Analysis of the Entire Sequence of a Single Photon Experiment on a Flavin Protein†

James B. Witkoskie and Jianshu Cao*

Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Received: July 28, 2007; In Final Form: December 18, 2007

The large amount of statistical data collected by single biomolecule experiments often demonstrates complex and non-Markovian relaxation over many time scales. Analyzing and interpreting these data is a major challenge because of the inherently statistical noise and the lack of definite theoretical descriptions or computer simulations on biologically relevant time scales. This paper reports one of the first complete sequence analyses of a single photon experiment on the flavin protein to determine an underlying physical picture for protein motions on the millisecond to second regimes. The robustness of Bayesian information analysis combined with the nonparametric maximum entropy method (MEM) incorporates all available information of the single-molecule data sequence and maximizes our ability to test the legitimacy of possible models. Our analysis of the experimental data is consistent with the stochastic Gaussian diffusion model where the slow protein motions are modeled as a collection of over-damped diffusive normal modes and reveals non-universal and distinct dynamic features that are specific for protein functions.

I. Introduction

Single-molecule methods are widely applied to the study of biomolecules.[1−11] The dynamics of biomolecules revealed by these single-molecule techniques are complex with fluctuations on many time scales. Data collected from these experiments are inherently noisy with contributions from background photons, the photon statistics of the system (shot noise), and the stochastic nature of protein dynamics.[12−17] These stochastic contributions to the data cause difficulties in interpreting single-molecule data and necessitates the applications of robust statistical methods.

This paper uses a Bayesian/information theory framework[18] to examine a possible model for a flavin protein (Fre) experiment by Xie and co-workers.[19] The experiment collects single photons emitted from the system and shows photon correlations up to 100 ms time scales. Maximum entropy analysis (MEM) shows long-time multiexponential relaxation.[20,21] The MEM fit avoids using a predetermined functional form, so it does not introduce artificial physics through a parametric fit. Since many models may result in the same correlation functions, such as a Gaussian diffusion model or a two-state model with complex waiting times, the correlation function only contains a limited amount of information relevant to understanding the mechanisms of biomolecules. Confirmation of the validity of the physical picture motivates the examination of the entire data sequence. By combining physical insight with statistical methods, this paper shows that modeling the protein's motion as a collection of over-damped diffusive normal modes is consistent with the entire data sequence. Although diffusion in multidimensional Gaussian potentials can be cast into a generalized Langevin equation with a smooth relaxation spectrum for the correlation function of the random force, the picture of a connected network of amino acids has physical appeal since it explains the Gaussian nature of the long-lived correlations. This harmonic diffusion picture is a dynamic analogue of the elastic network models used in determining the static root-mean-squared (rms) displacements of functional groups in proteins.[22−25]

This analysis is one of the first to utilize the entire data sequence of a single photon experiment on a single protein to determine a physical picture. Since the model fit incorporates all of the available information, one maximizes the ability to test the legitimacy of models. The robustness of Bayesian analysis combined with the nonparametric MEM analysis gives a complete description of the probed dynamics of Fre. After discussing the experiment in Section II, we perform simple preliminary statistical analysis in Section III to extract the limited information contained in correlation functions. This information is used to explore possible models in Section IV. We establish the Gaussian diffusion model as an appropriate model for the system and discuss the reasons that an *N* state or trapping model is not a natural choice. In Section VI, we determine the adequacy of the Gaussian diffusion model to fit all possible statistics through Bayesian analysis of the entire sequence.

II. Description of Experiment

The Fre experiment examines a single flavin protein attached to a cover slip by exciting an electron in the flavin with a repetitive sequence of laser pulses. As shown in Figure 1a, the excited electron can relax through the emission of a photon or through a two-step electron-transfer process between a nearby tyrosine, tyr^{35}, and the flavin molecule. The kinetic scheme associated with this system is shown in Figure 1b. The fluorescence rate is $k_f \approx 0.2\ ns^{-1}$. The first electron-transfer rate is a dynamic quantity that fluctuates around $k_{ET}(t) \approx 1.0\ ns^{-1}$. The second electron-transfer rate does not affect the ability to fluoresce and can be neglected. The experiment continually excites the flavin molecule with a pulse train separated by 13.2 ns, (Figure 1c). As depicted in Figure 1c, the experiment detects the first photon and records the arrival time of this photon, t_i, and the fluorescence lifetime, τ_i. The arrival time is the time difference between the excitation pulse time and the photon arrival time. Figure 1d is a record of the arrival time versus chronological time for a short piece of the time sequence. This

† Part of the "Attila Szabo Festschrift".

* Corresponding author.

10.1021/jp075980p CCC: $40.75
Published on Web 02/12/2008

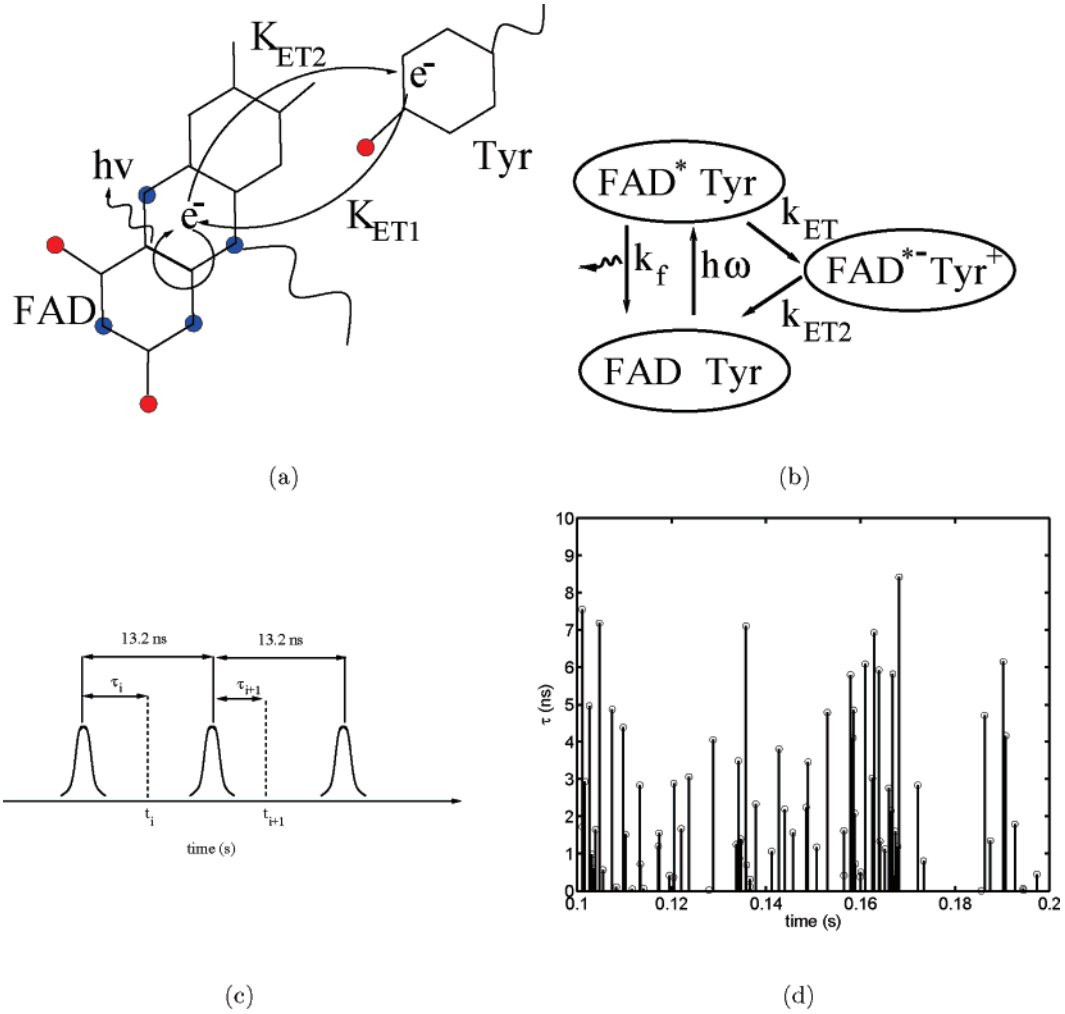

Figure 1. (a) The two competing mechanisms for relaxation of an excited electron to the ground-state−photon emission and electron transfer. (b) The corresponding kinetic scheme. (c) Schematic of the pulse trail that defines the chronological time, t_i, and the photon arrival time, τ_i. (d) Trace of the photon arrival time as a function of chronological time from the experiment.

sequence demonstrates the anticorrelation between the interarrival time $t_i - t_{i-1}$ (the inverse photon density) and the arrival time, τ_i. The probability of detecting a photon is proportional to the photon arrival time,[19]

$$P_{\text{photon}}(t|\tau(t)) \propto \frac{\tau(t)}{\tau_f} = \frac{(\tau_f^{-1} + \tau_{ET}^{-1}(t))^{-1}}{\tau_f} \qquad (1)$$

Experiments reveal the exponential dependence of the electron-transfer rate on the distance between the flavin and a specific tyrosine, tyr[35], $\tau_{ET}(t)/\text{ns} = e^{\beta r(t) - \beta r_0} = e^{R(t)-R_0}$, where R_0 accounts for the prefactor, and $\beta \approx 1.4$ Å^{-1} is the empirically determined scaling coefficient.[19] Following Xie and co-workers, $\tau_{ET} \ll \tau_f$ implies

$$\tau(t)/\text{ns} \approx e^{(R(t)-R_0)} \qquad (2)$$

The objective of this paper is the determination of the equations of motion for this coordinate $R(t)$.

III. Analysis with the Maximum Entropy Method (MEM)

To gain insight into viable models for $R(t)$, we visualize the data through one-dimensional measurements. The data is preprocessed to remove systematic errors, including monotonic intensity fluctuations in a peak corresponding to scattered photons from the laser source and a drift in the zero time baseline for the lifetime. Then, the photons are binned in 1 ms time bins. The lifetime is assumed to be static on this time scale. The background photon rate of $\lambda_b \approx 0.414$ photons/bin was measured after the experiment, and the rate for the molecule plus background is $\langle\lambda_s\rangle + \lambda_b \approx 0.781$ photons/bin. After preprocessing, measurements with different segments of the sequence are consistent (stationary), and the background measurements show no correlations.

A. Static Lifetime Distribution. The photon statistics are complicated by the background counts contributing over half of the photons ($\approx 58\%$), and by the photon's arrival time, τ_i, being a random variable that depends stochastically on the lifetime of the system

$$P_s(\tau_i|\tau_{ET}(t)) \approx \frac{1}{\tau_{ET}(t)} e^{-\tau_i/\tau_{ET}(t)} \qquad (3)$$

which is convolved with the instrument response. These complications necessitate the examination of several averaged measurements to develop insight into possible models for this system. The first averaged measurement is the static distribution of the fluorescence lifetimes. We perform this measurement for both the experiment and the background measurement to determine the typical lifetimes of photons emitted by the chromophore. The photon lifetime distributions for the experiment and background measurements are histogrammed in Figure

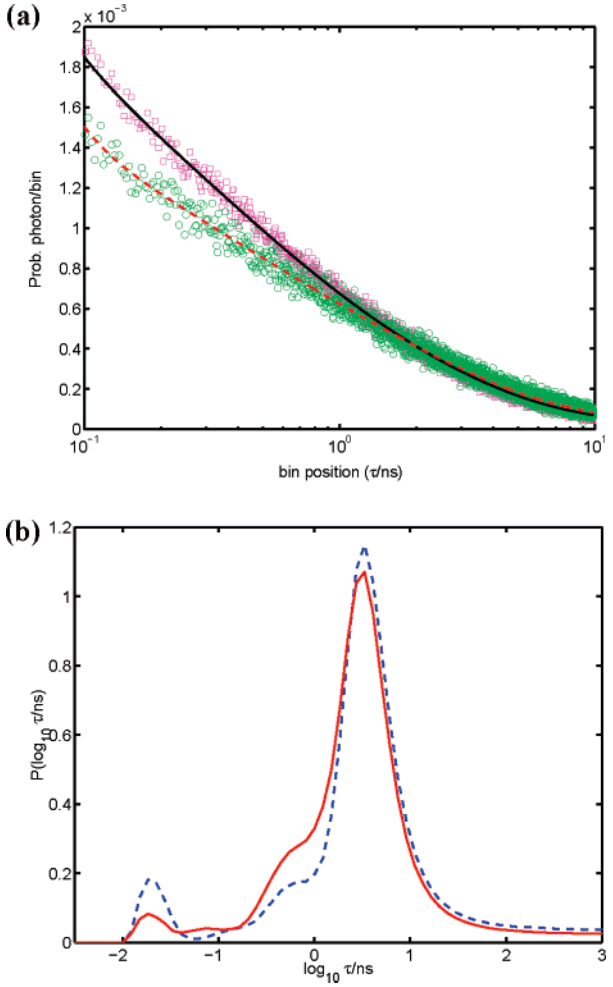

Figure 2. (a) The MEM fit to the experimental (squares = data, black solid line = fit) and background photon lifetime measurements (circles = data, gray dashed line = fit). The two curves show similar long lifetime behavior, but differ in the intermediate times. The scatter in the data gives a good indication of the error bars that are not plotted for visual clarity due to the large number of measurements. (b) The MEM spectrum for the fits. The fits differ in the amplitudes of the intermediate time shoulders, $0.1 \leq \tau/\text{ns} \leq 1.0$.

2, which shows maximum entropy fits (MEM) to sums of exponentials.[20,21] MEM attempts to balance the ability to fit the data with the desire to have a featureless spectrum. This fitting scheme does not impose an arbitrary functional form on the data.

The distributions and MEM fits of the single-molecule experiment and background measurement differ quantitatively, but no strong features differentiate the two lifetime distributions (See Figure 2). Both MEM fits show a small peak at short lifetimes, $\tau \ll 1$ ns and a broad peak at $\tau \approx 10$ ns caused by the instrument response with a shoulder at shorter lifetimes. The shoulder is larger for the single-molecule experiment's data and indicates that many of the photons from the protein occur on the short-time side of this peak, 0.1 ns $< \tau <$ 1.0 ns. Our MEM fits differ from those of Xie and co-workers, which show well-defined peaks corresponding to several different times. The disparities are the result of different data truncations and possible differences in the definition of σ_i used in the χ^2 statistics. The different σ_i values result in different stop criteria and affect the resolution of features. The lifetime analysis reveals the important range of lifetimes and the short-time shoulder of the distribution, but obvious signatures of the fluorophore remain hidden. Histogramming all of the lifetimes and performing a MEM fit is a more robust estimate of the lifetime than estimating the lifetimes from bins of 100 photons.[19] Binning every 100 photons reduces temporal resolution to approximately 0.1 seconds, and much of the photon correlation decays in this time period. Similar considerations of photon statistics and data binning have been addressed by Talaga and co-workers.[26]

B. Intensity Correlation Function. To gain insight into the dynamics of the system, we examine the 1 ms discretized trajectory to determine correlations between the number of photons in each bin. The objective of this paper is to relate these temporal correlations in the intensity to the underlying dynamics, which is dominated by the fluorophore−quencher distance, $R(t)$.[19] If the fluorophore−quencher distance, $R(t = j\Delta t) = R_j$ is constant over the $\Delta t = 1$ ms bin, the number of photons is Poisson, with parameter $\lambda(j) = \lambda_b + \lambda_s(j)$,

$$P(n|R_j) = \lambda(j)^n/n!\mathrm{e}^{-\lambda(j)} \tag{4}$$

where λ_b accounts for the background counts, and $\lambda_s(j) = A_0\mathrm{e}^{R_j}$ with prefactor A_0. Instrument considerations slightly modify these expressions. For a Poisson process, the second moment for the number of photons in any two bins, i and j, has the form

$$M_\lambda(j,k) = \langle n_j n_k \rangle = \langle \lambda(j)\lambda(k) \rangle + \delta_{jk}\langle \lambda \rangle \tag{5}$$

with Kronecker delta δ_{jk} denoting an additional white noise term, and $\langle ... \rangle$ is the expected value. After preprocessing to remove systematic instrument errors, the data is translationally invariant and allows time averaging. Subtracting the squared average, $\langle \lambda \rangle^2$ and $\delta_{jk}\langle \lambda \rangle$ gives the λ-correlation function, plotted in Figure 3a.

$$C_\lambda(j) = \frac{M_\lambda(0,j) - \delta_{0,j}\langle \lambda \rangle - \langle \lambda \rangle^2}{M_\lambda(0,0) - \langle \lambda \rangle - \langle \lambda \rangle^2} \tag{6}$$

To identify the time scales in this system, we perform a maximum entropy fit (MEM). The MEM fit avoids using parametrized models that can hide certain features in the data. The resulting spectrum reveals three time regimes (Figure 3c) The fastest time scales correspond to correlations that fall off within a few time bins (less than 20 ms). This time scale appears to be broadly distributed since few data points contribute to determining these parameters. A less broadly distributed second time scale decays around 50−100 ms, and a third narrowly distributed time scale decays around 400 ms. The narrow distribution at long times demonstrates an exponential decay of the correlation function at longer times. A fit to the fractional Gaussian noise (FGN) model to the first 3000 data points shows poor agreement, especially for the short-time behavior ($t <$ 100 ms)[27] (see Figure 3a). The $\chi^2 = N^{-1}\sum(x_i - \mu_i)^2/2\sigma_i^2 = 1.3$ for the FGN model, which shows that the fit is outside the 95% confidence interval, $\chi^2 = 1$, used to choose the MEM solution. The large number of data points in the tail force the FGN model to neglect the short-time correlations. The parameters predict that the mean squared (MS) fluorophore−quencher distance fluctuation is approximately 1.2 Å^2, which is significantly larger than experimental measurements.[19,28] This large MS displacement reduces the amplitude of the long-time power-law tail to less than 2% of the correlation function. As a result, exponential long-time relaxation fits the data better, but FGN has reasonable long-time agreement (see Figure 3b). As discussed in Section VI, multiple diffusive Gaussian modes are consistent with this MEM spectrum and gives MS displacements of 0.32 ± 0.021 Å^2, which agrees with experimental measurements. It is possible

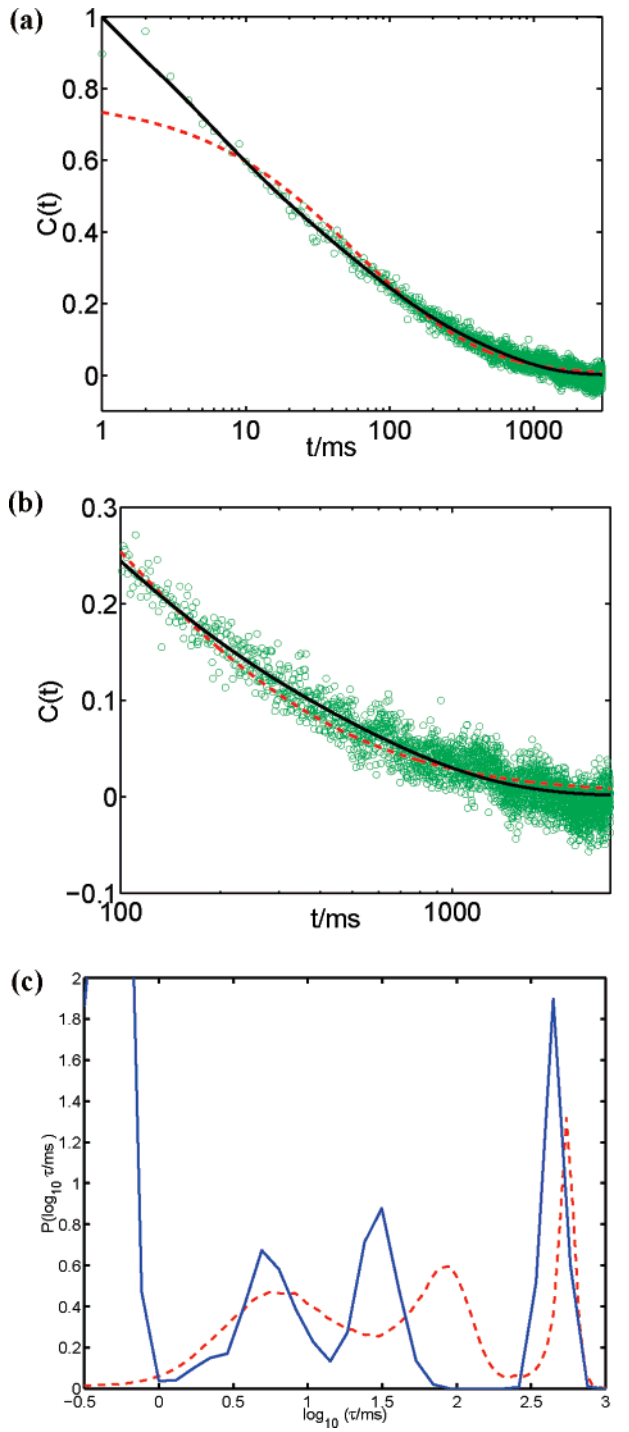

Figure 3. (a) Comparison of the MEM fit (solid line) to the FGN model (dashed line). The FGN model correlation function has the form, $C_\lambda = A(e^{C_{FGN}(t)} - 1)$ with $C_{FGN} = B(\sum_n((t/\tau)^{\gamma n}/\Gamma[n\gamma + 1]))$. The FGN fit gives $\tau = 263$ ms, $\gamma = 0.84$, $B = 2.32$ and $A = 0.084$. (b) The tail of panel a. (c) Comparison of the MEM spectrum (dashed) to the Bayes spectrum averaged over the MC data (solid). Except for the fast fluctuations that cannot be captured by the simple correlation analysis, the Bayesian peaks overlap with the peaks in the MEM spectrum, showing that they are consistent. We normalize the spectrum to the 1 ms bin contribution since the zero time correlation cannot be accurately measured.

to achieve a better fit with the FGN model to the first few hundred data points (up to tenths of a second), but not the entire time range of interest. We emphasize that the MEM fit does not assume a functional form and favors a less structured relaxation spectrum, such as a power-law or stretched exponential, over a structured spectrum. As a result, one should have confidence that the data reflects these structures.

Millisecond motions such as those captured by the MEM analysis have been observed in several fluorescence and NMR experiments and have been attributed to loop rearrangements, breathing motions in β sheets, rigid body motions of α-helices, and internally hindered rotations.[23,24,29] These low-frequency millisecond motions often play pivotal roles in protein function, so the ability to resolve and model these motions is important.[30] Although the tyr^{35}−flavin distance may not play a key role in functionality, other motions coupled to this displacement may. The MEM analysis suggests that any physical model for the tyr^{35}−flavin coordinate must reflect both the small-scale fluctuations of $R(t)$ and the structured relaxation spectrum. It is also important to capture the non-Markovian fluctuations in the intensity and fluorescence lifetimes (as demonstrated in Figure 1d). The model must also account for the distribution being stationary after preprocessing the data, and no aging effects are present. Armed with these insights, we are now able to explore physically feasible models for this system in Section IV.

IV. Slow Motions in Proteins: *N* State Models, Trapping Models, and the Gaussian Diffusion Model

Many candidate models can reproduce the intensity correlation function and the lifetime distribution, so the other physical attributes discussed above also need to be considered in selecting a model. The physical basis of a model depends on the level of coarse graining required to capture the essential physics of a system. This point is illustrated by the hierarchical tier picture of protein energy landscapes/surfaces (PESs).[31] The potential energy landscape is high dimensional and complex with motions on many length and time scales. The motions on tier m are generally faster than motions on tier $n > m$, but slower than motions on tier $n < m$, and time separation arguments generally apply. If the motions that we are monitoring occur at tier m, we can homogeneously average over the degrees of freedom associated with the lower tiers $n > m$ and need to perform a quenched average over the higher tiers $n < m$. The quenched average would result in heterogeneity in the behavior of single molecules. As shown in Figure 5, averaging over faster time scales results in a free energy potential instead of a detailed microscopic potential.[24] Considering that the experimental time scales of the Fre experiment range from milliseconds to tenths of seconds, these motions occur on the slowest time scales of the protein (no tiers $n < m$), so we do not expect additional slower motions that must be heterogeneously averaged over.

Three models that result from different coarse graining procedures include the *N* state model, the trapping model, and the Gaussian diffusion model. These models originate from different topologies of the protein potential energy surface. As elaborated below, the *N* state model results from the time scale of interest corresponding to motions over high-energy barriers, while the trapping model corresponds to hopping over many smaller structures. For the Gaussian diffusion model, the smaller scale structures result in a diffusion tensor.

A. The *N* State Model. The *N* state model results from the tier of interest containing multiple minima separated by high barriers (see Figure 5e,f). Averaging over the faster degrees of freedom results in Kramer's barrier crossing kinetics,

$$\dot{P}_{R(t)=R_i} = -KP_{R(t)=R_i} \tag{7}$$

where $P_{R(t)=R_i}$ is the probability that the particle is in minima i with a corresponding tyr^{35}−flavin distance of R_i. One may also

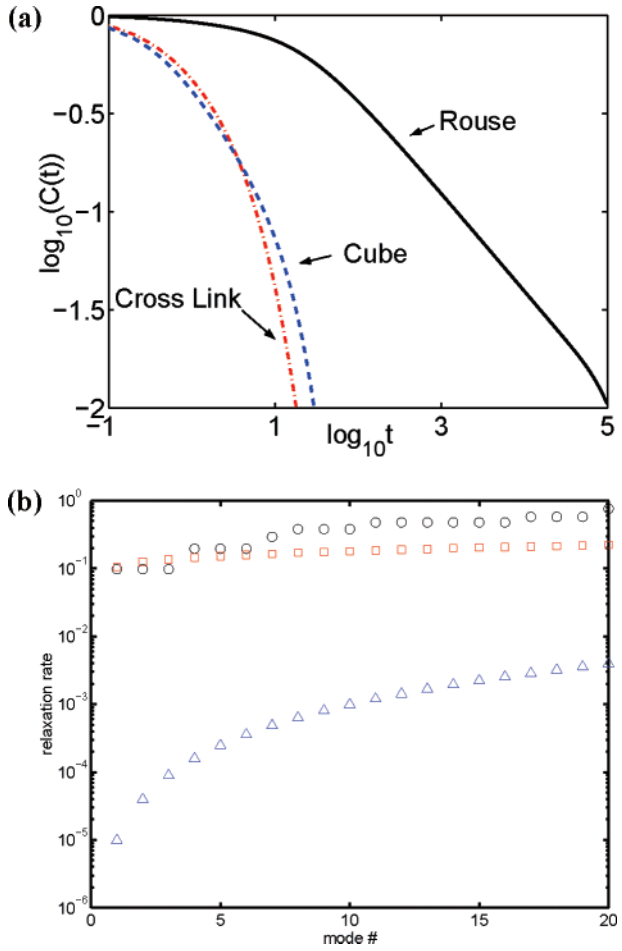

Figure 4. (a) The correlation function for $m = 493$ and $n = 507$ for a Rouse chain with 1000 beads, and the same correlation function averaged over random cross-linking of the Rouse chain polymer of 1000 beads with approximately 1 cross-link per polymer. These correlation functions are compared with the correlation function for the opposite corners, $i = j = k = 1,10$, of a $10 \times 10 \times 10$ cube. The cube also has 1000 beads, but it is a small object. The diffusivity, D and the force constants for all connected beads, k, are assumed to be unity $D,k = 1$. Except for the Rouse chain, the smallest relaxation rates $\lambda_1 \approx 0.1$, and the contribution of long-time exponential relaxation is significant. (b) The lowest eigenvalues for the Rouse (triangle), cross-linked polymer (square), and the cube (circle).

add fluctuating barrier heights that make K(t) time dependent.[32] The presence of three relaxation time scales suggests that a minimum of four states is necessary, but some of the kinetic rates would be slow, $\tau \sim 100$ ms, and one would expect deviations from Gaussian behavior that are not seen in the data. Instead, the data analysis below suggests that the additional non-Markovian fluctuations in $R(t)$ can be captured by a Gaussian model, although a model with both barrier crossing kinetics and intra-well relaxation may also be viable. These additional non-Markovian fluctuations can be captured by the inclusion of additional states. If enough states are included, the N state model can approximate any other model, but attempts to fit the data through complete sequence analysis with a reasonable number of states, $N > 6$, did not achieve a desired fit to the data. The identity of these states is also ambiguous since the apparent fluctuations in the chromophore quencher distance is rather small, on the order of tenths of an angstrom, compared to the larger scale motions that an N state model attempts to capture,[19] although these motions may simply have a weak projection onto the probed coordinate.

B. The Trapping Model. Unlike the N state model, where the barriers that dominate the dynamics occur on the same tier as the motion of interest, the trapping model has important contributions from the smaller scale motions (see Figure 5c,d). These small-scale structures trap or hinder the motions of the coordinates of interest, and the fluorescence lifetime becomes static for long periods of time. Fractional diffusion is an extreme example of this scenario, where the traps have energetic barriers that are exponentially distributed for large energy barriers, $P(E^{\dagger}) \sim \alpha e^{-\alpha E^{\dagger}}$.[33] The exponential decay of the energy barrier distribution is the result of extreme value arguments with a strong emphasis on the functional form in the tail of the distribution. This formulation is hindered by the tails being slow to converge to the universal form.[33] The scenario results in a long-time power-law decay and aging effects that are not seen in the data or the MEM fit.[33] We examined truncating the distribution of energy barriers, $P(E^{\dagger})$, but this truncation imposes an interrupted aging effect that removes contributions from the short-time trapping behavior and prevents the inclusion of a broad distribution of time scales in the stationary correlation function.[34] The trapping models also depend on the system being large so that correlations in the trapping times can be ignored, which probably does not apply to finite sized proteins. As a result, the data does not support a trapping model.

Since the exact distribution of configurations or traps that lead to trapping are unknown, one introduces the traps stochastically. The trapping model makes an annealed disorder assumption that neglects correlations in the trapping times or energy barriers. The assumption requires the degrees of freedom other than $R(t)$ to be weakly correlated to the coordinate of interest, except during trapping events. The trapping scenario is possible for the myoglobin experiments, where the CO molecule is moving in a cavity of the protein and occasionally encounters a site in the pocket that traps it for a period of time, but the FAD and tyrosine in this protein are connected through the scaffolding of the protein.[19,32] It is more appropriate to discuss motions of the entire protein including the FAD and tyrosine as a whole. Small-scale motions can create the a rough potential energy surface, which retards the motion of the protein, but allowing the lower tier structures to trap the system for extended periods of time is an extremely strong emphasis on these small-scale fluctuations.

C. Gaussian Diffusion Model. If the barriers are not high within the tier of interest, the system demonstrates a diffusive behavior. Averaging over faster degrees of freedom results in a smooth convex free energy landscape (see Figure 5a,b),[22,35] and the slow motions of the protein can be approximated by diffusion of a collection of independent normal modes whose correlation function is a simple exponential $\langle R_\mu(t)R_\nu(0)\rangle = \delta_{\mu\nu}(a_\mu^2/b_\mu^2)e^{-t/\gamma_\mu}$. The motion of interest is a weighted sum of modes, $R(t) = \sum_\mu b_\mu R_\mu(t)$, and the process becomes Gaussian with correlation function $\langle R(t)R(0)\rangle = \sum_\mu a_\mu^2 e^{-t/\gamma_\mu}$. By defining $\rho(\gamma) = \sum_\mu a_\mu^2 \delta(\gamma - \gamma_\mu)$, we can define a relaxation spectrum and write

$$\langle R(t)R(0)\rangle = C_R(t) = \int d\gamma \rho(\gamma) e^{-t/\gamma} \tag{8}$$

The probability distribution of $R(t)$ becomes a simple functional integral

$$P(\{R(t)\}) = \frac{1}{\sqrt{\mathrm{Det}(2\pi C_R(t))}} e^{-(1/2)\int dt_1 dt_2 R(t_1) C_R^{-1}(t_1 - t_2) R(t_2)} \tag{9}$$

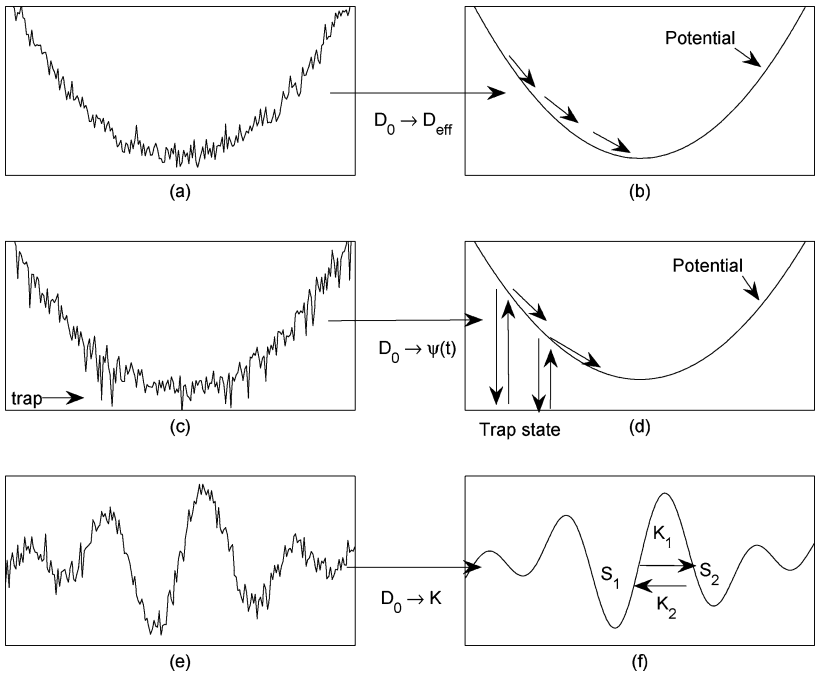

Figure 5. Schematic illustrations of the different potential energy surfaces (PESs) after averaging over faster time scales. (a) A PES with a global curvature and Gaussian roughness. Homogeneously averaging over the smaller length scales results in normal diffusion with a renormalized diffusion constant, $D_{\text{eff}} = D_0 e^{-\langle V2(0)\rangle/d}$, as shown in panel b. (c) A PES with an exponential distribution of traps, $P(E^{\dagger}) = \alpha e^{-\alpha E^{\dagger}}$, which results in deep spikes. Averaging over the smaller length scales replaces normal diffusion by a trapping process with a waiting time $\psi(t)$, as depicted in panel d. (e) A more complicated PES with several local minima. Averaging over small length scales results in an N state model with Poisson kinetics depicted in panel f.

This model can be cast into an N state model with many states, but the kinetic rates have simple relationships with each other, which reduces the number of parameters. Unlike the trapping model, which puts strong emphasis on small-scale fluctuations, the Gaussian diffusion model accounts for small-scale motions through the more modest means of a renormalized diffusion tensor. The diffusion tensor for a Gaussian distribution of energy barriers can be shown to be isotropic with $D_{\text{eff}} = D_0 e^{-\langle \beta^2 V(0)^2\rangle/d}$, where β is the inverse temperature, and $V(0)$ is the random potential. With this assumption for $R(t)$, the fluorescence intensity correlation function is approximately

$$C_{\lambda}(t) = A(e^{C_R(t)} - 1) \sim AC_R(t) + AC_R{}^2(t)/2! + \ldots \quad (10)$$

where A is a prefactor that accounts for A_0 and R_0, and the asymptotic relations result from the decay of the correlation function. This expression would be exact, except for experimental considerations such as instrument response and the finite detection window, but these deviations are small.

Large-scale coarse-grained normal modes have been used to study the dynamics and statistics of several systems.[24,36,37] The fast motions in proteins deviate from this model's simple harmonic motion, but the larger scale motions have been found to be independent of small-scale anharmonicities due to laws of large numbers and motional narrowing effects.[22,38] A Gaussian model is also the minimal information model consistent with the measured observables (the correlation function).[39] Even if deviations from harmonicity exist, the current data is not able to resolve characteristics of these deviations. As a result, the important features are characteristic relaxation times and the magnitude of the displacements of $R(t)$, which can be captured by a harmonic approximation.

1. Spectrum of the Correlation Function. The MEM analysis showed three time scales contributing to the system, which suggests that the major contribution to the long-time relaxation of the system can be captured by a Gaussian diffusion model with three relaxation time scales. Concurrent with our work, Kou and Xie suggested a Gaussian diffusion model with a power-law distribution of relaxation times, which they capture with the FGN model.[27] This model is interesting, but even ignoring the difficulty in fitting the dynamics, its application to proteins needs to be justified.

Gaussian noise assumptions have two possible sources. An unlikely source for this protein system is the bath being much faster than the time scales of interest, which leads to multiple collisions and a large number arguments. In this case, the structure of the bath does not matter, but long-lived correlations in the random force cannot be introduced. The other possibility is that the bath has an intrinsically harmonic structure, as argued above for coarse-grained descriptions of the protein. This coarse graining will add many time scales to the relaxation spectrum, and the major issue becomes the expected structure of the spectrum of relaxation times.

2. Power-Law Spectrum and Scaling. A power-law spectrum has been suggested for this system.[27] This power-law may be the result of the protein showing a self-similar structure. The simplest self-similar structure is the Rouse polymer chain of N beads at positions x_n with local connectivities.[40] The beads undergo diffusion in a potential of the form $V_{\text{Rouse}} = \sum_{i,a=x,y,z} (k/2)(a_i - a_{i+1})^2$. In the large monomer limit, $N \rightarrow \infty$, the eigenmodes $a_{\omega}(t) = (1/N)e^{i\omega n} a_n(t)$ have a correlation function

$$C_{\omega}(t) = \langle a_{-\omega}(t) a_{\omega}(0)\rangle \approx \frac{D}{2k(1 - \cos(\omega))} e^{-2k(1-\cos(\omega))t} \quad (11)$$

where D is the diffusion constant, k is the spring constant of the Rouse chain, the friction coefficient $\zeta = 1$, and the end effects were ignored.[40,41] The correlation function for $a_n - a_m$ is

$$C_{nm}(t) = \langle (a_n(t) - a_m(t))(a_n(0) - a_m(0)) \rangle = \int \frac{d\omega}{2\pi} \frac{D}{k} \frac{1 - \cos(\omega(m-n))}{1 - \cos(\omega)} e^{-2k(1-\cos(\omega))t} \sim \frac{D}{k}(m-n)^2 e^{-2kt} I_0(2kt) \quad (12)$$

which decays as $t^{-(1/2)}$ (see Figure 4).[40] More generally, the lattice could have a self-similar fractal structure, and the correlation would be expected to asymptotically decay as $t^{-d/2}$, where d is the fractal dimension. A physical model for fractal dimension was recently constructed by Klafter and co-workers to explore the dispersive dynamics of protein relaxation.[42] The power-law arises from the scale invariance of the system. By only having local connectivities, there is a translational invariance (if end effects are ignored). As a result, the relaxation times correspond to differing length scales resulting in the power-law. The power-law results will not be altered by adding bending rigidity or any other local interaction since the translational invariance will still hold in the large N limit.

3. Nonlocal Contacts, Finite Sizes, and Nonscaling Behavior. Proteins have additional interactions that are nonlocal with respect to the position along the protein sequence, so these self-similar structures are probably not a good model for a protein. These nonlocal interactions destroy the scale invariance, so a power-law spectrum cannot be universally applied to proteins. Although proteins may be large objects in terms of the one-dimensional sequence, the three-dimensional structure is much smaller than a crystal, so edge effects destroy the scale invariance necessary for a power-law. A protein with 1000 amino-acids (Rouse beads) would only form three-dimensional structure around 10 amino-acids across (on the order of nanometers).[43]

The smallest eigenvalues for the relaxation spectrum of a 1000-bead Rouse chain shows a near power-law behavior with very small eigenvalues, whereas the spectrum of an elastic body that is a 10 unit cube (also 1000 beads) shows a lower bound in the relaxation rates, (see Figure 4). The cube undergoes the same diffusion process as the Rouse chain, but the potential for the cube is

$$V_{\text{cube}} = \sum_{i,j,k=1\ldots n; a=x,y,z} \frac{k}{2} [(a_{i,j,k} - a_{i+1,j,k} - a^0_{i,j,k,i+1,j,k})^2 + (a_{i,j,k} - a_{i,j,+1,k} - a^0_{i,j,k,i,j+1,k})^2 + (a_{i,j,k} - a_{i,j,k+1} - a^0_{i,j,k,i,j,k+1})^2] \quad (13)$$

where $a^0_{i,j,k,i,j,k+1}$ denotes the equilibrium distance. In the large n limit, this expression is also exactly solvable with a t^{-1} power-law dependence, but the edge effects for $n = 10$ are quite significant. If force constant k and the diffusion constant D are assumed to be unity (unitless time and distance), the cube's spectrum (excluding rigid body motions) is bounded away from zero, $\lambda_1 \approx 0.1$, and long-time exponential relaxation is expected. Even a less structured connectivity such as the average eigenspectrum of a randomly cross-linked Rouse chain (1000 beads) shows an eigenspectrum that is bounded away from zero (Figure 4).[44] In this model, the potential is random with $V_{\text{cross-link}} = V_{\text{Rouse}} + \sum_{i>j; a=x,y,z} p_{ij}(k/2)(a_i - a_j)^2$, where $p_{ij} = 0,1$ with probabilities $(N-1)/N$ and $1/N$, respectively. This random linking results in a collapsed structure, so finite size effects are expected.[44] Proteins have specific nonlocal connectivities to allow them to perform their function, so a greater variety of behaviors and, subsequently, structures in the relaxation spectrum is expected relative to those discussed above. As a result, it should not be surprising if the relaxation spectra of proteins show structures that are unique to the protein.

V. Implementation of Bayesian Statistics

Applying Bayesian methods to physical systems often presents difficulties since models convenient for Bayesian analysis do not always have a desirable physical interpretation. We examine the diffusive harmonic model because it is physically justified, is computationally feasible, and does not require dramatic assumptions about the system. The model implies that the lifetime of the system at time $t = j\Delta t$ is given by a sum of harmonic modes, $\tau_j = e^{-R_0 + \sum a_\mu R_{\mu j}}$, and the intensity from the molecule during time bin j is $\lambda_s(j) = e^{-R' + \sum a_\mu R_{\mu j}}$, where a_μ is the weight of mode μ and $R_{\mu j}$ is the displacement of mode μ during time bin j. For this process, the C_λ correlation function is related to the correlation function of R_j, $C_\lambda(j) \propto e^{C_R(j)} - 1$, where $C_R(j) = \sum a_\mu^2 e^{-j/\gamma_\mu}$ is the correlation function for R, and the γ_μ values are the relaxation times. This expression is not valid for an N state or hopping model since they are not Gaussian. For the rest of the paper, i refers to a photon, j refers to a bin, and μ refers to a mode. Each mode has a relaxation time γ_μ. Our goal is to find the appropriate number of modes, M, their weights, $\{a_\mu\}$, their lifetimes, $\{\gamma_\mu\}$, and other parameters, R_0 and R' to best represent the data. We will denote these parameters by θ_M and perform the optimization through a Bayesian framework.[18] Following standard Bayesian modeling assumptions, the probability of a model, M, given the data, $D = \{D_j\} = \{\tau_{ij}\}$, is

$$P(M|D) \propto \sum_{s_j, \theta_M} P(D, \{s_j\}, \theta_M, \theta, M) \propto \sum_{s_j, \theta} \left[\prod_j P(D_j|s_j) \right] P(\{s_j\}|\theta_M M) P(M) \quad (14)$$

In the above expression, D_j denotes all photons in bin j, and s_j is the state during bin j. The probability for the data in a bin is the product of two components, the measured lifetimes of the photons (Section IIIA), and the number of photons in a bin (Section IIIB). These two contributions have complicated expressions due to background photon counts, the instrument response, and the truncated time windows, but numerically computing these probabilities can be implemented.[18]

If the Gaussian diffusion model is correct, the MEM analysis suggests that Gaussian diffusion with three distinct time scales can capture the long-time relaxation of the data. One can model the three time scales with three modes without introducing statistically significant errors. The remaining issue to address is the verification of the consistency of the model with the data set as a whole. Additional coarse-grained measures, including multiple time correlation functions, are too noisy to assess the model adequately, but the results are consistent with a Gaussian diffusion model for the motions of $R(t)$. Concurrent with our work, Kou and Xie also demonstrated this consistency by examining these averaged quantities, so we will not go into detail about these tests of the Gaussian hypothesis.[27] To strengthen the legitimacy of Gaussian diffusion, a full sequence Bayesian analysis is necessary.

The Bayesian analysis is implemented by fixing the number of modes, M, and performing a Monte Carlo (MC) simulation to sample the parameters of the model (the weights, a_μ, the relaxation times, γ_μ, and auxiliary parameters such as A and R_0) that determine the statistics of the system. For given parameters, the probability of having n_j photons with lifetimes $\tau_{j1} \ldots \tau_{jn}$ in bin j for $R(t) = R_j$ is computed as

$$P(\{\tau_{j1}...\tau_{jn}\}|R_j) = P(n|R_j)\prod_{m=1...n} P(\tau_{jm}|R_j) \quad (15)$$

where $P(n|R_j)$ is the probability of getting n photons as defined in eq 4, and $P(\tau_{jm}|R_j)$ is the probability of the arrival time of the photons given by R_j,

$$P(\tau_{jm}|R_j) = \frac{\lambda_b}{\lambda_b + \lambda_s(j)}P_b(\tau_{jm}) + \frac{\lambda_s(j)}{\lambda_b + \lambda_s(j)}P_{sys}(\tau_{jm}|\tau_{ET} = e^{(R_j - R_0)}) \quad (16)$$

Both sources of photons, the system, P_{sys}, and the background, P_b, are accounted for in this expression. The exact form of these probabilities is complicated by convolution with instrument response and other instrument considerations. Without these considerations, P_{sys} is equal to P_s in eq 3. Given the probabilities of the photon emission events for all R_j, the mode positions $R_\mu(t = j\Delta t)$ are varied by randomly choosing one mode and statistically choosing its positions $\{R_\mu(t = j\Delta t)\}$, keeping the other modes fixed through a forward−backward algorithm.[45]

From this simulation we estimate the Bayesian score (log of the probability that the model produced the data) to determine the optimal parameters.[45] The score includes how well the sampled paths fit the data and the probability that the paths are produced by the diffusion model. The fit to the data is estimated from the log of eq 15 for the selected sequence $\{R_j\}$, and the fit to the model is estimated from the Fourier components of $R_\mu(t = j\Delta t)$, $\sum R_\mu(t = j\Delta t)e^{i\omega j}$. The Bayesian score was computed for a different number of diffusive harmonic modes and compared to determine the appropriate number of modes.

VI. Results

The simulation found that a fourth mode is necessary to account for fast fluctuations that are not consistent with the stochastic fluctuations (including the background) and would not be represented in the correlation function. The time constants and weights of the four oscillators are $a_\mu^2 = 0.595 \pm 0.023$, 0.293 ± 0.040, 0.292 ± 0.028, and 0.324 ± 0.041 for $\gamma_\mu = 0.42 \pm 0.10$, 5.9 ± 2.8, 28.0 ± 8.2, and $400. \pm 57$, respectively. The exponential components discovered by the Bayesian simulation fall into the time scales revealed by the maximum entropy fits, which shows that Gaussian diffusion agrees with the basic features of the data and the Bayesian approach identifies the important time scales. The Bayesian spectrum is determined by averaging C_λ over the MC simulation and is compared against the MEM simulation in Figure 3c.

For $\beta = 1.4$ Å, the MS displacement of $r(t)$ is $\langle r^2(t)\rangle \approx 0.32 \pm 0.02$ Å^2, which is in agreement with other measurements.[19,28] Crystal structure data show that tyr[35] has an MS displacement of 0.25 Å^2 and that the isoalloxazine portion of FAD has an MS displacement of 0.10 Å^2, so fluctuations around 0.35 Å^2 are expected. For four modes, we predict that the average arrival time is around $\langle\tau\rangle \approx 0.310 \pm 0.011$ ns. Since the number of photons emitted depends on $\tau(t)$, we must weigh the probability of $\tau(t)$ by the expected number of photons given $\tau(t)$ to determine the average arrival time of a photon emitted by the FAD, $\langle\tau_{photon}\rangle \approx 0.410 \pm 0.028$ ns. As expected, this distribution suggests that the photons from the system occur on the short-time shoulder peak of the MEM distribution in Figure 2b.

Additional modes, beyond four, slightly improve the fit to the data and the correlation function, but the improvement cannot be justified statistically. For less modes, the paths selected by two- and three-mode models have similar likelihoods to the paths of the four-mode model, but the probability of these paths being produced by the harmonic model was much lower. In other words, following the variation in the data with less than four modes resulted in unlikely paths. The Fourier components of the sequence for the four-mode model, $\sum R_\mu(t = j\Delta t)e^{i\omega j}$, are within the expected variances of the model, so Gaussian diffusion is consistent with the paths that fit the data. Similarly, the photon emission events are consistent with the model. The Bayesian scores and parameters for four oscillators show time translational invariance, so the model is not over-fitting the data.

VII. Conclusion and Discussion

This paper examines a single photon experiment with a complex data set that is difficult to interpret from correlation analysis. Nonparametric fits by the MEM demonstrate a wide distribution of time scales with distinct structures in the relaxation spectrum that are neglected by phenomenological fits using smooth predetermined functional forms. The evidence for these structures (especially at long times) is strong, so it is appropriate to discuss distributed lifetimes, but the existence of a stretched exponential, power-law, or other phenomenological functional form cannot be fully supported by the data. Our analysis demonstrates the importance of introducing nonparametric methods into single-molecule data analysis and the need for caution in interpreting model features such as power-law tails. Models without these features may also be consistent with the data, so it may be difficult to assign a physical meaning to the predicted power-law. From the nonparametric analysis of coarse-grained measures, such as correlation functions, one can develop legitimate models to describe the behavior of the system. Although models should be consistent with the correlation analysis, correlation functions provide only one- or two-dimensional information and generally fail to distinguish different models. The desire for a comprehensive test motivated the use of Bayesian methods in analyzing the entire data sequence. These tests are more time-consuming than the correlation analysis, but the conclusions are more reliable.[27] However, one must remember that the excellent agreement between the data and the harmonic diffusion experiment may simply be the result of the insensitivity of measurements to non-harmonic features. Through a complete sequence analysis on a single molecular trajectory, this paper demonstrates that the Gaussian diffusive model with a few well-defined long time scales is a viable candidate for describing this system. These slow time modes may correspond to motions that influence protein structure and function.

Proteins are specific entities that perform specific tasks. The complexity of proteins may cause a broad distribution of time scales, but it is important to understand how the motions are associated with the specific tasks of the protein. Coarse-grained diffusive harmonic modes incorporate the universality of large numbers by averaging over small scale fast fluctuations while maintaining features that are specific to the protein's structure and function. The fact that coarse-grained models and simulations can capture these slower time scale motions while also being computationally tractable is a desirable feature that may allow comparison of simulation to experiment.

Acknowledgment. We dedicate the paper to Attila Szabo with admiration. The research is supported by the AT&T Research Fund Award, the NSF Career Award (No. Che-0093210), and the Camille and Henry Dreyfus Teacher-Scholar Award. We would like to thank X. S. Xie and G. Luo for

providing the data, R. J. Silbey for useful discussions, and Y. C. Cheng and T. van Voorhis for computer time. The work was first completed in 2004 and reported thereafter.

References and Notes

(1) Moerner, W. E.; Orrit, M. *Science* **1999**, *283*, 1670.
(2) Edman, L.; Mets, U.; Rigler, R. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 6710.
(3) Jia, Y.; Sytnik, A.; Li, L.; Vladimirov, S.; Cooperman, B. S.; Hochstrasser, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 7932.
(4) Lu, H. P.; Xun, L. Y.; Xie, X. S. *Science* **1998**, *282*, 1877.
(5) Weiss, S. *Nat. Struct. Biol.* **2000**, *7*, 724.
(6) Zhuang, X. W.; Bartley, L. E.; Babcock, H. P.; Russell, R.; Ha, T. J.; Herschlag, D.; Chu, S. *Science* **2000**, *288*, 2048.
(7) Lipman, E.; Bakajin, O.; Eaton, W. A. *Science* **2003**, *301*, 1233.
(8) Rhoades, E.; Gussakovsky, E.; Haran, G. *Proc. Natl. Acad. Sci.* **2003**, *18*, 3197.
(9) Bokinsky, G.; Rueda, D.; Misra, V. K.; Gordus, A.; Rhodes, M. M.; Babcock, H. P.; Walter, N. G.; Zhuang, X. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 9302.
(10) Xie, Z.; Srividya, N.; Sosnick, T. R.; Pan, T.; Scherer, N. F. P. *Natl. Acad. Sci. U.S.A.* **2004**, *101*, 334.
(11) Lippitz, M.; Kulzer, F.; Orrit, M. *ChemPhysChem* **2005**, *6*, 770.
(12) Chernyak, V.; Schulz, M.; Mukamel, S. *J. Chem. Phys*. **1999**, *111*, 7416.
(13) Agmon, N. *J. Phys. Chem.* **2000**, *104*, 7830.
(14) Cao, J. S. *Chem. Phys. Lett.* **2000**, *327*, 38.
(15) Gopich I. V.; Szabo, A. *J. Chem. Phys.* **2003**, *118*, 454.
(16) Barkai, E.; Jung, Y.; Silbey, R. J. *Phys. Rev. Lett.* **2001**, *87*, 207403.
(17) Flomenbom, O.; Klafter, J.; Szabo, A. *Biophys. J.* **2005**, *88*, 3780.
(18) Witkoskie, J. B.; Cao, J. S. *J. Chem. Phys*. **2004**, *121*, 6373.
(19) Yang, H.; Luo, G.; Karnchanaphanurach, P.; Louie, T. M.; Rech, I.; Cova, S.; Xun, L.; Xie, X. S. *Science* **2003**, *302*, 262.
(20) Skilling, J.; Bryan, R. K. *Mon. Not. R. Astron. Soc*. **1984**, *211*, 111.
(21) Steinbach, P. J.; Chu, K.; Frauenfelder, H.; Johnson, J. B.; Lamb, D. C.; Nienhaus, G. U.; Sauke, T. B.; Young, R. D. *Biophys. J.* **1992**, *61*, 235.
(22) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Dermirel, M. C.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505.
(23) Kuwata, K.; Kamatari, Y. O.; Akasaka, K.; James, T. L. *Biochemistry* **2004**, *43*, 4439.
(24) Doruker, P.; Atligan, A. R.; Bahar, I. *Proteins: Struct., Funct., Genet*. **2000**, *40*, 512.
(25) Garcia, A. E.; Hummer, G. *Proteins. Struct., Funct., Genet.* **1999**, *36*, 175
(26) Andrec, M.; Levy, R. M.; Talaga, D. S. *J. Phys. Chem. A* **2003**, *107*, 7454.
(27) Kou, S. C.; Xie, X. S. *Phys. Rev. Lett.* **2004**, *93*, 180603.
(28) Ingelman, M.; Ramaswamy, S.; Niviére, V.; Fontecave, M.; Eklund, H. *Proc. R. Soc. London, Ser. A* **1999a**, *455*, 3425.
(29) McCammon, J. A. *Rep. Prog. Phys.* **1984**, *47*, 1.
(30) Doniach, S.; Eastman, P. *Curr. Opin. Struct. Biol.* **1999**, *9*, 157.
(31) Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G. *Science* **1991**, *254*, 1598.
(32) Agmon, N.; Hopfield, J. J. *J. Chem. Phys.* **1983**, *79*, 2042.
(33) Monthus, C.; Bouchaud, J. P. *J. Phys. A: Math. Gen.* **1996**, *29*, 3847.
(34) Witkoskie, J. B.; Cao, J. S. *J. Chem. Phys.* **2006**, *125*, 244511.
(35) Karplus, M. *J. Phys. Chem. B* **2000**, *104*, 11.
(36) Bahar, I.; Wallquist, A.; Covell, D. G.; Jernigan, R. L. *Biochemistry* **1998**, *37*, 1067.
(37) Roitberg, A. E.; Gerber, R. B.; Ratner, M. A. *J. Phys. Chem. B* **1997**, *101*, 1700.
(38) Shen, M. Y.; Freed, K. F. *J. Chem. Phys*. **2003**, *118*, 5143.
(39) Politis, D. N. *IEEE Trans. Image Process.* **1994**, *4*, 865.
(40) Shore, J. E.; Zwanzig, R. *J. Chem. Phys.* **1975**, *63*, 5445.
(41) Witkoskie, J. B.; Wu, J.; Cao, J. S. *J. Chem. Phys.* **2004**, *120*, 5696.
(42) Granek, R.; Klafter, J. *Phys. Rev. Lett.* **2005**, *95*, 098106.
(43) Ingelman, M.; Ramaswamy, S.; Niviére, V.; Fontecave, M.; Eklund, H. *Biochemistry* **1999b**, *28*, 7040.
(44) Bryngelson, J. D.; Thirumalai, D. *Phys. Rev. Lett.* **1996**, *76*, 542.
(45) Venkataramanan, L.; Sigworth, F. J. *Biophys. J.* **2002**, *82*, 1930.