Science in China Series A: Mathematics
Jun., 2009, Vol. 52, No. 6, 1181–1211
www.scichina.com math.scichina.com
www.springerlink.com

A selective view of stochastic inference and modeling problems in nanoscale biophysics

KOU S. C.

Department of Statistics, Harvard University, Cambridge, MA 02138, USA
(email: kou@stat.harvard.edu)

Abstract Advances in nanotechnology enable scientists for the first time to study biological processes on a nanoscale molecule-by-molecule basis. They also raise challenges and opportunities for statisticians and applied probabilists. To exemplify the stochastic inference and modeling problems in the field, this paper discusses a few selected cases, ranging from likelihood inference, Bayesian data augmentation, and semi- and non-parametric inference of nanometric biochemical systems to the utilization of stochastic integro-differential equations and stochastic networks to model single-molecule biophysical processes. We discuss the statistical and probabilistic issues as well as the biophysical motivation and physical meaning behind the problems, emphasizing the analysis and modeling of real experimental data.

Keywords: likelihood analysis, Bayesian data augmentation, semi- and non-parametric inference, single-molecule experiment, subdiffusion, generalized Langevin equation, fractional Brownian motion, stochastic network, enzymatic reaction

MSC(2000): 60G35, 62F15, 62G05, 62P10, 92C05, 92C45

1 Introduction

The renowned physicist Richard Feynman once said that "everything that living things can do can be understood in terms of the jigglings and wigglings of atoms"[1]. Advances in nanotechnology of the last two decades have brought scientists closer to this "holy grail" than ever before. For the first time scientists were able to study biological processes on an unprecedented nanoscale molecule-by-molecule basis[2−7], opening the door to addressing many problems that were inaccessible just a few decades ago.

The new field of nanoscale single-molecule biophysics has attracted much attention from biologists, chemists and biophysicists because nanoscale *single-molecule* experiments offer many advantages over the traditional experiments involving a *population* of molecules. First, by "zooming in" on individual molecules, single-molecule experiments provide data with more accuracy and higher resolution. Second, by isolating, tracking and manipulating individual molecules, single-molecule experiments capture transient intermediates and detailed dynamics of a biological process, the type of information rarely available from traditional population experiments. Third, by following single molecules, scientists can study biological processes

Received December 1, 2008; accepted January 14, 2009
DOI: 10.1007/s11425-009-0074-y
This work was supported by the United States National Science Fundation Career Award (Grant No. DMS-0449204)

Citation: Kou S C. A selective view of stochastic inference and modeling problems in nanoscale biophysics. Sci China Ser A, 2009, 52(6): 1181–1211, DOI: 10.1007/s11425-009-0074-y

directly on the individual molecule level, instead of relying on the extremely difficult task of synchronizing the actions of a population of biomolecules. Fourth, since many important biological functions in a living cell are carried out by single molecules, understanding the behavior of individual biomolecules is a crucial task, for which single-molecule experiments are specifically designed. Many new scientific discoveries (see, for example, [8–10]) have emerged from the nanoscale single-molecule studies.

Advances in nanoscale single-molecule biophysics also bring opportunities and challenges for statisticians and stochastic modelers because of the stochastic nature of single-molecule experiments. First, on the single-molecule level, the laws of statistical and quantum mechanics fundamentally dictate the underlying biological dynamics/processes to be stochastic; their characterization thus requires stochastic models. Second, since the experiments focus on and study only one molecule at a time, the data from single-molecule experiments tend to be much noisier than those from the traditional population experiments because one cannot use the actions of thousands of molecules to average out the noise. Third, in most biophysical experiments, single-molecule experiments in particular, inference of the underlying stochastic dynamics is usually complicated by the presence of latent processes, which are unobserved but affect the data collection. Fourth, in addition to the preference of analytical tractability, it is important that the stochastic models constructed for biophysical processes should agree with fundamental physical laws and have a sound physical foundation.

In this article, to exemplify the stochastic inference and modeling problems in the field, we will look at a few selected cases, ranging from likelihood inference (of single-molecule fluorescence experiments), Bayesian data augmentation (to handle latent processes), and semi- and non-parametric inference (of nanometric biochemical systems) to the utilization of stochastic integro-differential equations (to model single-protein conformational fluctuation) and stochastic networks (to model single-enzyme reaction dynamics).

The paper is organized as follows. Section 2 considers the likelihood and Bayesian analysis of single-molecule experimental data, outlining, in particular, a Bayesian data augmentation method to handle latent processes. Sections 3 and 4 discuss non- and semi-parametric inference of nanoscale data. The second half of the paper (Sections 5 and 6) focuses on stochastic modeling problems. Section 5 considers the modeling of subdiffusive motion within single proteins, formulating a stochastic integro-differential equation framework. Section 6 studies enzymatic reactions of single proteins, proposing a stochastic network model to describe the reaction kinetics. Section 7 makes a few concluding remarks.

2 Likelihood and Bayesian inference

2.1 Likelihood analysis of single-molecule fluorescence experiments

One of the most important experimental tools to probe biological systems is fluorescence imaging, where biologists and chemists use stochastic signals extracted from the experimental videos to infer the function and mechanism of proteins and enzymes alike. In the context of single-molecule biophysics, fluorescence experiments[11] play an indispensable role. In such experiments, the system of interest is placed in a focal volume under a laser beam. The laser excites the molecule under study, which then emits photons. These photons are recorded by a pho-

ton detector. Because the molecule's photon emission pattern depends on its underlying state (e.g., the active and inactive states of an enzyme could have different photon emission intensity), by tracking how the photon emission pattern fluctuates over time, one can investigate the underlying (stochastic) dynamics itself.

A typical biological process/dynamics involves two entities A and B:

$$A \underset{k_{21}}{\overset{k_{12}}{\rightleftarrows}} B, \tag{2.1}$$

where A and B could be two different states of a protein or DNA molecule, k_{12} and k_{21} are the transition rates between the two states[12]. In the familiar statistics language, (2.1) translates to a two-state continuous-time Markov chain with the infinitesimal generator $Q = \left(\begin{smallmatrix} -k_{12} & k_{12} \\ k_{21} & -k_{21} \end{smallmatrix}\right)$. The corresponding transition matrix is

$$P(t) = e^{Qt} = \begin{pmatrix} \pi_1 + \pi_2 e^{-kt} & \pi_2(1 - e^{-kt}) \\ \pi_1(1 - e^{-kt}) & \pi_2 + \pi_1 e^{-kt} \end{pmatrix}, \tag{2.2}$$

where $k = k_{12} + k_{21}$, and $(\pi_1, \pi_2) = (k_{21}/(k_{12} + k_{21}), k_{12}/(k_{12} + k_{21}))$ denotes the stationary distribution of the two-state Markov chain.

In most experiments, however, the states A and B are not directly observed; they have to be inferred from the photon observations. The photon emission of the molecule follows a doubly stochastic Poisson process: During the period that the molecule is in state A, the photons are emitted (and subsequently arrive at the detector) according to a Poisson process with arrival rate λ_A; during the period that the molecule is in state B, the photon arrival is also Poisson but with a different arrival rate λ_B. The term "doubly stochastic" comes from the fact that both the arrival time and the arrival rate are stochastic. Let $\gamma(t)$ denote the two-state Markov process (2.1), taking values λ_A and λ_B. Define $Y(t)$ to be the total number of photons arrived at the detector up to time t. Then the dependence of $Y(t)$ on $\gamma(t)$ can be written as

$$\begin{aligned} &P(Y(t+h) - Y(t) = 1 | \gamma(t)) = \gamma(t)\, h + o(h), \\ &P(Y(t+h) - Y(t) = 0 | \gamma(t)) = 1 - \gamma(t)\, h + o(h). \end{aligned} \tag{2.3}$$

In the fluorescence experiments successive photon arrival times $T_1, T_2, \ldots, T_n$ are recorded. The goal is to infer from them the transition dynamics between A and B[10, 13]. The parameters are $\boldsymbol{\theta} = (k_{12}, k_{21}, \lambda_A, \lambda_B)$, of which the transition rates k_{12} and k_{21} are of special importance because (i) they are directly linked with the energy barrier height between A and B[14], and (ii) the energy barrier height in turn marks the energy landscape and the time scale of the reaction.

Using an infinitesimal discretization approach formulated in [15], the exact likelihood function can be obtained in closed form. First, the time interval $[T_1, T_n]$ can be divided into infinitesimal pieces, each with length h. Second, within each small interval, one can approximate the conditional likelihood of $T_1, T_2, \ldots, T_n$ given γ by successive Bernoulli trials (2.3). Finally, combining this approximation with the transition probability (2.2) of γ, applying matrix algebra and taking the limit of $h \to 0$ give the *exact* likelihood $L(T_1, \ldots, T_n | \boldsymbol{\theta}) = (\pi_1, \pi_2)\Lambda(\prod_{i=1}^{n-1} e^{(Q-\Lambda)(T_{i+1}-T_i)}\Lambda)\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$, where the matrix $\Lambda = \mathrm{diag}(\lambda_A, \lambda_B)$.

Likelihood for single-molecule fluorescence lifetime experiments. A special subclass of fluorescence experiments is the so-called single-molecule fluorescence lifetime experiments, in which, *in addition to* the photon arrival time T_i, the detector registers another quantity termed fluorescence lifetime for each photon. The *fluorescence lifetime* (measuring the molecule's response time to laser excitation) has an exponential distribution with rate depending also on the underlying state of the molecule. For the two-state process (2.1), the fluorescence lifetime has an exponential distribution with rate a when the molecule is in state A, and an exponential distribution with a different rate b when the molecule is in state B. An interesting fact from the quantum theory is that at any molecular state the mean of the fluorescence lifetime and the photon's Poisson arrival rate are proportional to each other. We thus have $\lambda_A = C/a$, and $\lambda_B = C/b$ for some proportional constant C.

Let τ_i denote the fluorescence lifetime associated with the i-th photon. The data pairs $\{(T_i, \tau_i)\}_{i=1}^n$ are collected in fluorescence lifetime experiments. They depend on the underlying γ process via

$$\begin{aligned}
&P(Y(t+h) - Y(t) = 1|\gamma(t)) = \gamma(t)\, h + o(h),\\
&P(Y(t+h) - Y(t) = 0|\gamma(t)) = 1 - \gamma(t)\, h + o(h),\\
&[\tau|\gamma(t), Y(t+h) - Y(t) = 1] \sim C\,\gamma^{-1}(t) \cdot \exp(-C\,\gamma^{-1}(t)\tau).
\end{aligned}$$

The parameter of interest is $\boldsymbol{\theta} = (k_{12}, k_{21}, a, b, C)$. Fluorescence *lifetime* experiments usually require more sophisticated equipment than ordinary fluorescence experiments, but the gain is that both T_i and τ_i provide information about the underlying stochastic dynamics of the molecule. Using the same infinitesimal discretization method, one can derive the closed-form likelihood for $\{(T_i, \tau_i)\}_{i=1}^n$.

2.2 Bayesian data augmentation for latent processes

The closed-form likelihood in principle allows efficient inference of the parameters. In real experiments, however, there are usually latent processes that complicate the inference. One substantial complication arises from the molecule's diffusion[11]. In the experiments, as the laser-excited molecule emits photons, it also diffuses in the laser's focal volume. As a result, since the laser illuminating intensity that the molecule receives is the highest at the center of the focal volume and decreases from center outward, the actual photon arrival rate depends not only on the underlying state of the molecule, but also on the molecule's location. The photon arrival rate can be expressed as $\gamma(t)\alpha(t)$ with

$$\alpha(t) = \exp\left[-\frac{B_x^2(t) + B_y^2(t)}{w_{xy}^2} - \frac{B_z^2(t)}{w_z^2} \right], \tag{2.4}$$

where $(B_x(t), B_y(t), B_z(t))$ is the location of the molecule, following a three-dimensional Brownian motion, and the known constants w_{xy} and w_z specify the x-, y- and z-axes of the ellipsoidal focal volume. The presence of molecular diffusion changes the statistical structure from (2.3) to

$$\begin{aligned}
&P(Y(t+h) - Y(t) = 1|\gamma(t), \alpha(t)) = \gamma(t)\alpha(t)h + o(h),\\
&P(Y(t+h) - Y(t) = 0|\gamma(t), \alpha(t)) = 1 - \gamma(t)\alpha(t)h + o(h).
\end{aligned}$$

The extra conditionality on the diffusion factor $\alpha(t)$ alters the likelihood as well. Using the infinitesimal discretization technique, we can derive the *conditional* likelihood of $T_1, \ldots, T_n$ conditioning on $\alpha(t)$ as $L(T_1, \ldots, T_n|\boldsymbol{\theta}, \alpha_t) = (\pi_1, \pi_2)\Lambda_1(\prod_{i=1}^{n-1} e^{(Q-\Lambda_i)(T_{i+1}-T_i)}\Lambda_{i+1})\binom{1}{1}$, where $\Lambda_i = \mathrm{diag}(\lambda_A\alpha(T_i), \lambda_B\alpha(T_i))$. Since the molecule's 3-D Brownian diffusion is not observed, in the full likelihood the latent diffusion factor $\alpha(t)$ has to be integrated out: $L(T_1, \ldots, T_n|\boldsymbol{\theta}) = \int L(T_1, \ldots, T_n|\boldsymbol{\theta}, \alpha_t)P(\alpha_t)\, d\alpha_t$, where $P(\alpha_t)$ denotes the probability law of $\alpha(t)$. This path integral is analytically intractable. The difficulty carries over to the Bayesian inference: with prior distribution $\eta(\boldsymbol{\theta})$, evaluating the posterior distribution

$$P(\boldsymbol{\theta}|T_1, \ldots, T_n) \propto \eta(\boldsymbol{\theta}) \int L(T_1, \ldots, T_n|\boldsymbol{\theta}, \alpha_t)P(\alpha_t)\, d\alpha_t \tag{2.5}$$

is very challenging even numerically.

Bayesian data augmentation. One method to address this difficulty is Bayesian data augmentation[16]. The intuitive idea is that if we can augment the unobserved diffusion (B_x, B_y, B_z), the inference and computation will be much easier. Statistically speaking, instead of focusing on the marginal distribution (2.5) of $\boldsymbol{\theta}$, we consider the joint posterior distribution of $\boldsymbol{\theta}$ and (B_x, B_y, B_z),

$$P(\boldsymbol{\theta}, B_x, B_y, B_z|T_1, \ldots, T_n) \propto \eta(\boldsymbol{\theta})L(T_1, \ldots, T_n|\boldsymbol{\theta}, \alpha_t)P(B_x, B_y, B_z),$$

where $P(B_x, B_y, B_z)$ is the law of (B_x, B_y, B_z). The difficult path integral disappears from this joint distribution. By sampling from it, one effectively marginalizes out the hidden diffusion process and obtains the correct inference.

To draw the samples, one can start with the Gibbs sampler. However, since the 3-D diffusion (B_x, B_y, B_z) has to be updated time point by time point, which is lengthy, a dynamic programming idea of forward-backward sampling[17, 18] can be applied to reduce the computation cost: backward compute the partial matrix products in the likelihood, forward update the diffusion chain. The detailed implementation of the forward-backward sampling is given in [15]. The computation cost is reduced by an order of magnitude: from $O(n^2)$ to $O(n)$.

For fluorescence lifetime experiments (where, recall, one also observes the τ_i's) the data augmentation procedure for inferring the parameters can be carried out in the same way except with conditional likelihood $L(T_1, \tau_1, \ldots, T_n, \tau_n|\alpha(t), \boldsymbol{\theta}) = (\pi_1, \pi_2)D_1\Lambda_1(\prod_{i=1}^{n-1} e^{(Q-\Lambda_i)(T_{i+1}-T_i)} D_{i+1}\Lambda_{i+1})\binom{1}{1}$, where $D_i = \mathrm{diag}(a\exp(-a\tau_i), b\exp(-b\tau_i))$, and $\Lambda_i = \mathrm{diag}(\lambda_A\alpha(T_i), \lambda_B\alpha(T_i))$.

A simulation example. We use simulated fluorescence lifetime data sets from [15] to illustrate the data augmentation approach. Each data set contains pairs $\{(T_i, \tau_i)\}_{i=1}^n$ generated from the latent Brownian diffusion (B_x, B_y, B_z) and the two-state process γ. Applying the data augmentation approach, we imputed Brownian diffusion for each data set. Figure 1(a) shows the sample posterior distribution (with a flat prior) of the parameters from a typical Monte Carlo run (the vertical bars are the true values). The method is seen to correctly identify all the parameters. Figure 1(b) compares the augmented Brownian diffusion factor $\alpha(t)$ with the actual one, displaying the multiple augmented $\alpha(t)$ (the light curves) with the true $\alpha(t)$ (the thick curve) for four representative data sets. The data augmentation technique appears to recover the hidden factor well.

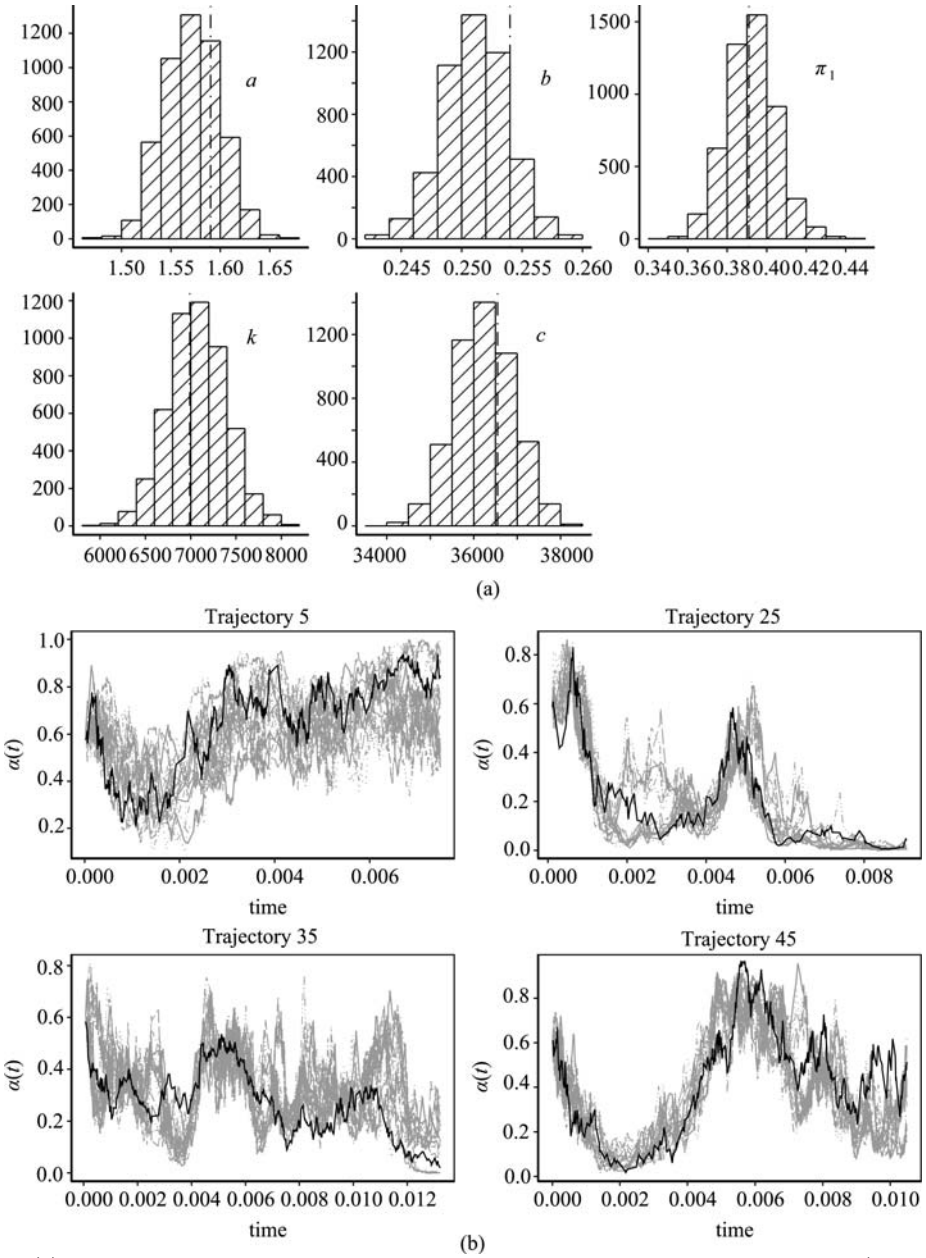

Figure 1 (a) Histograms of the posterior samples. Vertical bar in each panel is the true value ($k = k_{12} + k_{21}$, $\pi_1 = k_{21}/k$). (b) Comparison of the augmented Brownian factors $\alpha(t)$ with the actual one. The thick curve is the actual Brownian factor. The light curves are the augmented ones.

A real experimental data set. Let us now consider single-molecule experiments carried by the Xie group at Chemistry Department of Harvard University to study a DNA hairpin molecule, which is a single stranded nucleic acid structure, participating in many important biological processes including the regulation of gene expression[19], DNA recombination[20], and the facilitation of mutagenic events[21]. Because of the biological relevance, studying DNA hairpin's conformational properties, such as the conformational fluctuation and energy barrier between the different states, is of current interest. A DNA hairpin has two states, open and closed, illustrated in Figure 2. In the experiments, a fluorescence donor dye and a quencher

are attached to the two ends of the molecule (Figure 2). The donor dye emits photons under laser excitation, while the quencher annihilates the excitation. In the hairpin's closed state, the quenching is strong and very few photons from the donor are detected; in the open state, because the donor and quencher are far away from each other, the quenching is weak and many photons from the donor are detected.

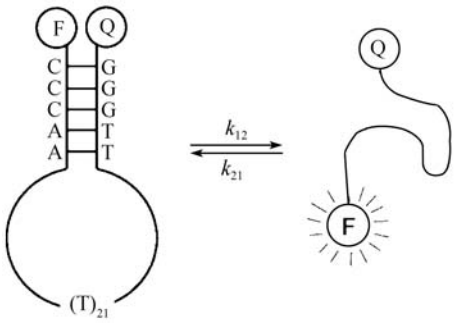

Figure 2 The closed and open states of a DNA hairpin. To infer the states, a fluorescence donor (F) and a quencher (Q) are attached to the two ends of the DNA hairpin.

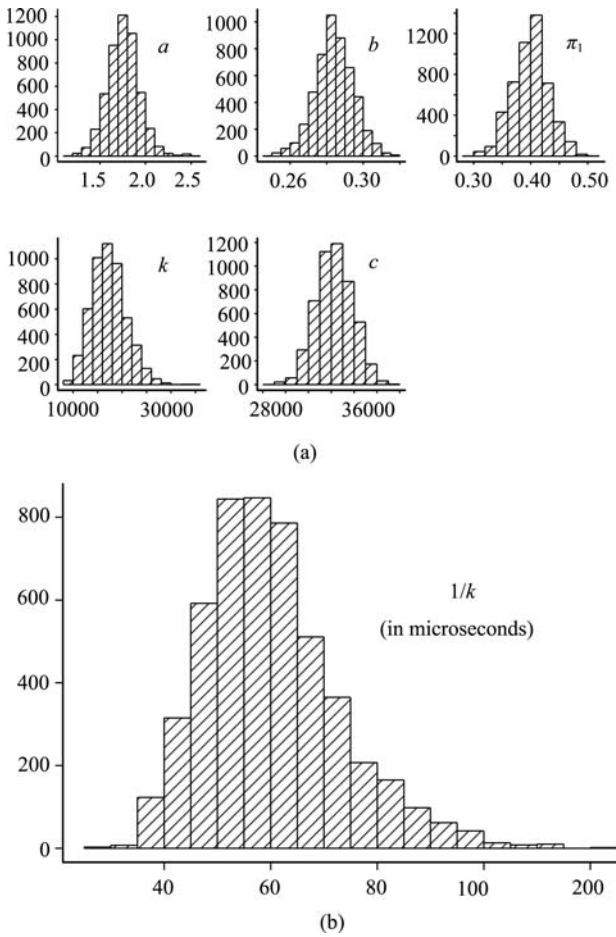

Figure 3 (a) Posterior histograms from the experimental data ($k = k_{12} + k_{21}$, $\pi_1 = k_{21}/k$). (b) The decay time constant $1/k$ for the experimental data.

Applying the data augmentation method to the fluorescence lifetime data yields the posterior distribution of the five parameters, shown in Figure 3(a). Figure 3(b) shows the posterior distribution of $1/k = 1/(k_{12} + k_{21})$, which, termed the *decay-time constant*, tells the energy barrier height between the two states of a DNA hairpin. Since our method uses the likelihood, it is more efficient than the available method-of-moment type estimation methods in the chemistry literature. In those approaches photon arrival times have to be first "binned" together to smooth out the effect of unobserved Brownian diffusion, and then the binned arrival times are used to fit certain moment equations to estimate the parameters[22, 23]. Because what happens inside the binning time-window is lost once the arrival times are binned together, the binning approaches suffer a significant loss of accuracy (i.e., time resolution). For the same data we analyze here, the binning methods have a maximum time resolution of 280 microseconds (μs). In contrast, our analysis provides a much sharper inference: the posterior median of $1/k$ is 59μs; a 95% symmetric posterior interval is (39, 91) μs.

2.3 Inference of complex models

Our discussion so far focuses on the two-state model (2.1). For some systems, scientists have proposed more complex models. For example, a two-by-two model

$$\begin{array}{ccc} A_1 & \overset{k_{12}}{\underset{k_{21}}{\rightleftarrows}} & B_1 \\ \alpha \downarrow\uparrow \alpha' & & \beta \downarrow\uparrow \beta' \\ A_2 & \overset{k'_{12}}{\underset{k'_{21}}{\rightleftarrows}} & B_2 \end{array} \tag{2.6}$$

has been used to describe the dynamics of some proteins[24]. The underlying $\gamma(t)$ process is now a four-state continuous-time Markov chain, taking values λ_{A_1}, λ_{A_2}, λ_{B_1}, and λ_{B_2} respectively in the four states. To infer this type of model, the likelihood analysis and Bayesian data augmentation can be straightforwardly generalized.

The two-by-two model can be further generalized to a continuous diffusive model

$$A \overset{k_{12}e^{-x(t)}}{\underset{k_{21}e^{-x(t)}}{\rightleftarrows}} B, \tag{2.7}$$

where an Ornstein-Uhlenbeck process $x(t)$ satisfying $dx(t) = -\rho x(t)dt + \sqrt{2\xi\rho}\, dW(t)$ is used to control the transition rates[25]. Physically, the time-varying and stochastic transition rates can be pictured as the result of a dynamically fluctuating energy barrier between the two states. Inference of this model is more challenging. First, the presence of the control process $x(t)$ implies that the likelihood must undergo another layer of conditioning: conditioning on $x(t)$. Second, the control process $x(t)$ is not observed and cannot be analytically integrated out. Third, the parameters ρ and ξ of the $x(t)$ are unknown and they couple strongly with $x(t)$. The data augmentation idea can be applied to address the first two difficulties: augmenting both $x(t)$ and the Brownian diffusion. The strong correlation between (ρ, ξ) and $x(t)$, however, creates a complication: Conditioning on large values of $x(t)$, posterior draws of (ρ, ξ) tend to be large, and, conversely, large values of (ρ, ϕ) tend to induce large posterior draws of $x(t)$.

Thus, to draw posterior samples, in addition to the forward-backward sampling of Subsection 2.2, more powerful Monte Carlo methods have to be used. Kou, Xie and Liu[15] introduced a group-move Monte Carlo, which works to move the highly correlated (ρ, ξ) and $x(t)$ together.

By breaking the high-correlation bottleneck that seriously limits the Gibbs sampler, the group-move Monte Carlo substantially improves the sampling efficiency.

Since different models convey different physical/biological pictures of the underlying systems, a natural scientific question is to discriminate the competing models from the experimental data. For instance, for the DNA hairpin molecule (of Figure 2), there were many debates about whether the continuous diffusive model is more appropriate than the two-state model[26−29]. However, because of the difficulty of data analysis, no clear consensus had been previously reached. The likelihood-based data augmentation approach provides a means to address it. First, since the two-state model can be viewed as a degenerate case of the diffusive model with $\sqrt{\xi} = 0$, one can examine the posterior distribution of $\sqrt{\xi}$ from the diffusive model: if it is sufficiently far from 0, then the data is supportive of the diffusive model. Second, the data augmentation method can be used together with the following result from [15] to compute the Bayes factor[30] for Bayesian model selection.

For two nested models $M_1 \subset M_2$, where M_1 has parameters $\boldsymbol{\mu}$, and the larger model M_2 has parameters $(\boldsymbol{\mu}, \boldsymbol{\zeta})$, if the prior distributions are consistent, i.e., $P(\boldsymbol{\mu}|M_1) = \int P(\boldsymbol{\mu}, \boldsymbol{\zeta}|M_2)d\boldsymbol{\zeta}$, then the Bayes factor can be expressed as the posterior mean of the likelihood ratio:

$$BF = \frac{P(\mathbf{y}|M_1)}{P(\mathbf{y}|M_2)} = E\left[\frac{P(\mathbf{y}|M_1, \boldsymbol{\mu})}{P(\mathbf{y}|M_2, \boldsymbol{\mu}, \boldsymbol{\zeta})} \;\middle|\; \mathbf{y}, M_2\right]. \tag{2.8}$$

Here $\mathbf{y}$ refers to the data. This identity implies that if we have posterior samples of the parameters $(\boldsymbol{\mu}^{(i)}, \boldsymbol{\zeta}^{(i)})$, $i = 1, \ldots, N$, drawn from the larger model M_2, we can then estimate the Bayes factor by the sample average of $\widehat{BF} = \frac{1}{N}\sum_{i=1}^{N}[P(\mathbf{y} \mid M_1, \boldsymbol{\mu}^{(i)})/P(\mathbf{y} \mid M_2, \boldsymbol{\mu}^{(i)}, \boldsymbol{\zeta}^{(i)})]$.

Results for the DNA hairpin data. Applying the group-move Monte Carlo method with data augmentation to the diffusive model, we obtained the posterior distribution of the parameters from the DNA hairpin data, shown in Figure 4. The posterior samples of $\sqrt{\xi}$

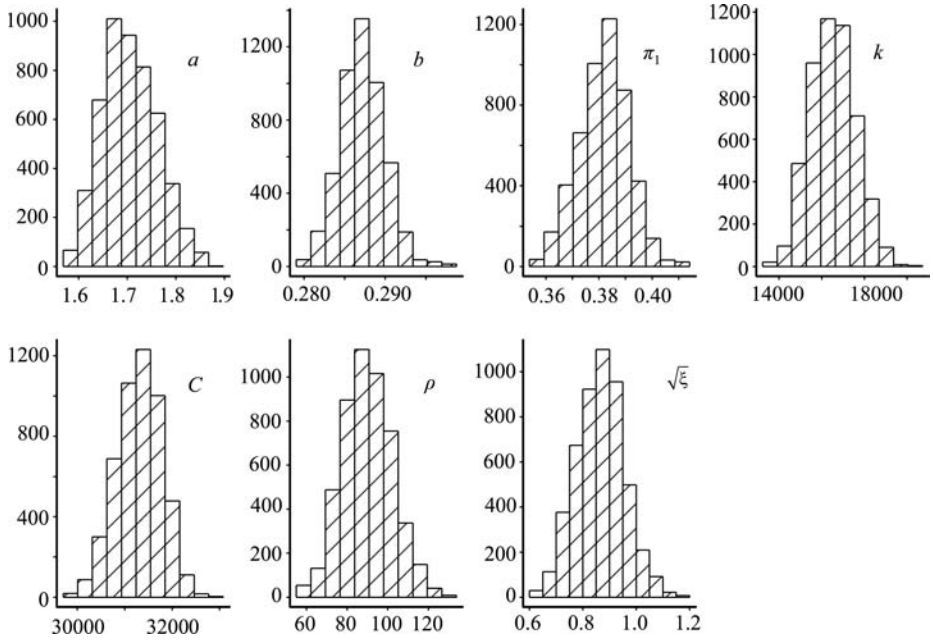

Figure 4 Posterior histograms from the diffusive model.

are concentrated far away from 0, which indicates strongly that the two-state model does not fit the data. The estimated Bayes factor of $\widehat{BF} = (3.43 \pm 0.29) \times 10^{-9}$ corroborates the graphical message of Figure 4. For the DNA hairpin data, $\boldsymbol{\mu}$ and $\boldsymbol{\zeta}$ in (2.8) correspond to $\boldsymbol{\mu} = (\boldsymbol{\theta}, B_x, B_y, B_z)$ and $\boldsymbol{\zeta} = (\rho, \xi, x(t))$. From a scientific point of view, the preference of the diffusive model implies that the energy barrier between the two states of the DNA hairpin has more complex behavior than the simple static picture depicted in the two-state model. The fluctuation of the energy barrier in this case could be attributed to conformational flexibility of the DNA molecule.

3 Nonparametric inference of single-molecule fluorescence experiments

The preceding discussion illustrates that, under the specification of parametric models, Bayesian and likelihood methods are quite effective for inferring the underlying biological dynamics from fluorescence photon arrival data, and for discriminating between competing models even when latent processes are present. Parametric models, however, are not always available for data analysis, especially when scientists are in the early exploration of a new biological process. The intuitive idea of "learning" directly from the data makes nonparametric inference appealing here.

Let us recall in single-molecule fluorescence experiments photon arrival times $T_1, T_2, \ldots, T_n$ within an observational time window $[0, T]$ are recorded; they follow a doubly stochastic Poisson process where the stochastic arrival rate $\gamma(t)$ depends on the underlying biological process. In the nonparametric case, the Poisson setting (2.3) still holds, but $\gamma(t)$ no longer has a specific parametric form (such as a continuous-time Markov chain that we encountered previously). The goal is to infer the properties of $\gamma(t)$ nonparametrically.

One important characteristic is the autocorrelation function (ACF) of $\gamma(t)$, which measures the strength of dependence. A fast decay of the ACF, such as an exponential decay, indicates that the underlying biological process is Markovian and that the biomolecule under study has a relatively simple conformation dynamics, whereas a slow decay of ACF signifies a complicated process and points to an intricate internal structure/conformational dynamics of the biomolecule.

To estimate the ACF nonparametrically, we first approximate $\gamma(t)$ from the photon arrival data by a kernel estimate

$$\hat{\gamma}(t) = \sum_{i=1}^{n} \frac{1}{h} f\left(\frac{T_i - t}{h}\right), \tag{3.1}$$

where f is a symmetric smooth density function with bounded support $[-b, b]$, and h is the bandwidth. Under stationarity and ergodicity assumptions on $\gamma(t)$, the ACF $C(t) = \mathrm{Cov}(\gamma(0), \gamma(t))$ is estimated by

$$\hat{C}(t) = \frac{1}{T - 2bh - t} \int_{bh}^{T-bh-t} (\hat{\gamma}(t+s) - \hat{\mu})(\hat{\gamma}(s) - \hat{\mu}) ds, \tag{3.2}$$

where $\hat{\mu} = n/T$ estimates $\mu = E[\gamma(t)]$, the mean photon arrival intensity, and the integral is taken over the range of $[bh, T - bh - t]$ instead of $[0, T - t]$ to avoid the bias at the boundary of the observational window. Zhang and Kou[31] showed that under mind regularity conditions (such as stationarity and ergodicity) $\hat{C}(t)$ converges to $C(t)$ in L^2 as $T h \to \infty$, $h \to 0$.

The practical application of (3.1) and (3.2) requires the choice of the bandwidth h. It can be shown that the optimal one (in the mean square error sense) for $\hat{\gamma}$ is

$$h_{\text{opt}} = D_f \left[\frac{\mu}{C'(0^+)} \right]^{1/2}, \tag{3.3}$$

where the constant $D_f = \{\int_{-b}^{b} f^2(s)ds \,/(\int_{-b}^{b}\int_{-b}^{b} |s_1 - s_2| f(s_1)f(s_2)ds_1 ds_2 - 2\int_{-b}^{b} |s| f(s)ds)\}^{1/2}$. Based on this result, a plug-in regression method for choosing h was proposed in [31]. We can start with an initial h and use it to estimate $C(t)$ and $C'(0^+)$, then replace the unknown $C'(0^+)$ and μ in (3.3) with $\hat{C}'(0^+)$ and $\hat{\mu} = n/T$, which gives a better $\hat{h}$.

Under the assumption of short range dependence of $\gamma(t)$, we can further establish the asymptotic normality for $\hat{C}(t)$. Detailed (but lengthy) analysis[31] shows that the asymptotic variance can be well approximated by

$$\hat{V}(t) = \frac{2}{(T - 2bh - t)^2} \int_0^{T-2bh-t} (T - 2bh - t - s)\widehat{\text{Cov}}(t, s)ds, \tag{3.4}$$

where $\widehat{\text{Cov}}(t, s)$ involves the fourth central moment of γ,

$$\widehat{\text{Cov}}(t, s) = \max\left\{ \int_{bh}^{T-bh-s-t} \frac{\hat{\gamma}_c(r)\,\hat{\gamma}_c(r + t)\,\hat{\gamma}_c(r + s)\,\hat{\gamma}_c(r + s + t)}{T - 2bh - s - t} dr - \hat{C}^2(t), 0 \right\}$$

with $\hat{\gamma}_c(t) \equiv \hat{\gamma}(t) - \hat{\mu}$ denoting the centralized $\hat{\gamma}(t)$. Equation (3.4) provides a practical method to construct confidence intervals for the ACF. For example, an asymptotic 95% confidence interval is $\hat{C}(t) \pm 1.96\sqrt{\hat{V}(t)}$.

Two simulation examples. As a first illustration, we simulate the photon arrival times from the two-by-two model (2.6), where $\gamma(t)$ follows a four-state continuous-time Markov chain with transition rates $k_{12} = 3$, $k_{21} = 4$, $k'_{12} = 4$, $k'_{21} = 5$, $\alpha = 1$, $\alpha' = 2$, $\beta = 2.5$, and $\beta' = 4$, and $\gamma(t)$ takes values 100000, 90000, 5000, and 4500, respectively, at states A_1, A_2, B_1, and B_2. The observational time window is $[0, T] = [0, 100]$. The true ACF $C(t)$ in this case is a mixture of three exponential functions. Figure 5(a) shows the estimated $\hat{C}(t)$ and the 95% confidence interval based on one data set, compared with the true $C(t)$. The ACF is well recovered. As a further checkup for the accuracy of the confidence interval, we repeat the data generation 100 times (each time generating $\gamma(t)$ first and then $T_1, T_2, \ldots$). For each data set we calculate $\hat{C}(t)$. The 2.5 and 97.5 percentile of these repeated estimates gives the real 95% coverage of $\hat{C}(t)$, which is shown in Figure 5(b). Comparing the two panels, we can see that the approximate confidence interval based on just one realization is close to the true one.

We next consider an example where the arrival rate $\gamma(t)$ takes continuous values: $\gamma(t) = K \exp(G(t))$, where $G(t)$ is a stationary Gaussian process with mean 0 and covariance function $\varpi(t) = 1/(1+|t|)^6$. It is straightforward to show that the true ACF is $C(t) = K^2(\exp(\varpi(t)) - 1)$ here. We generate photon arrival times from this model with $K = 10000$. Figure 6(a) shows $\hat{C}(t)$ and the 95% confidence interval based on one data set, compared with the true $C(t)$. The ACF is well recovered. To further check up, we repeat the simulation 100 times. Figure 6(b) shows the 2.5 and 97.5 percentile from the repeated estimates $\hat{C}(t)$. It is evident that the approximate confidence interval (based on one realization alone) in Figure 6(a) is quite close to the true one in Figure 6(b).

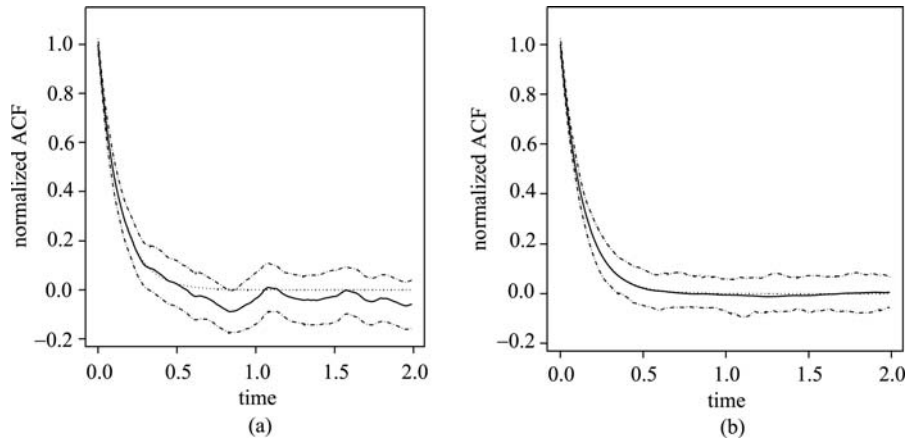

Figure 5 Estimated ACF for data generated from the two-by-two model. (a) $\hat{C}(t)$ and the 95% confidence interval based on one simulated data set. The dotted curve is the true $C(t)$. (b) The true 95% coverage of $\hat{C}(t)$ based on 100 repetitions. The solid line is the average over the 100 estimates.

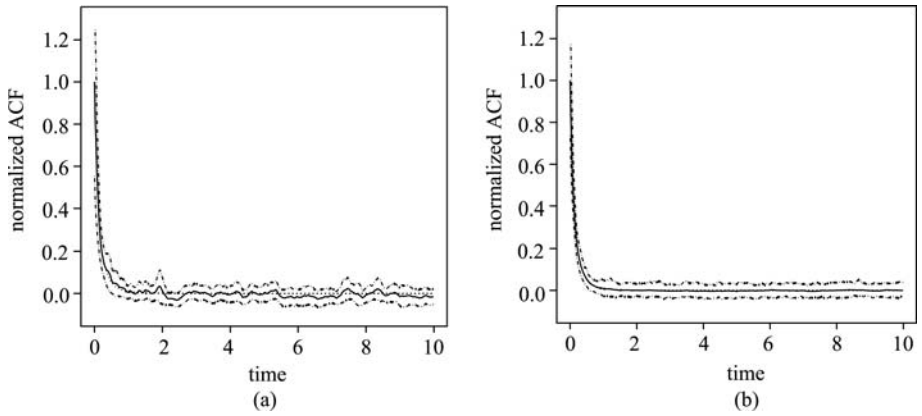

Figure 6 Estimated ACF for data generated from the exponential-Gaussian model. (a) $\hat{C}(t)$ and the 95% confidence interval based on one simulated data set. The dotted curve is the true $C(t)$. (b) The true 95% coverage of $\hat{C}(t)$ based on 100 repetitions. The solid line is the average over the 100 estimates.

Real experimental data. A recent single-molecule experiment[32] studied a protein complex formed by fluorescein and monoclonal antifluorescein. This is an antibody-antigen system. In the experiment, the 3-D conformation of the molecule spontaneously fluctuates over time. To study the conformational dynamics, the immobilized protein complex was placed under a laser beam. Photons from the laser-excited molecule are collected with the photon arrival rate depending on the molecule's time-varying conformation.

Applying the nonparametric estimator to the experimental data gives $\hat{C}(t)$ and the 95% confidence interval, shown in Figure 7. The graph was plotted on a log-log scale; the approximate (log-log) linear trend reveals the slow decay of $C(t)$ (more specifically, a power law type of decay), and suggests that the underlying $\gamma(t)$ has a long memory. The presence of long memory indicates the complexity of the protein complex' conformational fluctuation, which might be due to the molecule's intricate structure. For detailed discussion about the biological implications, such as its effect on enzyme activity, see [33].

4 Semi-parametric inference of nanometric biochemical systems

Studying the chemical kinetics behind biochemical reactions is of great importance to chemists and biologists because chemical kinetics often governs the reactions' biological functions. Understanding the detailed chemical kinetics of enzyme reactions, in particular, is of considerable current interest. A single enzyme molecule in a living cell is a nanometric system; it catalyzes biochemical reactions by first binding to the substrate (i.e., the reactant), then turning the substrate into reaction product, and finally coming back to its initial state. The corresponding kinetics can be depicted as

$$\begin{array}{ccc} & ES & \\ \nearrow\swarrow & & \\ E & \uparrow\downarrow & \\ \nwarrow\searrow & & \\ & EP & \end{array} \quad \text{with transition rates} \quad \begin{cases} E \underset{\kappa_{-1}}{\overset{\kappa_1}{\rightleftarrows}} ES \\ ES \underset{\kappa_{-2}}{\overset{\kappa_2}{\rightleftarrows}} EP \\ EP \underset{\kappa_{-3}}{\overset{\kappa_3}{\rightleftarrows}} E \end{cases}. \tag{4.1}$$

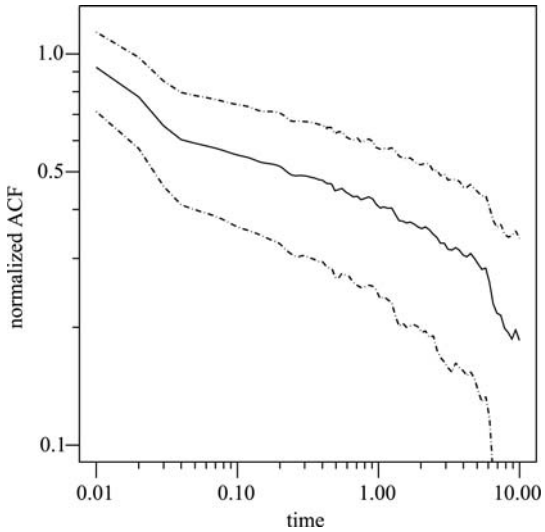

Figure 7 ACF for a real experimental data set. $\hat{C}(t)$ and the 95% confidence interval are plotted on a log-log scale. The approximate power law type of decay suggests a long memory of $\gamma(t)$, which indicates that the conformational fluctuation of the protein complex is very complicated.

When an enzyme completes a forward reaction cycle (i.e., clockwise in the above diagram), it successfully converts a substrate molecule to a product. However, since the reactions are reversible, once in a while an enzyme can cycle backward (i.e., counterclockwise), in which case it (counterproductively) turns a product back to a substrate. The tradeoff between the forward and backward reaction flow is characterized by the thermodynamic driving force $\triangle\mu$, defined as

$$\triangle\mu = \log\left(\frac{\kappa_1}{\kappa_{-1}}\right) + \log\left(\frac{\kappa_2}{\kappa_{-2}}\right) + \log\left(\frac{\kappa_3}{\kappa_{-3}}\right).$$

Since the thermodynamic driving force $\triangle\mu$, intrinsic to the (enzyme) system, directly measures the system's thermodynamic tendency towards its chemical equilibrium[34], accurate estimation of $\triangle\mu$ is central to the understanding of biochemical processes, in particular, the actual function of the reaction in live cells.

In single-molecule experiments, the turnover cycles of a single enzyme molecule are followed over time. In particular, the number of *net* cycles (the full forward cycles minus the full backward cycles) accomplished by the enzyme molecule within a time window can be determined. Let Z_i denote the number of net cycles in the i-th time window ($1 \leqslant i \leqslant n$); it can take both positive and negative values. The task is to estimate $\triangle\mu$ from the cycle data $Z_1, Z_2, \ldots, Z_n$.

To do so, one natural approach is to calculate the probability distribution of Z_i from the three-state continuous-time Markov chain (4.1), estimate the parameters κ_1, κ_{-1}, etc., and then obtain $\triangle\mu$ through its definition. This full parametric approach, however, has a serious limitation: it relies on the correctness of the three-state Markov chain model. In fact, for many enzyme reactions, scientists have suggested and sometimes deduced the existence of reaction intermediates, which means that there could be four or more states in the reaction cycle, such as $E \rightleftarrows ES \rightleftarrows ES' \rightleftarrows EP' \rightleftarrows EP \rightleftarrows E$. The thermodynamic driving force $\triangle\mu$ in these cases is still defined as the sum of log-ratios between the forward and backward rates (except there might be κ_4, κ_5, etc.). The question is how to estimate $\triangle\mu$ from $Z_1, \ldots, Z_n$ without assuming a specific model.

A very useful result for the estimation is the following relationship: for any positive integer j,

$$\frac{P(Z_i = j)}{P(Z_i = -j)} = \exp(j\,\triangle\mu). \tag{4.2}$$

It can be shown[35] that this relationship, which is termed fluctuation theorem, holds irrespective of the exact model. Therefore, it is very desirable to estimate $\triangle\mu$ based on this model-independent semi-parametric result. One method-of-moment approach used by biophysicists is based on a corollary of (4.2): $E[\exp(-\triangle\mu\, Z)] = \sum_{j=-\infty}^{\infty} P(Z = j)e^{-j\,\triangle\mu} = \sum_{j=-\infty}^{\infty} P(Z = j) = 1$, which leads to the estimator $\triangle\hat{\mu}_{\rm MM}$

$$\triangle\hat{\mu}_{\rm MM} = \text{the nonzero solution of} \quad \frac{1}{n}\sum_{i=1}^{n} \exp(-Z_i\,\triangle\hat{\mu}_{\rm MM}) = 1.$$

This method-of-moment estimator, though intuitively simple, does not utilize the full potential of (4.2). A more efficient semi-parametric maximum likelihood estimator was proposed by [36].

The probabilities $p_j = P(Z = j)$ are treated as nuisance parameters but with the important link of $p_{-j} = p_j \exp(-j\,\triangle\mu)$, $j \geqslant 1$. Let N_j denote the number of occurrences (out of the n observations) that exactly j net turnover cycles are observed, i.e., $N_j = \sharp\{i : Z_i = j\}$, $j = 0, \pm1, \pm2, \ldots$. Clearly, $\sum_{j=-\infty}^{\infty} N_j = n$. The likelihood of observing the outcome $\{N_0, N_{\pm1}, N_{\pm2}, \ldots\}$ is then given by

$$L(N_0, N_{\pm1}, \ldots \mid \triangle\mu, p_0, p_1, \ldots) = \frac{n!}{\prod_{j=-\infty}^{\infty} N_j!} \prod_{j=-\infty}^{\infty} p_j^{N_j} \propto p_0^{N_0} \prod_{j=1}^{\infty} [p_j^{N_j} (p_j e^{-j\,\triangle\mu})^{N_{-j}}].$$

The semi-parametric MLE $\triangle\hat{\mu}_{\rm MLE}$ is obtained by maximizing the log-likelihood

$$\log L(N_0, N_{\pm1}, \ldots \mid \triangle\mu, p_0, \ldots) = \text{const} + N_0 \log p_0 + \sum_{j=1}^{\infty} [(N_j + N_{-j}) \log p_j - \triangle\mu\, j N_{-j}],$$

subject to the constraint $p_0 + \sum_{j=1}^{\infty} p_j(1 + e^{-j\,\triangle\mu}) = 1$ (i.e., the total probability should be one).

Using the Lagrange multiplier and solving the corresponding first-order conditions, we find the final expression

$$\triangle\hat{\mu}_{\mathrm{MLE}} = \text{the solution of} \quad \sum_{j=1}^{\infty} j\frac{N_j \exp(-j\,\triangle\hat{\mu}_{\mathrm{MLE}}) - N_{-j}}{1+\exp(-j\,\triangle\hat{\mu}_{\mathrm{MLE}})} = 0, \tag{4.3}$$

where the nuisance parameters automatically dropped out. It is instructive to compare this new semi-parametric MLE with the method-of-moment approach. We can rewrite $\triangle\hat{\mu}_{\mathrm{MM}}$ as

$$\triangle\hat{\mu}_{\mathrm{MM}} = \text{the nonzero solution of} \quad \sum_{j=1}^{\infty}[N_j \exp(-j\,\triangle\hat{\mu}_{\mathrm{MM}}) - N_{-j}][\exp(j\,\triangle\hat{\mu}_{\mathrm{MM}}) - 1] = 0, \tag{4.4}$$

which tells us that $\triangle\hat{\mu}_{\mathrm{MLE}}$ and $\triangle\hat{\mu}_{\mathrm{MM}}$ differ essentially by their assignments of weights on the individual equations of $N_j \exp(-j\,\triangle\hat{\mu}) - N_{-j} = 0$. The (asymptotically) optimal weights of $j/(1+\exp(-j\,\triangle\hat{\mu}))$ are obtained only by working on the (semi-parametric) likelihood.

An illustration. Let us consider a simulation. We generated enzyme turnover data from the model (4.1). The parameters were taken to be $\kappa_1 = 430$, $\kappa_2 = \kappa_3 = \kappa_{-1} = \kappa_{-2} = 1000$, $\kappa_{-3} = 4.3$ (with unit sec^{-1}); the numbers were chosen so that their ranges are representative of real enzyme reactions. The number of net cycles Z_i within the time window of 0.01 sec were collected.

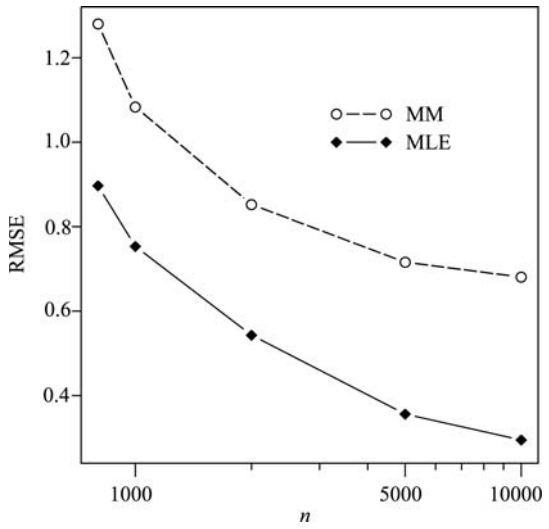

Figure 8 The root mean squared errors (RMSE) for $\triangle\hat{\mu}_{\mathrm{MM}}$ and $\triangle\hat{\mu}_{\mathrm{MLE}}$ as functions of the sample size n. The data generation and estimation are repeated 5000 times at each sample size.

The true $\triangle\mu$ is 4.6 in this case. To compare the two estimators, we examine the estimation with different sample sizes, ranging from several hundred to 10,000. The root mean squared error (RMSE), $[E(\triangle\hat{\mu} - \triangle\mu)^2]^{1/2}$, were calculated by repeating the data generation and estimation 5000 times at each sample size. Figure 8 plots the RMSE of $\triangle\hat{\mu}_{\mathrm{MLE}}$ and $\triangle\hat{\mu}_{\mathrm{MM}}$ as functions of n, the sample size. The superiority of $\triangle\hat{\mu}_{\mathrm{MLE}}$ over $\triangle\hat{\mu}_{\mathrm{MM}}$ is evident. For example, at $n = 10,000$, $\triangle\hat{\mu}_{\mathrm{MLE}}$ is more than five times as efficient as $\triangle\hat{\mu}_{\mathrm{MM}}$. An inspection of (4.4) reveals the root of $\triangle\hat{\mu}_{\mathrm{MM}}$'s problem. The method-of-moment estimator assigns exponential weight on the j-th equation. However, as j gets larger, having j or $-j$ net cycles becomes

rarer and rarer. Thus, by putting exponentially high weights on rare events, $\triangle\hat{\mu}_{\mathrm{MM}}$ loses its stability and, subsequently, its efficiency. This simulation example illustrates the efficacy of the semi-parametric approach. We expect its application to real single-molecule experimental data in the near future.

5 Modeling subdiffusion within proteins

We have so far discussed statistical inference problems in single-molecule biophysics. From this section on, we will turn to stochastic modeling problems in the field. We start from the modeling of subdiffusion within single proteins.

Since Einstein's and Wiener's ground breaking works in the early 20th century, the theory of Brownian motion and diffusion processes has revolutionized not only physics, chemistry and biology, but also probability and statistics. A key characteristic of Brownian motion is that the second moment $E[x^2(t)]$, which in physics corresponds to the mean squared displacement (location) of a Brownian particle, is proportional to time t. In some systems[37−39], scientists, however, have discovered a puzzling departure from Brownian diffusion: The mean squared displacement $E[x^2(t)]$ there is no longer proportional to t, but rather $E[x^2(t)] \propto t^{\alpha}$, where $0 < \alpha < 1$. Because $\alpha < 1$, these movements are defined as subdiffusion. Recent single-molecule experiments[9, 32] reveal that subdiffusion may be quite prevalent in biological systems.

In a 2003 *Science* paper[9], scientists conducting single-molecule experiments on a protein-enzyme system, called Fre, observed this subdiffusion phenomenon. The Fre system is involved in the DNA synthesis of *E. Coli*. (where Fre catalyzes a reaction involving the protein Flavin). Figure 9 shows the crystal structure of Fre, which contains two smaller structures: FAD (an electron carrier) and Tyr (an amino acid). The 3-D conformation of Fre spontaneously fluctuates, and consequently, the distance between the two substructures FAD and Tyr varies over time. It was found that the stochastic distance fluctuation between FAD and Tyr undergoes a subdiffusion.

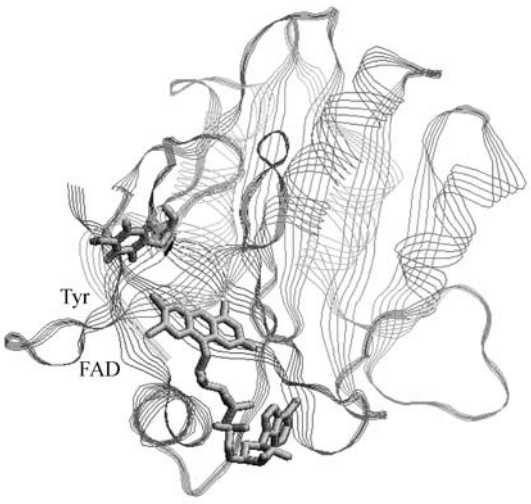

Figure 9 The crystal structure of Fre. The two substructures FAD and Tyr are highlighted.

To explain this subdiffusion phenomenon, Kou and Xie[40] formulated a stochastic integro-differential equation model based on fractional Gaussian noise and the generalized Langevin

equation.

Since the model utilizes concepts from statistical mechanics, to facilitate the discussion, let us first review how the law of Brownian diffusion was derived in physics. Suppose we have a Brownian particle with mass m suspended in water. The physical description of the particle's motion starts from the Langevin equation[41, 14]: $m\frac{dv(t)}{dt} = -\zeta v(t) + F(t)$, where $v(t)$ is the velocity of the particle at time t, and $dv(t)/dt$ is the acceleration of the particle. On the right hand side, ζ is the friction constant, and $F(t)$ is the white noise (formally the derivative of the Wiener process). Because both the movement of the particle and the friction originate from the particle's collision with water molecules, the Langevin equation has an important physical constraint $E[F(t)F(s)] = 2\zeta k_B T \cdot \delta(t-s)$, where k_B is the Boltzmann constant, T is the underlying temperature and $\delta(\cdot)$ is Dirac's delta function. This proportional relationship between the noise level and the friction constant is a consequence of the *fluctuation-dissipation* theorem in statistical mechanics[42, 43]. In the more familiar probability notation, the Langevin equation translates to $m\,dv(t) = -\zeta v(t)dt + \sqrt{2\zeta k_B T}\,dB(t)$, where $B(t)$ is the Wiener process, and the formal association of "$F(t) = \sqrt{2\zeta k_B T}\,dB(t)/dt$" is recognized.

The solution $v(t)$ to the Langevin equation is the Ornstein-Uhlenbeck process, which is stationary Gaussian with mean $E[v(t)] = 0$ and covariance function $E[v(t)v(s)] = \frac{k_B T}{m}\exp\left(-\frac{\zeta}{m}|t-s|\right)$. It follows that for the displacement, $x(t) = \int_0^t v(s)ds$, which is the actual observed motion, the second moment is

$$E[x^2(t)] = \mathrm{Var}[x(t)] \sim 2\frac{k_B T}{\zeta}t, \quad \text{for large } t, \tag{5.1}$$

which is known in physics as Einstein's Brownian diffusion law.

The classical theory of Brownian diffusion, however, fails to explain subdiffusion, which satisfies, instead, $E[x^2(t)] \propto t^\alpha$ with $0 < \alpha < 1$ for large t. The starting point of our model to account for subdiffusion is the *generalized Langevin equation* (GLE)[42, 44]

$$m\frac{dv(t)}{dt} = -\zeta\int_{-\infty}^{t} v(u)K(t-u)du + G(t), \tag{5.2}$$

where, in comparison with the Langevin equation, (i) a noise $G(t)$ having memory replaces the white noise, and (ii) the memory kernel K convoluted with the velocity makes the process non-Markovian. The reason that both K and $G(t)$ appear in the equation is that any closed (equilibrium) physical system must satisfy the fluctuation-dissipation theorem, which requires the memory kernel $K(t)$ and the noise to be linked by $E[G(t)G(s)] = k_B T\zeta \cdot K(t-s)$. In an intuitive sense, this relationship arises because both the friction and the motion of the particle originate from the collision between it and its surrounding media. The GLE reduces to the Langevin equation when K is the delta function.

Under the GLE framework, the key question is: Is there a combination of kernel function and noise structure that can lead to subdiffusion? To answer this question, we note that the white noise is mathematically interpreted as the formal derivative of a Wiener process, which is the unique process that satisfies (i) being Gaussian, (ii) having independent increment, (iii) having stationary increment, and (iv) being self-similar. To generalize the white noise, we want to maintain as many nice properties as possible and at the same time introduce memory. This

leads to processes with the following three property: (i) Gaussian, (ii) stationary increment and (iii) self-similar. The only class of process that embodies all three properties is the fractional Brownian motion (fBm) process $B_H(t)$[45, 46], which is Gaussian with mean $E[B_H(t)] = 0$, and covariance function $E[B_H(t)B_H(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H})$. H, between 0 and 1, is the so-called Hurst parameter; $B_H(t)$ reduces to the Wiener process when $H = 1/2$.

Taking $G(t)$ in (5.2) to be the (formal) derivative of fBm, $F_H(t) = \sqrt{2\zeta k_B T} dB_H(t)/dt$, we reach the model $m\frac{dv(t)}{dt} = -\zeta \int_{-\infty}^{t} v(u)K_H(t-u)du + F_H(t)$, where the kernel $K_H(t)$ is given by

$$K_H(t) = E[F_H(0)F_H(t)]/(k_B T\zeta) = 2H(2H-1)|t|^{2H-2}, \quad \text{for} \quad t \neq 0. \tag{5.3}$$

In the more familiar probability notation, the model can be written as

$$m\, dv(t) = -\zeta \left(\int_{-\infty}^{t} v(u)K_H(t-u)du \right) dt + \sqrt{2\zeta k_B T}\, dB_H(t). \tag{5.4}$$

The presence of the convolution term and the $dB_H(t)$ term makes this equation non-Markovian and nonstandard. It, nevertheless, can be solved in closed form via a Fourier analysis. The solution $v(t)$ is a stationary Gaussian process. See [47] for details. One can further show that the displacement $x(t) = \int_0^t v(s)ds$ of equation (5.4) satisfies

$$E[x(t)^2] = \mathrm{Var}[x(t)] \sim \frac{k_B T}{\zeta} \frac{2\sin(2H\pi)}{\pi H(2H-1)(2H-2)} t^{2-2H} \propto t^{2-2H}, \quad \text{for large } t.$$

This result tells us that the model with $H > 1/2$ leads to an explanation of subdiffusion.

Modeling subdiffusion under external potential. The model so far considers subdiffusion of a free particle, i.e., the motion of a particle without the influence of an outside force (or potential). If there exists an external potential $U(x)$ (e.g., a magnetic field), which is a function of the displacement $x(t)$, the model has to be modified. More specifically, the term $-U'(x(t))$ will be added to the right hand side of (5.2)[14, 42, 44]. Thus, to describe the movement of a subdiffusive particle under an external potential $U(x)$, the model becomes

$$dx(t) = v(t)dt, m\, dv(t) = -\zeta \left(\int_{-\infty}^{t} v(u)K_H(t-u)du \right) dt - U'(x(t))dt + \sqrt{2\zeta k_B T} dB_H(t). \tag{5.5}$$

In the special case of a harmonic potential $U(x) = \frac{1}{2}m\psi x^2$, where m is the mass of the particle and the constant ψ reflects the potential's strength, the model is

$$dx(t) = v(t)dt, m\, dv(t) = -\zeta \left(\int_{-\infty}^{t} v(u)K_H(t-u)du \right) dt - m\psi x(t)dt + \sqrt{2\zeta k_B T} dB_H(t), \tag{5.6}$$

which can also be solved by the Fourier transform method[47].

When the acceleration term $m\, dv(t)/dt$ is negligible, which corresponds to the so-called overdamped condition in physics[14], equation (5.6) reduces to

$$dx(t) = v(t)dt, \quad m\psi x(t)dt = -\zeta \left(\int_{-\infty}^{t} v(u)K_H(t-u)du \right) dt + \sqrt{2\zeta k_B T} dB_H(t), \tag{5.7}$$

which can be solved in closed form via the Fourier method[47].

Physical meaning of the model. A key requirement for biophysical models is that the model must make physical sense: It must agree with fundamental physical laws and should have a sound physical basis. For the motion of a free particle, the law of thermal dynamics[42, 43, 48] requires that the stationary variance of the velocity should be k_BT/m. It can be shown[47] that indeed our model (5.4) satisfies $E[v^2(0)] = \mathrm{Var}[v(0)] = k_BT/m$. For particles moving under a harmonic potential $U(x) = \frac{1}{2}m\psi x^2$, the law of thermal dynamics asserts that the stationary variance of the displacement should be $E[x^2(0)] = \frac{k_BT}{m\psi}$. Notably, our model (5.6) and its overdamped version (5.7) both satisfy this requirement[47].

Furthermore, the models can be derived from the physical microscopic interaction between the particle and its surrounding media. We start from the Hamiltonian (which is essentially the total energy) of the particle $H_s = \frac{p^2}{2m} + \frac{1}{2}m\psi x^2$, where $p = mv$ is the momentum, $p^2/(2m)$ is the kinetic energy, x is the displacement, and $m\psi x^2/2$ is the potential energy under the harmonic case. The surrounding media, consisting of $N \sim 10^{23}$ small molecules, has its own Hamiltonian

$$H_B = \sum_{j=1}^{N}\left(\frac{p_j^2}{2m_b} + \frac{1}{2}m_b\omega_j^2\left(q_j - \frac{\gamma_j}{\omega_j^2}x\right)^2\right),$$

where m_b is the (common) mass of an individual molecule in the media, p_j, q_j and ω_j are, respectively, the momentum, location and oscillation frequency of the j-th molecule in the media, and γ_j is the coupling strength between the particle of interest and the j-th molecule. Once the total Hamiltonian $H_s + H_B$ is given, the classical theory of mechanics[49] states that the motion of the particle as well as that of the media molecules is given by

$$\frac{dx}{dt} = \frac{\partial(H_s + H_B)}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial(H_s + H_B)}{\partial x},$$
$$\frac{dq_j}{dt} = \frac{\partial(H_s + H_B)}{\partial p_j}, \quad \frac{dp_j}{dt} = -\frac{\partial(H_s + H_B)}{\partial q_j}.$$

Solving them[47, 44] leads to

$$m\frac{dv(t)}{dt} = -m\psi x(t) - \zeta\int_0^t K(t-s)v(s)ds + G(t), \quad x(t) = \int_0^t v(s)ds, \tag{5.8}$$

where $K(t) \propto m_b\sum_{j=1}^{N}\gamma_j^2\cos(\omega_j t)/\omega_j^2$. Equation (5.8) is of exactly the same form as (5.6). If the media molecules are such that $\sum_j m_b\gamma_j^2\cos(\omega_j t)/\omega_j^2 \to 2H(2H-1)t^{2H-2}$, then the memory kernel (5.3) is also obtained.

Explaining the single-molecule experimental results. A recent single-molecule experiment[9] studied a protein-enzyme compound Fre, which is involved in the DNA synthesis of *E. Coli*. As shown in Figure 9, Fre contains two subunits: FAD and Tyr. Because the 3-D conformation of Fre spontaneously fluctuates over time, the (edge-to-edge) distance between FAD and Tyr varies. This distance fluctuation was probed in the experiment. Fre is placed under a laser beam. The laser excites FAD to be fluorescent. By recording the fluorescence lifetime of FAD, one can trace the distance between FAD and Tyr, because at any time t the fluorescence lifetime $\lambda(t)$ of FAD is a function of the distance[50, 51]:

$$\lambda(t) = k_0e^{\beta(x_{eq}+x(t))},$$

where k_0 and β are known constants[51], x_{eq} is the mean distance, and $x(t)$ with mean 0 is the distance fluctuation at time t.

To model $x(t)$, researchers used to describe $x(t)$ as a Brownian diffusion process under the harmonic potential $m\frac{d}{dt}v(t) = -\zeta v(t) - m\psi x(t) + F(t),\ x(t) = \int_0^t v(s)ds$, or by its overdamped version $m\psi x(t) = -\zeta v(t) + F(t),\ x(t) = \int_0^t v(s)ds$, where $F(t)$ is the white noise.

The nanoscale single-molecule experimental data of $\lambda(t)$ provides the means to test the model. One can calculate the empirical autocorrelation function of $\lambda(t)$ from the experimental data and compare it with the theoretical autocorrelation function from the model. The autocorrelation function is used as the test statistic because the experimentally recorded fluorescence lifetime is actually the true $\lambda(t)$ plus background and equipment noise. Using autocorrelation effectively removes the noise, since the noise is uncorrelated. Figure 10 shows the empirical autocorrelation function (the open circles) compared with the best fitting from the Brownian diffusion model (the dashed curve). A clear discrepancy is seen. On the other hand, the solid line in Figure 10 is the result from modeling $x(t)$ by our subdiffusive process (5.7). The curve is fitted by using the Hurst parameter $H = 0.74$, and the analytical solution of (5.7). A very close agreement with the experimental autocorrelation function is seen.

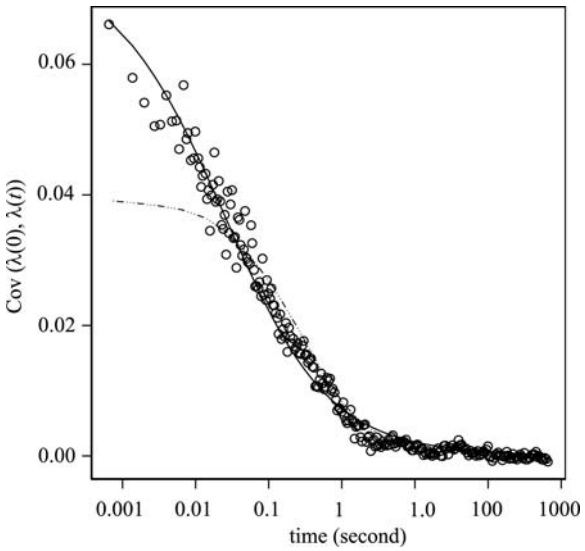

Figure 10 Autocorrelation function of the fluorescence lifetime $\lambda(t)$. The open circles represent the empirical autocorrelation from the experimental data. The dashed line is the best fit from the classical Brownian diffusion model. The solid line is the fit ($H = 0.74$, $\zeta/(m\psi) = 0.40$, $\beta^2 k_B T/(m\psi) = 0.81$) from our model (5.7).

As a model checking, we make predictions about the distance fluctuation and test whether these predictions can be confirmed by the experiments. The first set of predictions involves higher order autocorrelation functions because they are very sensitive to distinguishing models[52]. With the values of the fitting parameters fixed to those in Figure 10, we compute from the model the predicted three-step and four-step autocorrelation functions $E[\Delta\lambda(0)\,\Delta\lambda(t_1)\,\Delta\lambda(t_1{+}t_2)]$ and $E[\Delta\lambda(0)\,\Delta\lambda(t_1)\,\Delta\lambda(t_1+t_2)\,\Delta\lambda(t_1+t_2+t_3)]$, where $\Delta\lambda(t) = \lambda(t) - E[\lambda(t)]$, and compare them with their experimental counterparts. Figure 11(a) and (b) show, respectively, the three-step and four-step autocorrelation functions $E[\Delta\lambda(0)\,\Delta\lambda(t)\,\Delta\lambda(2t)]$ and $E[\Delta\lambda(0)\,\Delta\lambda(t)\,\Delta\lambda(2t)\,\Delta$

$\lambda(3t)]$ as functions of time t. The theoretical curves (the solid lines) from our model agree well with the experimental values (the open circles).

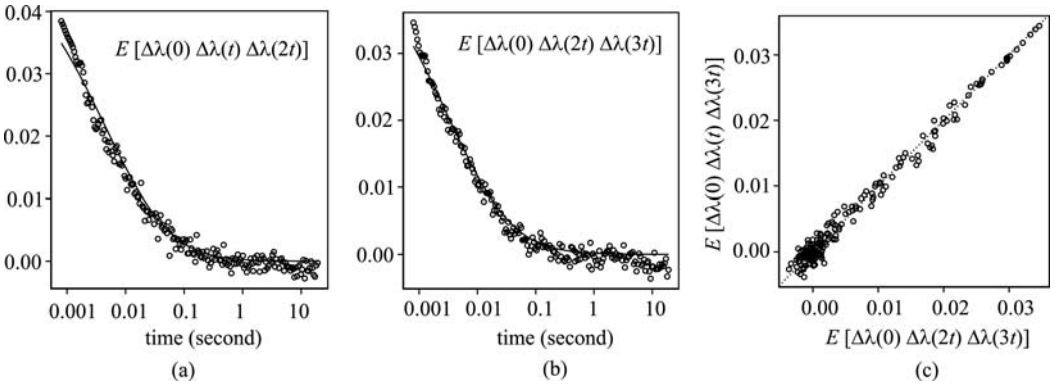

Figure 11 Model predictions compared with experimental data. (a) and (b): The experimentally obtained autocorrelation functions $E[\Delta\lambda(0)\,\Delta\lambda(t)\,\Delta\lambda(2t)]$ and $E[\Delta\lambda(0)\,\Delta\lambda(t)\,\Delta\lambda(2t)\,\Delta\lambda(3t)]$ overlaid with the model predictions for various t. The theoretical curves from our model are calculated using the same parameter values as in Figure 10. (c): A test for time-symmetry. The experimental three-step correlations $E[\Delta\lambda(0)\,\Delta\lambda(t)\,\Delta\lambda(3t)]$ and $E[\Delta\lambda(0)\,\Delta\lambda(2t)\,\Delta\lambda(3t)]$ are plotted again each other for various t. A 45° line is predicted by our model.

The second prediction from the model is time-symmetry. For any t_1 and t_2, the model predicts $E[\Delta\lambda(0)\,\Delta\lambda(t_1)\,\Delta\lambda(t_1+t_2)] = E[\Delta\lambda(0)\,\Delta\lambda(t_2)\,\Delta\lambda(t_1+t_2)]$. It says that if our model is true, then one can swap the order of the time lags without changing the correlation value. This can be tested by taking $t_1 = t$, $t_2 = 2t$ and plotting the experimentally obtained three-time correlation $E[\Delta\lambda(0)\,\Delta\lambda(t)\,\Delta\lambda(3t)]$ against $E[\Delta\lambda(0)\,\Delta\lambda(2t)\,\Delta\lambda(3t)]$ for various t. A 45° line is predicted by the model. The experimental plot in Figure 11(c) indeed confirms the prediction.

6 Modeling enzymatic reaction of single proteins

In this section we will consider the stochastic modeling problems raised in recent single-molecule experiments on enzymatic reactions, where the high resolution experimental results showed a surprising departure from what classical theory predicts.

According to the classical Michaelis-Menten (MM) model of enzymatic reaction in biochemistry[53], an enzyme catalyzes a reaction in the following way. First, the enzyme binds to the substrate (i.e., the reactant molecule) and forms an enzyme-substrate complex. The complex then undergoes a decomposition to generate the reaction product and release the enzyme to its original form to catalyze the next substrate. In the biochemistry literature, this process is typically diagrammed as

$$E + S \underset{k_{-1}}{\overset{k_1[S]}{\rightleftarrows}} ES \overset{k_2}{\to} E^0 + P, \quad E^0 \overset{\delta}{\to} E, \tag{6.1}$$

where E (and E^0), S, ES, and P stand for the enzyme, the substrate, the enzyme-substrate complex, and the reaction product, respectively. The symbol $[S]$ denotes the substrate concentration; k_1 is the association rate (per unit substrate concentration); k_{-1} and k_2 are, respectively, the dissociation and catalytic rate, and δ is the rate of E^0's return to E. In our familiar statistics language, diagram (6.1) corresponds to a three-state continuous-time Markov chain

with the infinitesimal generator

$$Q_{\mathrm{MM}} = \begin{pmatrix} -k_1[S] & k_1[S] & 0 \\ k_{-1} & -(k_{-1}+k_2) & k_2 \\ \delta & 0 & -\delta \end{pmatrix}.$$

An enzyme molecule switches continuously among the three states E, ES, and E^0 according to it.

The time needed for an enzyme to complete one catalytic cycle is called the *turnover time*. The reciprocal of the mean turnover time is defined as the enzymatic *reaction rate*[54−56].

In the MM model, the turnover time is the first passage time from E to E^0. It can be shown[55, 57] that the density function of this first passage time is given by

$$f(t) = \frac{k_1 k_2[S]}{2p}(e^{-(q-p)t} - e^{-(q+p)t}), \tag{6.2}$$

where $p = \sqrt{(k_1[S]+k_2+k_{-1})^2/4 - k_1k_2[S]}$ and $q = (k_1[S]+k_2+k_{-1})/2$. This explicit description, together with (6.1), has important experimental implications for the MM model. First, (6.2) says that the turnover time's distribution should have an exponential decay with rate $q-p$. In addition, due to the exponential nature, for most values of t, $e^{-(q-p)t}$ easily overwhelms $e^{-(q+p)t}$; thus, $f(t)$ is almost a purely exponential distribution, and will yield a practically straight line on a log-linear plot. Figure 12 provides an illustration, plotting $f(t)$ on a log-linear scale for typical values of $[S]$, k_1, k_2, and k_{-1}; a clear linear pattern is shown.

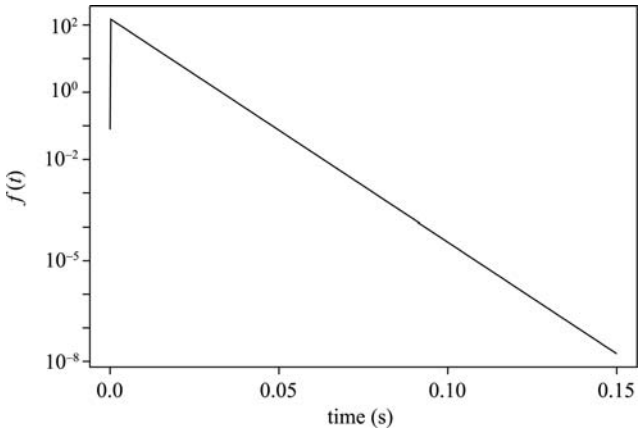

Figure 12 The density function $f(t)$ of the turnover time from the MM model plotted on a log-linear scale. $[S] = 100$ μM (micro-molar), $k_1 = 5 \times 10^7$ M^{-1}s^{-1}, $k_2 = 730$ s^{-1}, $k_{-1} = 18300$ s^{-1}.

Second, since an enzyme behaves as a Markov chain in the MM model, it follows immediately from the Markov property that an enzyme's successive turnover times should be independently and identically distributed of each other. No memory should be found among the turnover times.

Third, from (6.2) we know that the enzymatic *reaction rate* under the MM model is

$$v = \frac{1}{\int_0^\infty f(t)t\,dt} = \frac{k_2[S]}{[S] + (k_2+k_{-1})/k_1}. \tag{6.3}$$

This relationship, referred to as the Michaelis-Menten equation, is of fundamental importance in the biochemistry literature[53, 58]: It gives an explicit hyperbolic dependence of the reaction rate v on the substrate concentration $[S]$.

Before the single-molecule experiments were possible, numerous researchers had studied different enzyme systems under the traditional experimental approach. Unable to track an individual enzyme molecule, the traditional experiments relied on a population of enzymes, and by measuring the accumulation of reaction products over time, they estimated the reaction rate for various substrate concentrations. It was found in these traditional experiments that the hyperbolic form in (6.3), i.e.,

$$v \propto \frac{[S]}{[S]+C} \quad \text{with some constant } C$$

appeared to hold for many enzymes. Thus for decades the MM model has been featured in textbooks as the fundamental mechanism for enzymatic reactions[58–60].

Advances in nanotechnology have made it possible to study enzymatic reactions at the single-molecule level[61]. English et al.[62] recently carried out single-molecule experiments on β-galactosidase, an essential enzyme that catalyzes the breakdown of the sugar lactose[63, 64]. The experimental results surprised researchers, as the high resolution data clearly demonstrated that: (a) the empirical distribution of the experimentally recorded turnover times is much heavier (skewed) than an exponential one; (b) a single enzyme's successive turnover times are highly correlated; and (c) The hyperbolic relationship of $v \propto [S]/([S]+C)$ appears to hold for the single-molecule data.

Some questions immediately arise from these observations. First, what causes the turnover time's heavier-than-exponential distribution? Second, how can an enzyme "remember" its past, and from where does the memory come? Third, given that findings (a) and (b) have contradicted the fundamentals of the MM model, how can the hyperbolic formula derived from it still hold?

A stochastic network model was introduced in [55] to answer these questions. The model is based on the experimental insight that enzymes are not rigid entities but rather *dynamic* biomolecules, experiencing constant fluctuations in their 3-D conformations[10, 9, 33, 32], which suggests that one should not treat an enzyme as an object with a fixed state, but as a collection of (interconverting) states (with each state being a distinct conformation of the enzyme). We thus propose the following stochastic network model for enzymatic reactions[55], diagramed as:

$$\begin{array}{ccccc}
S+E_1 & \underset{k_{-11}}{\overset{k_{11}[S]}{\rightleftarrows}} & ES_1 & \overset{k_{21}}{\to} P+E_1^0 & E_1^0 \overset{\delta_1}{\to} E_1 \\
\downarrow\uparrow & & \downarrow\uparrow & \downarrow\uparrow & \vdots \\
S+E_2 & \underset{k_{-12}}{\overset{k_{12}[S]}{\rightleftarrows}} & ES_2 & \overset{k_{22}}{\to} P+E_2^0 & E_2^0 \overset{\delta_2}{\to} E_2 \\
\vdots & & \vdots & \vdots & \vdots \\
\downarrow\uparrow & & \downarrow\uparrow & \downarrow\uparrow & \\
S+E_n & \underset{k_{-1n}}{\overset{k_{1n}[S]}{\rightleftarrows}} & ES_n & \overset{k_{2n}}{\to} P+E_n^0 & E_n^0 \overset{\delta_n}{\to} E_n,
\end{array} \tag{6.4}$$

where $E_1, E_2, \ldots$ represent the different states (conformations) of the original enzyme, and ES_i

and E_i^0 are the states corresponding to subsequent enzyme-substrate binding and decomposition. k_{1i} is the association rate (per unit concentration) for the i-th state E_i; k_{-1i}, k_{2i}, and δ_i are, respectively, the dissociation, catalytic, and returning rates.

The transitions between E_i and E_j $(i \neq j)$ in the model capture the (conformational) fluctuation of the enzyme. Different states, due to their specific spatial conformation, could have different reactivity levels. This is embodied in the model by allowing k_{1i}, k_{-1i}, k_{2i}, and δ_i to take distinct values for different i.

The model (6.4) generalizes the classical MM model to a stochastic network[65–67]. Let α_{ij}, β_{ij} and γ_{ij} $(i \neq j)$ be the transition rates of $E_i \to E_j$, $ES_i \to ES_j$ and $E_i^0 \to E_j^0$, respectively. Then the stochastic network (6.4) can be described as a continuous-time Markov chain with infinitesimal generator

$$Q_{net} = \begin{pmatrix} Q_{AA} - Q_{AB} & Q_{AB} & 0 \\ Q_{BA} & Q_{BB} - (Q_{BA} + Q_{BC}) & Q_{BC} \\ Q_{CA} & 0 & Q_{CC} - Q_{CA} \end{pmatrix}, \tag{6.5}$$

where the square matrix Q_{AA} represents the transition rates among the E_i states: $(Q_{AA})_{ij} = \alpha_{ij}$ for $i \neq j$, $(Q_{AA})_{ii} = -\sum_{j \neq i} \alpha_{ij}$. Likewise the matrices Q_{BB} and Q_{CC} represent the transition rates among the ES_i states and E_i^0 states, respectively. The diagonal matrices Q_{AB}, Q_{BA}, Q_{BC} and Q_{CA} denote the transition rates of $E_i \to ES_i$, $ES_i \to E_i$, $ES_i \to E_i^0$, and $E_i^0 \to E_i$: $Q_{AB} = \mathrm{diag}(k_{11}[S], \ldots, k_{1n}[S])$, $Q_{BA} = \mathrm{diag}(k_{-11}, \ldots, k_{-1n})$, $Q_{BC} = \mathrm{diag}(k_{21}, \ldots, k_{2n})$ and $Q_{CA} = \mathrm{diag}(\delta_1, \ldots, \delta_n)$.

In model (6.4), an enzyme's turnover time is the first passage time from the first reaction stage to the third stage, i.e., from any E_i state to any E_j^0 state. For example, suppose an enzyme goes through the following path: $E_1^0 \to E_1 \to E_2 \to ES_2 \to E_2^0 \to E_2 \to E_3 \to ES_3 \to ES_1 \to S_1 \to ES_1 \to E_1^0$, then the first turnover time corresponds to $E_1 \to E_2 \to ES_2 \to E_2^0$, and the second corresponds to $E_2 \to E_3 \to ES_3 \to ES_1 \to S_1 \to ES_1 \to E_1^0$. The feature that a turnover event can start from any E_i and end in any E_j^0 in our model captures the fact that in a single-molecule experiment, instead of observing the specific enzyme conformations and their interconversions, one can record only the time for an enzyme to complete a reaction cycle (see the description of real experiments below). In other words, on the network (6.4), the exact states are not observed, and only transitions from the set $\{E_1, \ldots, E_n\}$ to the set $\{E_1^0, \ldots, E_n^0\}$ are observed.

The stochastic model (6.4) provides an explanation to the experimental puzzles. First, it can be shown[57] that under it the stationary turnover time distribution $f_{eq}(t)$ can be expressed as a mixture of exponentials

$$f_{eq}(t) = \sum_{i=1}^{2n} \sigma_i \lambda_i e^{-\lambda_i t},$$

where σ_i and λ_i are related to the eigen-values and vectors of the matrix (6.5). See [57] for the detailed (but lengthy) expressions. This result implies that as long as there are multiple states (conformations) in the network ($n > 1$), the distribution in general would be heavier (skewer) than a single exponential one; thus, in particular, if one plots the empirical distribution of successive turnover times on a logarithmic scale, instead of observing a straight line indicating

a single-exponential tail, one would find a line skewed to the right, which is exactly the first puzzle observed in the single-molecule experiments.

Second, the stochastic network model leads to a direct explanation of the memory between successive turnover times. To make the idea transparent, imagine there are only two states E_1 and E_2 for illustration *per se*:

$$\begin{array}{ccccc} S+E_1 & \underset{k_{-11}}{\overset{k_{11}[S]}{\rightleftarrows}} & ES_1 \overset{k_{21}}{\to} P+E_1^0 & \qquad & E_1^0 \overset{\delta_1}{\to} E_1 \\ \downarrow\uparrow & & \downarrow\uparrow & & \downarrow\uparrow \qquad\qquad \vdots \\ S+E_2 & \underset{k_{-12}}{\overset{k_{12}[S]}{\rightleftarrows}} & ES_2 \overset{k_{22}}{\to} P+E_2^0 & \qquad & E_2^0 \overset{\delta_2}{\to} E_2. \end{array}$$

Suppose the transitions of $E_1 \leftrightarrow E_2$, $ES_1 \leftrightarrow ES_2$, and $E_1^0 \leftrightarrow E_2^0$ are all infrequent. Then it is easily seen that if a turnover event starts from E_1 (E_2) it is highly likely that the next turnover event will also start from E_1 (E_2). Now imagine furthermore that E_1 and E_2 have different reactivity levels; for example, the transitions of $S+E_1 \rightleftarrows ES_1 \to P+E_1^0$ and $E_1^0 \to E_1$ are all fast, while the transitions of $S+E_2 \rightleftarrows ES_2 \to P+E_2^0$ and $E_2^0 \to E_2$ are all slow. Then it is clear that a slow (fast) turnover will likely be followed by another slow (fast) turnover, naturally producing the correlation between successive turnover times.

Third, it can also be shown[57] that under general conditions (such as slow conformational fluctuation) the hyperbolic relationship of $v \propto [S]/([S]+C)$ between the reaction rate $v = 1/\int_0^\infty t f_{eq}(t)dt$ and the substrate concentration $[S]$ holds in the stochastic network model.

In summary, the stochastic network model offers a resolution of the experimental puzzles. Although the classical MM model gives the description of $v \propto [S]/([S]+C)$, observing such a relationship in experiments by no means implies that the MM model is the underlying mechanism because the MM model is only one of many that display such a relationship – the discovery of memory and heavier-than-exponential distribution of the turnover times in fact points to the opposite direction.

Real single-molecule experimental data. English et al.[62] studied β-galactosidase (β-gal). In the experiment a single β-gal molecule is immobilized (to a bead) so that its enzymatic turnovers can be continuously monitored under a fluorescence microscope. To detect the individual turnovers, careful treatments were carried out so that once the experimental system was placed under a laser beam the reaction product and *only* the reaction product was fluorescent. In other words, as the β-gal enzyme catalyzes one substrate molecule after another, strong fluorescence signal is emitted and detected only when a product is released. Recording the fluorescence signals over time thus enables the experimental determination of individual turnovers.

Figure 13(a) presents a schematic picture of the experimental setup. Figure 13(b) shows a typical fluorescence intensity reading from the experiment: each vertical bar is a fluorescence intensity spike generated by the release of *one* reaction product. The time lag between two adjacent fluorescence spikes is the enzymatic turnover time. Thus taking the time lag between every two consecutive fluorescence spikes gives the successive turnover times of the β-gal molecule. To investigate how substrate concentration $[S]$ affects the turnover times, the experiment was

repeated at different levels of $[S]$; throughout each repetition the substrate concentration $[S]$ is held at a fixed level.

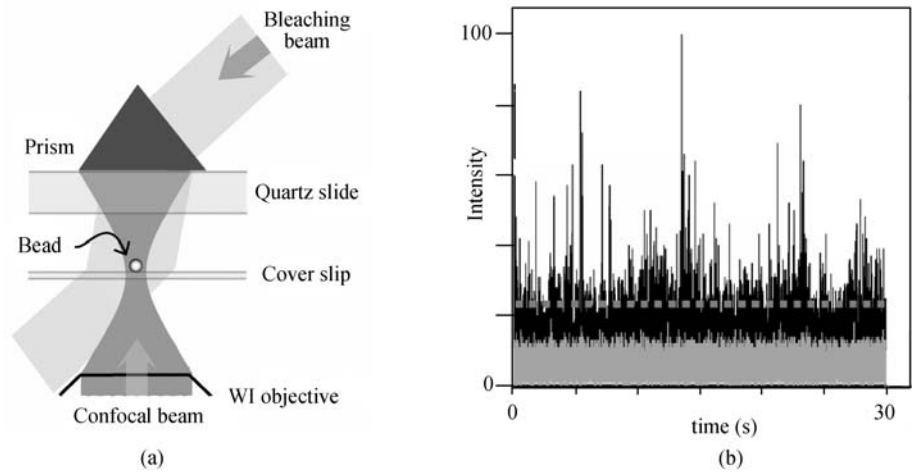

Figure 13 (a) Schematic presentation of the experimental setup. A single β-gal molecule is immobilized to a bead on a glass coverslip. Two laser beams (confocal beam and bleaching beam) are applied to collect the fluorescence signals. (b) Experimental fluorescence intensity reading. Each fluorescence intensity spike is caused by the release of one reaction product.

The empirical distributions of the experimental turnover times obtained at four substrate concentrations $[S] = 10\ \mu$M, $20\ \mu$M, $50\ \mu$M, and $100\ \mu$M (micro-molar) are plotted in Figure 14 on a log-linear scale (open circles, filled circles, open squares, and filled squares, respectively). Rather than following straight lines on the logarithmic scale as the MM model predicts, the empirical distributions have skewed tails at high substrate concentrations.

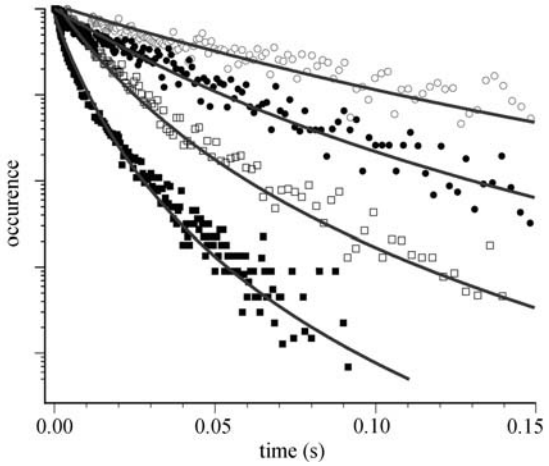

Figure 14 Empirical distributions of the turnover times on a log-linear scale. The open circles, filled circles, open squares, and filled squares represent experimental data obtained at substrate concentrations $10\ \mu$M, $20\ \mu$M, $50\ \mu$M, and $100\ \mu$M respectively. The solid curves are the fittings from our model using equation (6.6), where the fitted parameter $\hat{k}_1 = 5.01 \times 10^7\,\mathrm{M}^{-1}\mathrm{s}^{-1}$, $\hat{k}_{-1} = 1.83 \times 10^5\,\mathrm{s}^{-1}$, $\hat{a} = 4.25$ and $\hat{b} = 220\,\mathrm{s}^{-1}$.

As a check of our stochastic network model, we fit it to the empirical turnover time distributions. It can be shown[55, 57] that under the assumptions of homogeneous enzyme binding (i.e., $k_{11} = k_{12} = \cdots = k_{1n} \equiv k_1$, $k_{-11} = k_{-12} = \cdots = k_{-1n} \equiv k_{-1}$) and slow (enzyme) conformational fluctuation[10, 9, 32], the stationary turnover time distribution in the model is

$$f_{eq}(t) = \int_0^\infty w(k_2)\frac{k_1 k_2[S]}{2p(k_2)}(e^{-[q(k_2)-p(k_2)]t} - e^{-[q(k_2)+p(k_2)]t})dk_2 \quad \text{with} \tag{6.6}$$

$$p(k_2) = \sqrt{\frac{1}{4}(k_1[S] + k_2 + k_{-1})^2 - k_1 k_2[S]}, \quad q(k_2) = \frac{1}{2}(k_1[S] + k_2 + k_{-1}),$$

where $w(k_2)$ is the distribution of the k_{2i}'s. Since the simplest distribution over the positive real line is the gamma distribution, for model fitting we take $w(k_2) = k_2^{a-1}\exp(-k_2/b)/[b^a\Gamma(a)]$, a gamma density. Compared with (6.4), now (6.6) only has four parameters: k_1, k_{-1}, a and b.

The maximum likelihood fitting of (6.6) to the experimental data is shown in Figure 14 as the solid curves (overlaid on the empirical distributions). The fitted parameter values are given in the Figure caption. For all four substrate concentrations, close agreement between the theoretical curves and the experimental values is evident.

Experimental relationship between reaction rate and substrate concentration. At each substrate concentration $[S]$, the reaction rate can be directly estimated from the experimental turnover times $\tau_1, \tau_2, \ldots, \tau_N$ via $\hat{v} = 1/\bar{\tau}$. If the hyperbolic relationship of $v = \chi[S]/([S] + C_M)$ holds, then a plot of $1/v$ versus $1/[S]$ should yield a straight line with slope C_M/χ and intercept $1/\chi$. Figure 15(a) graphs $1/\hat{v}$ versus $1/[S]$ from experimental data. Notably a linear pattern indeed emerges. A simple least-square fit (the black line in Figure 15(a)) gives $\hat{\chi} = 730 \pm 40\,\mathrm{s}^{-1}$ and $\hat{C}_M = 390 \pm 30\,\mu$M.

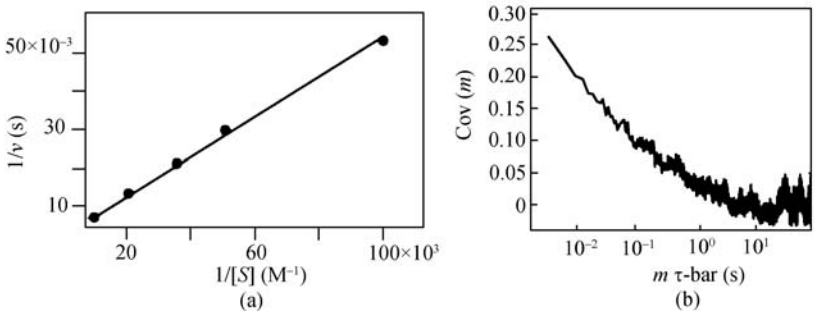

Figure 15 (a) A plot of $1/\hat{v}$ versus $1/[S]$ from the experimental data. The reaction rate $\hat{v}$ at each point is calculated from the experimental data at the corresponding substrate concentration. The black line is the least-square fit with $\hat{\chi} = 730 \pm 40\,\mathrm{s}^{-1}$ and $\hat{C}_M = 390 \pm 30\,\mu$M. (b) Turnover time autocorrelation function. $\mathrm{Cov}(m)$ is plotted against $m\bar{\tau}$ from the experimental data at substrate concentration $[S] = 100\,\mu$M.

As a consistency check of the model, we compute from formula (6.6) that

$$v = \frac{1}{\int_0^\infty t\, f_{eq}(t)dt} = \frac{b(a-1)\,[S]}{[S] + (k_{-1} + b(a-1))/k_1} \equiv \frac{\chi'[S]}{[S] + C'_M}.$$

Plugging in the MLEs of Figure 14, we note that $\hat{\chi}' = \hat{b}(\hat{a}-1) = 715\,\mathrm{s}^{-1}$ and $\hat{C}'_M = (\hat{k}_{-1} + \hat{b}(\hat{a}-1))/\hat{k}_1 = 380\,\mu$M agree well with the least-square nonparametric estimates of $\hat{\chi} = 730 \pm 40\,\mathrm{s}^{-1}$ and $\hat{C}_M = 390 \pm 30\,\mu$M above.

Experimental autocorrelation of turnover times. From the experimental successive turnover times $\tau_1, \tau_2, \ldots, \tau_N$, one can calculate their empirical autocovariance

$$\mathrm{Cov}(m) = \frac{1}{N-m}\sum_i (\tau_i - \bar{\tau})(\tau_{i+m} - \bar{\tau}).$$

Figure 15(b) shows the empirical autocorrelation function, plotting the normalized $\mathrm{Cov}(m)$ against $m\bar{\tau}$ for $m = 1, 2, \ldots$ at the substrate concentration $[S] = 100\,\mu$M. Instead of a flat horizontal line at zero as the MM model would predict, a clear memory effect is seen. The experimental data at other substrate concentrations showed similar correlation picture. The evident memory indicates strongly that the classical MM missed important aspects of real enzymatic reactions and that models that can account for the memory are necessary.

7 Discussion

The advances in nanoscale (single-molecule) biophysics have generated much excitement from biologists, chemists, and biophysicists, as they hold promise for new scientific discoveries. They also raise many interesting statistical inference and stochastic modeling problems, owing to the stochastic nature of the single-molecule world. If in the past some physical scientists had been resistant to advanced statistical methods (due to the remarkably high signal-to-noise ratio in classical experiments), the nanoscale development has significantly altered the landscape.

The statistical inference problems include both parametric and nonparametric ones, as illustrated in this paper. Parametric inference questions arise because in many cases there are well established models out of the basic understanding of physics, chemistry and biology. Nonparametric inference, on the other hand, is well suited for studying new or complex phenomena, where comprehensive theory is yet to be established, and for testing/validating existing theories. Three distinctive features underlie both the parametric and nonparametric analyses of single-molecule data. First, the data collected in the experiments are usually not of our familiar i.i.d. or independence type. They are rather inherently stochastic. Second, the inference is often complicated by the presence of latent noise, which itself can possess stochastic structures (such as governed by unobserved molecular Brownian motion). Third, since fluorescence technique is widely used in biophysical experiments to investigate the biological processes of interest, (doubly stochastic) Poisson process or general point process type of arrival or spike data are broadly present. Each of these features raises distinct statistical inference problems, as we have seen in this paper.

The study of nanoscale biophysics also brings many new stochastic modeling problems. While some can be addressed by applying existing stochastic tools, such as the utilization of stochastic network to model single-molecule enzymatic reaction in Section 6, others require new theoretical frameworks, as the modeling of subdiffusive motion within a single protein molecule in Section 5 illustrates. Nanoscale biophysics, hence, presents opportunities for both applied and theoretical probabilists. For instance, from a pure theoretical angle, how to solve the GLE (5.5) with an arbitrary potential $U(x)$ is an important open problem; its answer directly relates to the understanding of many biological and chemical systems. One distinct feature underlying the construction of biophysical models is the requirement that the models should have sound physical meaning and must agree with fundamental physical laws, since the randomness in

individual molecules' behavior is, after all, governed by statistical and quantum mechanics.

The problems presented in this paper exemplify only a few instances of the numerous and growing research opportunities in nanoscale biophysics. We hope they will serve to generate further interest in applying modern statistical and probabilistic methodology to interesting biophysical and scientific problems.

Acknowledgements The author is grateful to the Xie group at the Department of Chemistry and Chemical Biology of Harvard University for sharing the experimental data and for fruitful discussions, and to Professor Jun Liu for helpful discussions.

References

1 Feynman R P, Leighton R B, Sands M. The Feynman Lectures on Physics, Vol. 1. Reading, Massachusetts: Addison-Wesley, 1963
2 Moerner W. A dozen years of single-molecule spectroscopy in physics, chemistry, and biophysics. *J Phys Chem B*, **106**: 910–927 (2002)
3 Nie S, Zare R. Optical detection of single molecules. *Ann Rev Biophys Biomol Struct*, **26**: 567–596 (1997)
4 Tamarat P, Maali A, Lounis B, et al. Ten years of single-molecule spectroscopy. *J Phys Chem A*, **104**: 1–16 (2000)
5 Weiss S. Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature Struct Biol*, **7**: 724–729 (2000)
6 Xie X S, Lu H P. Single-molecule enzymology. *J Bio Chem*, **274**: 15967–15970 (1999)
7 Xie X S, Trautman J K. Optical studies of single molecules at room temperature. *Ann Rev Phys Chem*, **49**: 441–480 (1998)
8 Asbury C, Fehr A, Block S M. Kinesin moves by an asymmetric hand-over-hand mechanism. *Science*, **302**: 2130–2134 (2003)
9 Yang H, Luo G, Karnchanaphanurach P, et al. Protein conformational dynamics probed by single-molecule electron transfer. *Science*, **302**: 262–266 (2003)
10 Lu H P, Xun L, Xie X S. Single-molecule enzymatic dynamics. *Science*, **282**: 1877–1882 (1998)
11 Krichevsky O, Bonnet G. Fluorescence correlation spectroscopy: the technique and its applications. *Rep Progr Phys*, **65**: 251–297 (2002)
12 Reilly P D, Skinner J L. Spectroscopy of a chromophore coupled to a lattice of dynamic two-level systems. *J Chem Phys*, **101**: 959–973 (1994)
13 Eggeling C, Fries J, Brand L, et al. Monitoring conformational dynamics of a single molecule by selective fluorescence spectroscopy. *Proc Natl Acad Sci*, **95**: 1556–1561 (1998)
14 Van Kampen N G. Stochastic Processes in Physics and Chemistry. New York: Elsevier Science, 2001
15 Kou S C, Xie X S, Liu J S. Bayesian analysis of single-molecule experimental data (with discussion). *J Roy Statist Soc Ser C*, **54**: 469–506 (2005)
16 Tanner M A, Wong W H. The calculation of posterior distributions by data augmentation (with discussion). *J Amer Statist Assoc*, **82**: 528–540 (1987)
17 Lauritzen S L, Spiegelhalter D J. Local computations with probabilities on graphical structures and their application to expert systems. *J R Stat Soc Ser B Stat Methodol*, **50**: 157–224 (1988)
18 Liu J S. Monte Carlo Strategies in Scientific Computing. New York: Springer, 2001
19 Zazopoulos E, Lalli E, Stocco D, et al. DNA binding and transcriptional repression by DAX-1 blocks steroidogenesis. *Nature*, **390**: 311–315 (1997)
20 Froelich-Ammon S, Gale K, Osheroff N. Site-specific cleavage of a DNA hairpin by topoisomerase II. DNA secondary structure as a determinant of enzyme recognition/cleavage. *J Biol Chem*, **269**: 7719–7725 (1994)
21 Trinh T, Sinden R. The influence of primary and secondary DNA structure in deletion and duplication between direct repeats in Escherichia coli. *Genetics*, **134**: 409–422 (1993)
22 Pfluegl W, Brown F L, Silbey R J. Variance and width of absorption lines of single molecules in low temperature glasses. *J Chem Phys*, **108**: 6876–6883 (1998)
23 Brown F L, Silbey R J. An investigation of the effects of two level system coupling on single molecule lineshapes in low temperature glasses. *J Chem Phys*, **108**: 7434–7450 (1998)
24 Schenter G K, Lu H P, Xie X S. Statistical analyses and theoretical models of single-molecule enzymatic

dynamics. *J Phys Chem A*, **103**: 10477–10488 (1999)
25 Agmon N, Hopfield J J. Transient kinetics of chemical reactions with bounded diffusion perpendicular to the reaction coordinate: Intramolecular processes with slow conformational changes. *J Chem Phys*, **78**: 6947–6959 (1983)
26 Bonnet G, Krichevsky O, Libchaber A. Kinetics of conformational fluctuations in DNA hairpin-loops. *Proc Natl Acad Sci*, **95**: 8602–8606 (1998)
27 Ying L, Wallace M, Klenerman D. Two-state model of conformational fluctuation in a DNA hairpin-loop. *Chem Phys Lett*, **334**: 145–150 (2001)
28 Grunwell J, Glass J, Lacoste T, et al. Monitoring the conformational fluctuations of DNA hairpins using single-pair fluorescence energy transfer. *J Amer Chem Soc*, **123**: 4295–4303 (2001)
29 Ansari A, Kuznetsov S V, Shen Y. Configurational diffusion down a folding funnel describes the dynamics of DNA hairpins. *Proc Natl Acad Sci*, **98**: 7771–7776 (2001)
30 Kass R, Raftery A. Bayes factors and model uncertainty. *J Amer Statist Assoc*, **90**: 773–795 (1995)
31 Zhang T, Kou S C. Nonparametric Inference of Doubly Stochastic Poisson Process Data via Kernel Method. Preprint, 2009
32 Min W, Luo G, Cherayil B, et al. Observation of a power law memory kernel for fluctuations within a single protein molecule. *Phys Rev Lett*, **94**: 198302(1)–198302(4) (2005)
33 Min W, English B, Luo G, et al. Fluctuating enzymes: lessons from single-molecule studies. *Acc Chem Res*, **38**: 923–931 (2005)
34 Hill T L. Free Energy Transduction and Biochemical Cycle Kinetics. New York: Springer, 1989
35 Gaspard P. Fluctuation theorem for nonequilibrium reactions. *J Chem Phys*, **120**: 8898–8905 (2004)
36 Min W, Jiang L, Yu J, et al. Nonequilibrium steady state of a nanometric biochemical system: determining the thermodynamic driving force from single enzyme turnover time traces. *Nano Letters*, **5**: 2373–2378 (2005)
37 Bouchaud J, Georges A. Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications. *Phys Rep*, **195**: 127–293 (1990)
38 Klafter J, Shlesinger M, Zumofen G. Beyond Brownian motion. *Physics Today*, **49**: 33–39 (1996)
39 Sokolov I, Klafter J, Blumen A. Fractional kinetics. *Physics Today*, **55**: 48–54 (2002)
40 Kou S C, Xie X S. Generalized Langevin equation with fractional Gaussian noise: subdiffusion within a single protein molecule. *Phys Rev Lett*, **93**: 180603(1)–180603(4) (2004)
41 Risken H. The Fokker-Planck Equation: Methods of Solution and Applications. Berlin: Springer, 1989
42 Chandler D. Introduction to Modern Statistical Mechanics. New York: Oxford University Press, 1987
43 Hill T. An Introduction to Statistical Thermodynamics. New York: Dover, 1986
44 Zwanzig R. Nonequilibrium Statistical Mechanics. New York: Oxford University Press, 2001
45 Embrechts P, Maejima M. Selfsimilar Processes. Princeton, New Jersey: Princeton University Press, 2002
46 Samorodnitsky G, Taqqu M. Stable Non-Gaussian Random Processes. New York: Chapman & Hall, 1994
47 Kou S C. Stochastic modeling in nanoscale biophysics: subdiffusion within proteins. *Ann Appl Statist*, **2**: 501–535 (2008)
48 Reif F. Fundamentals of Statistical and Thermal Physics. Columbus: McGraw-Hill, 1965
49 Corben H C, Stehle P. Classical Mechanics. New York: Dover Publications, 1995
50 Gray H, Winkler J. Electron transfer in proteins. *Annu Rev Biochem*, **65**: 537–561 (1996)
51 Moser C, Keske J, Warncke K, et al. Nature of biological electron transfer. *Nature*, **355**: 796–802 (1992)
52 Mukamel S. Principle of Nonlinear Optical Spectroscopy. New York: Oxford University Press, 1995
53 Atkins P, de Paula J. Physical Chemistry. 7th ed. New York: W. H. Freeman, 2002
54 Yang S, Cao J. Two-event echos in single-molecule kinetics: a signature of conformational fluctuations. *J Phys Chem B*, **105**: 6536–6549 (2001)
55 Kou S C, Cherayil B, Min W, et al. Single-molecule Michaelis-Menten equations. *J Phys Chem B*, **109**: 19068–19081 (2005)
56 Min W, Gopich I V, English B, et al. When does the Michaelis-Menten equation hold for fluctuating enzymes? *J Phys Chem B*, **110**: 20093–20097 (2006)
57 Kou S C. Stochastic networks in nanoscale biophysics: modeling enzymatic reaction of a single protein. *J Amer Statist Assoc*, **103**: 961–975 (2008)
58 Segel I H. Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems. New York: Wiley, 1993
59 Fersht A. Enzyme Structure and Mechanism. 2nd ed. New York: W. H. Freeman, 1985
60 Hammes G G. Enzymatic Catalysis and Regulation. New York: Academic Press, 1982

61 Flomembom O, Klafter J. Stretched exponential decay and correlations in the catalytic activity of fluctuating single lipase molecules. *Proc Natl Acad Sci*, **102**: 2368–2372 (2005)
62 English B, Min W, van Oijen A M, et al. Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited. *Nature Chem Biol*, **2**: 87–94 (2006)
63 Dorland W A. Dorland's Illustrated Medical Dictionary. 30th Ed. Philadelphia: W. B. Sauders, 2003
64 Jacobson R H, Zhang X J, DuBose R F, et al. Three-dimensional structure of β-galactosidase from *E. Coli*. *Nature*, **369**: 761–766 (1994)
65 Ball K, Kurtz T G, Popovic L, et al. Asymptotic analysis of multiscale approximations to reaction networks. *Ann Appl Prob*, **14**: 1925–1961 (2006)
66 Glasserman P, Sigman K, Yao D, eds. Stochastic Networks. Stability and Rare Events. Lecture Notes in Statistics, 117. New York: Springer, 1996
67 Kelly F P, Williams R J, eds. Stochastic Networks. The IMA Volumes in Mathematics and Its Applications, 71. New York: Springer, 1995