Poisson distribution, Figure 2B, for example, shows the advantage of using LE. Using this statistic, the curves in Figure 2A around the emission value of 120 were determined to be consistent with the single curve centered at 120 in Figure 2B. An approximation for asymmetric distributions, the localization error works very well for higher intensities (> ∼20 counts/bin), where the differences between Poisson and Gaussian distributions are relatively small. This method still gives good results for low intensities and is readily adapted for other common distributions, if necessary. Together, this gives a meaningful method to determine the best fit for a given number of states. Comparing the goodness-of-fit for different numbers of states, however, demands inclusion of additional criteria.

Figure 3 shows the histogram of fitting results from 1250 data sets generated using Poissonian emission matrices with randomly assigned mean values. Both BIC and chi-square probability (eq 3) were compared as criteria for goodness-of-fit when varying the number of hidden states. Both incorporate more penalizing terms as more states are included to yield better-defined stopping points than maximum likelihood alone. Further, the incorporation of LE significantly reduced the tendency of PB and PC to overfit the number of states, especially for short data sets. The number of hidden states was consequently estimated by the PBL (Figure 3A,C) and the PCL algorithms (Figure 3B,D), respectively, for a large number of simulated data sets. Initial parameters were determined in two ways. First, we used built-in peak finding codes in Matlab to identify initial emission levels from the histogram of each data set. Alternatively, we set the initial levels by eye from the histogram. The first method is automatic and much faster than the second one. However, frequently, the automated method did a poor job in predicting emission levels, especially when levels had close

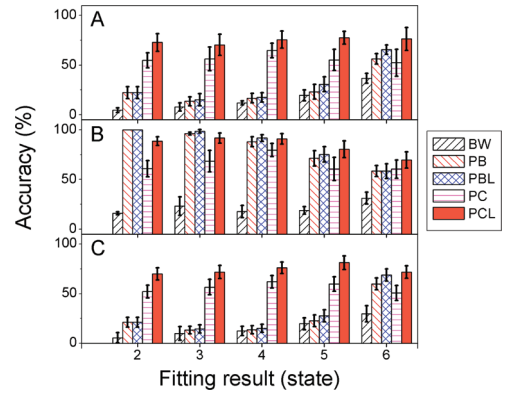

Figure 5. Accuracy bar plots of five kinds of criteria, BIC with Baum−Welch (BW), PB, PC, PBL, and PCL algorithms. 1250 data sets were analyzed using automatically (A) and manually (B) defined initial parameters. (C) 10000 data sets were also computed with automatic initial parameters. The error bars show the standard deviation calculated from 40 sets of 50 accuracy results.

average values or for low numbers of counts per bin. Therefore, the accuracies of fitting results by manual initiation (Figure 3C,D) are much higher than that by automatic initiation (Figure 3A,B). Interestingly, Figure 3A and B demonstrates that, in the case of automatic initiation, the determined "true" number of states by the PBL algorithm is larger than that from the PCL algorithm. These results are an example of the property of BIC; i.e., BIC predicts the maximum number of probable dimensions in the limit of infinitely long data sets.[18,28]

Not surprisingly, the fitting performance of all algorithms is highly dependent upon initial parameters, but the PCL algorithm appears the least sensitive to poor initial guesses (Figure 3). Additionally, the algorithm to automatically generate a new emission matrix with one more or one less state did not work as well as did manual input. When we manually input a suspected level in every simulation, all Poissonian-modified algorithms including PB, PBL, PC, and PCL algorithms tended to predict the correct dimension more frequently.

A significant fraction of incorrect fitting results for all algorithms arises from the shorter data sets. The dependence of accuracy on the length of the six-state data sequence is shown in Figure 3C. A solid black line in this figure means the proportion of data sets which were generated using six-state emission matrices and can be considered truly to have six hidden states (by localization error). The fraction is lower than 50% when the number of data points is smaller than 1000. This result can be explained in two ways. First, if we have too few data points, the system or molecule being measured may not access every possible state. Second, poorly sampled states have very large localization errors. Therefore, two low-occupancy adjacent curves in the emission matrix are likely to overlap, and two states are then consistent with a single level. In such cases, the fitting result tends to be smaller than the real answer. For these reasons, LE informs the setting of proper experimental conditions such as data collection time, bin width, or incident laser intensity. The accuracies of the PB and PC algorithms decrease with a decreasing number of data points. However, the PCL algorithm appears largely unaffected by the N_p, relative to the PB, PBL, and PC algorithms. As shown in Figure 4B−F, the accuracy of the PCL algorithm is larger than 60% even if the number of data points is smaller than 1000. These results illustrate the robustness of the PCL algorithm, suggesting great utility for short data sets like those that plague single molecule studies.

Figure 5 shows all fitting results and the percentage of correct answers from five different conditions: the Baum−Welch algorithm with BIC and the Poisson-modified algorithms using BIC and chi-square probability with and without LE. The error bars were determined by the standard deviation of 40 sets, each including 50 fitting results. The PCL algorithm (solid pattern) shows the best performance by t test with 95% confidence interval for short or Automatically Initiated Trajectories.

In conclusion, by modifying the Baum−Welch algorithm to introduce chi-square probability and localization error, we have improved HMM performance both in determining the dimensions of unknown systems and in robustness even if the intensity trajectory is very short, or if poor initial conditions are used. In these common experimental situations, the newly generated PCL algorithm can outperform even BIC with localization error for hidden Markov analysis of short trajectories.

Acknowledgment. The authors gratefully acknowledge financial support from NIH R01 GM068732.

References and Notes

(1) Rabiner, L. R. *Proc. IEEE* **1989**, *77*, 257.
(2) Baum, L. E.; Petrie, T.; Soules, G.; Weiss, N. *Ann. Math. Stat.* **1970**, *41*, 164.
(3) Durbin, R. *Biological sequence analysis: probalistic models of proteins and nucleic acids*; Cambridge University Press: Cambridge, U.K., New York, 1998.
(4) Beausang, J. F.; Zurla, C.; Manzo, C.; Dunlap, D.; Finzi, L.; Nelson, P. C. *Biophys. J.* **2007**, *92*, L64.
(5) Smith, D. A.; Steffen, W.; Simmons, R. M.; Sleep, J. *Biophys. J.* **2001**, *81*, 2795.
(6) Qin, F. *Biophys. J.* **2004**, *86*, 1488.
(7) Milescu, L. S.; Yildiz, A.; Selvin, P. R.; Sachs, F. *Biophys. J.* **2006**, *91*, 3135.
(8) McKinney, S. A.; Joo, C.; Ha, T. *Biophys. J.* **2006**, *91*, 1941.
(9) Lee, T.-H. *J. Phys. Chem. B* **2009**, *113*, 11535.
(10) Messina, T. C.; Kim, H.; Giurleo, J. T.; Talaga, D. S. *J. Phys. Chem. B* **2006**, *110*, 16366.
(11) Andrec, M.; Levy, R. M.; Talaga, D. S. *J. Phys. Chem. A* **2003**, *107*, 7454.
(12) Talaga, D. S. *Curr. Opin. Colloid Interface Sci.* **2007**, *12*, 285.
(13) Hu, D.; Liu, R.; Zeng, X.; Kaplan, S.; Lu, H. P. *J. Phys. Chem. C* **2007**, *111*, 8948.
(14) Akaike, H. *IEEE Trans. Autom. Control* **1974**, *19*, 716.
(15) Hannan, E. J.; Quinn, B. G. *J. Roy, Statist. Soc. B* **1979**, *41*, 190.
(16) Schwarz, G. *Ann. Stat.* **1978**, *6*, 461.
(17) Zhang, K.; Chang, H.; Fu, A.; Alivisatos, A. P.; Yang, H. *Nano Lett.* **2006**, *6*, 843.
(18) Watkins, L. P.; Yang, H. *J. Phys. Chem. B* **2005**, *109*, 617.
(19) Lanterman, A. D. *Int. Stat. Rev.* **2001**, *69*, 185.
(20) Xiaolin, L.; Parizeau, M.; Plamondon, R. *IEEE Trans. Pattern Anal. Machine Intell.* **2000**, *22*, 371.
(21) Levinson, S. E.; Rabiner, L. R.; Sondhi, M. M. *Bell Syst. Tech. J.* **1983**, *62*, 1035.
(22) Bevington, P. R.; Robinson, D. K. *Data reduction and error analysis for the physical sciences*, 2nd ed.; McGraw-Hill: New York, 1992.
(23) Ghosh, R. N.; Webb, W. W. *Biophys. J.* **1994**, *66*, 1301.
(24) Thompson, R. E.; Larson, D. R.; Webb, W. W. *Biophys. J.* **2002**, *82*, 2775.
(25) Barnes, J. W. *Statistical analysis for engineers and scientists: a computer-based approach*; McGraw-Hill: New York, 1994.
(26) Baldi, P.; Chauvin, Y. *Neural Comput.* **1994**, *6*, 307.
(27) Shatkay, H.; Kaelbling, L. P. *Proc. IJCAI* **1997**, 920.
(28) Leroux, B. G. *Ann. Stat.* **1992**, *20*, 1350.

JP907019P