Eur Biophys J (2006) 35:633–645
DOI 10.1007/s00249-006-0067-5

ARTICLE

Simulated data sets for single molecule kinetics: some limitations and complications of data analysis

Jue Shi · Ari Gafni · Duncan Steel

Received: 24 January 2006 / Revised: 30 March 2006 / Accepted: 4 April 2006 / Published online: 5 May 2006

Abstract When the fluorescence intensity of a chromophore attached to or bound in an enzyme relates to a specific reactive step in the enzymatic reaction, a single molecule fluorescence study of the process reveals a time sequence in the fluorescence emission that can be analyzed to derive kinetic and mechanistic information. Reports of various experimental results and corresponding theoretical studies have provided a basis for interpreting these data and understanding the methodology. We have found it useful to parallel experiments with Monte Carlo simulations of potential models hypothesized to describe the reaction kinetics. The simulations can be adapted to include experimental limitations, such as limited data sets, and complexities such as dynamic disorder, where reaction rates appear to change over time. By using models that are known a priori, the simulations reveal some of the challenges of interpreting finite single-molecule data sets by employing various statistical signatures that have been identified.

J. Shi · A. Gafni · D. Steel
Biophysics Research Division, University of Michigan,
Ann Arbor, MI 48109, USA

A. Gafni
Department of Biological Chemistry,
University of Michigan, Ann Arbor, MI 48109, USA

D. Steel (✉)
Department of Physics and EECS, University of Michigan,
Ann Arbor, MI 48109, USA
e-mail: dst@umich.edu

Present address: J. Shi
Department of Systems Biology, Harvard Medical School,
Boston, MA 02115, USA

Introduction

Recent experimental and theoretical studies show kinetics of single-molecules provide new information for the study of reaction mechanisms that compliment traditional ensemble methods such as steady-state assays and stopped-flow measurements. New insights into molecular motions (Ishijima et al. 1991; Ariga et al. 2002; Forkey et al. 2003), nucleic acid and protein folding (Ha et al. 1999; Schuler et al. 2002; Yang et al. 2003), protein–DNA interactions (Bianco et al. 2001; Ha et al. 2002; Pease et al. 2005) and enzyme kinetics (Lu et al. 1998; Shi et al. 2004; Zhuang et al. 2002; Oijen et al. 2003) at the single-molecule level have continued to extend our understanding of the mechanical, structural as well as functional dynamics in biological systems. In single molecule spectroscopy, the molecule to be studied has an intrinsically bound or covalently attached chromophore, whose fluorescence intensity differs in different states of the reaction process. By following the time evolution of the fluorescence, the kinetics of single-molecule events, e.g. enzyme catalysis or DNA/RNA folding, can then be monitored. A real-time single-molecule trajectory is a sequence of dwell times of the molecule residing in the various states. These times are the waiting times for a molecule to leave the respective reactive state. Rate constants of the corresponding reaction can be obtained by plotting the distribution of the individual dwell times extracted from the whole single-molecule trajectory.

As an example for enzyme kinetics, flavoenzymes are a convenient class of enzymes for single-molecule studies because the oxidized and reduced states of flavin can be distinguished by the large differences in fluorescence (Lu et al. 1998; Shi et al. 2004). Generally, flavin is fluorescent in the oxidized state but not fluorescent in the reduced state, therefore single flavoenzymes give an on–off blinking fluorescence signal as the enzyme cycles between the oxidized and reduced state. Because the fluorescent on-times and non-fluorescent off-times are the waiting times for flavin reduction and oxidation, respectively, the decay constants derived from the on-time and off-time distributions are the observed kinetic rate constants for the reductive and oxidative half reactions.

Raw single-molecule kinetic data are in the form of a time series of discrete events and appear in a data set fundamentally different from the traditional ensemble data that appear in the form of a continuous function of time, typically given by exponential decays. While studying reaction kinetics in a time series is common in areas such as patch-clamp study of single ion channels (Sakmann and Neher 1995), it is relatively new to the determination of single-molecule kinetics. Moreover, the distinct problems of interest in single-molecule studies, i.e. conformational fluctuations and environmental modulations, entail evaluation of statistical parameters not common in the single ion channel studies. Hence, an in-depth understanding of the kinetic time series and proper statistical analysis requires familiarization with a wide variety of anticipated single-molecule data sets under different reaction models.

In our single-molecule studies of a flavoenzyme, dihydroorotate dehydrogenase from *Escherichia coli*, we observed interesting static heterogeneity in the catalytic rates among different molecules (Shi et al. 2004). However, due to the dissociation of flavin from the holoenzyme, the single-molecule turnover traces that could be obtained experimentally were generally too short to allow statistical calculations for individual traces. It remains unclear whether ambiguity in interpretation of the results from statistical analysis could arise as a result of a small data set and whether the kinetic information derived from various statistical analyses faithfully relate to the reaction model.

To understand the features and characteristics of single-molecule data sets obtained in the laboratory and to provide an additional tool to experimentalists working to interpret single-molecule kinetic data from various systems, e.g., enzymes, DNA, or RNA, we use the Monte Carlo method in this paper to simulate experimental data based on two simple reaction models (measurement noise is assumed to be sufficiently small that it does not cause ambiguity in distinguishing the distinct fluorescent states). The simulated trajectories provide a direct visualization of the distinct characteristics of the single-molecule time series under different reaction models. The simulation results illustrate limitations of standard single-molecule data analysis methods in recovering kinetic parameters from the data. Calculations of various statistical parameters with different values of reaction rates as well as both small and large simulated single-molecule data sets provide new insight into the validity of statistical calculations of different parameters and demonstrate the challenges of using proper statistical analysis techniques when applied to truncated data sets resulting from processes such as photobleaching or dissociation of the fluorophore in single-molecule measurements. This paper extends our previous report of Monte Carlo simulation of enzyme kinetics (Shi et al. 2004), and differs from the recent work of Witkoskie and Cao (2004), who used Monte Carlo simulation to examine the usefulness of various predictors without consideration of experimental limitations.

The specific statistical parameters examined in this simulation study include the dwell time distribution function, the autocorrelation function of dwell times, the correlation function of the fluorescence state, the joint distribution functions of pairs of dwell times, and the difference function of dwell times, each discussed in detail below. We compare the results of Monte Carlo simulation of these statistical parameters with the kinetic parameters we know a priori and with theoretical predictions available in the literatures to investigate their efficacy and usefulness in practice. In particular, the two types of correlation functions, the joint distribution function and the difference function, have easily identifiable signatures for detecting the existence of temporal fluctuations of the reaction rate, sometimes referred to as dynamic disorder (Lu et al. 1998; Edman and Rigler 2000; Kleinekathofer et al. 2003). Kinetic analysis of such dynamic disorder is of great interest because it provides information about relatively slow (compared to the inverse reaction rate) environmental modulation or conformational changes of the molecule under study during the reaction (Zwanzig 1990; Wang and Wolynes 1995; Schenter et al. 1999; Cao 2000; Yang and Cao 2001, 2002; Barsegov et al. 2002; Barsegov and Mukamel 2002; Boguna et al. 2000; Lerch et al. 2002; Vlad et al. 2002; Brown 2003; Chakrabarti and Bagchi 2003).

Monte Carlo simulation of single-molecule enzyme kinetics

In the present study we employ a simple first-order reaction model shown in Fig. 1a, where the molecule leaves state A for state B with a forward reaction rate,

k_a, and converts back to state A with a backward reaction rate, k_b. State A and B are assumed to be distinguishable spectroscopically in single-molecule experiments, i.e. state A is fluorescent (on) and state B is non-fluorescent (off). The on-time that a single-molecule resides in state A is a random variable, dependent on the forward rate k_a. Similarly, the off-time in state B is a random variable depending on the backward rate k_b. Single-molecule events of chemical reactions are easily simulated using the Monte Carlo method to randomly generate a sequence of alternating on- and off-times that conforms to the assumed models.

It is well-known that the dwell time of a molecule being in either state A or B is exponentially distributed. For example, consider the dwell time that the molecule resides in the fluorescent state A, $k_a\Delta t$ is the probability that the molecule converts from A to B during Δt, while $(1 - k_a\Delta t)$ is the probability that the molecule remains in state A. Therefore, $P_A(t + \Delta t)$, the probability of a molecule being in state A for $t + \Delta t$, is related to $P_A(t)$, the probability of a molecule being in A for time t, by the following equation,

$$P_A(t + \Delta t) = P_A(t)(1 - k_a\Delta t).$$

In the limit of $\Delta t \rightarrow 0$, this defines the differential equation for $P_A(t)$,

$$\frac{\mathrm{d}P_A(t)}{\mathrm{d}t} = -k_a P_A.$$

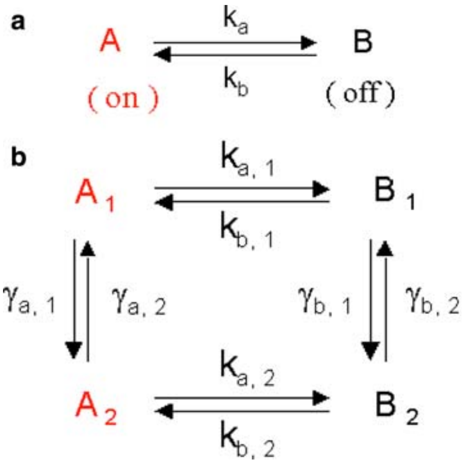

Fig. 1 **a** Reaction scheme of a simple first-order reaction. k_a is the reaction rate leaving state A for state B and k_b is the reaction rate leaving state B for state A. **b** Reaction scheme of a discrete two-conformation model. $k_{a,\ 1}$, $k_{a,\ 2}$, $k_{b,\ 1}$ and $k_{b,\ 2}$ are the reaction rates between state A and B. $\gamma_{a,\ 1}$, $\gamma_{a,\ 2}$, $\gamma_{b,\ 1}$ and $\gamma_{b,\ 2}$ are the interconversion rates between the different conformational states

The solution decays as $\exp(-k_a t)$. Similar rationale can be applied to obtain $P_B(t)$, the probability that the molecule resides in the non-fluorescent state B for time t, as $\exp(-k_b t)$. For more complex reaction models, the dwell time histograms of both the on-times and off-times would fit to multiple exponential decays and the decay constants are functions of the rate constants of distinct reaction steps. Comparison of such equations for probability with equations for the concentrations (i.e., the traditional rate equations) begin to reveal the significant differences between data sets obtained, for example in stopped-flow experiments and single molecule experiments.

The algorithm for randomly generating exponentially distributed on-times (off-times) corresponding to a first-order reaction as shown in Fig. 1a is given by the following: Rate constants are used to calculate the probability, α, of the occurrence of a reaction leaving the respective state. For example, the probability of leaving state A is $\alpha_a = (1 - \exp(-k_a\Delta t))$, when Δt is the time step of data collection and is regarded as the unit time in the simulation. In all of the following simulations, Δt is small so $\alpha = k\ \Delta t$. α is compared to a uniformly distributed random number R, generated by the Random function in the standard Mathematica library, whose value ranges between 0 and 1. Reactions occur when $R < \alpha$; if $R \geq \alpha$, the molecule remains unperturbed and another value of R is generated for the next time step. The number of iterations of random number generation before the occurrence of a reaction, N (iteration counter), is a random variable and has a geometric probability distribution as $\alpha(1 - \alpha)^N$. In the discrete analysis the on-time (off-time) is given by $N \times \Delta t$. While there are numerous Monte Carlo algorithms for generating exponentially-distributed dwell times, for example, the Gillespie simulation algorithm (Gillespie 1977), the current algorithm is chosen because of its simplicity and ease of implementation for the non-expert, and also because it reflects the reality of data acquisition in single-molecule experiments, e.g., the time-step is generally fixed in the laboratory and not adjustable under a feedback mechanism. In this paper, all results are simulated with time step of $\Delta t = 0.01$ s, which is generally the time resolution in single-molecule experiments and is very small compared to the time scale chosen for the reaction kinetics (k_a, k_b < 10 s^{-1}). Simulations have also been run with other lengths of time step ($\Delta t = 0.001$ and 0.005 s) and the results confirm the insensitivity of our results to the time step chosen. In addition, we varied the number of turnovers in the single molecule data sets to evaluate the acceptable lower limit of data size for deriving information from various statistical functions, as discussed below. All results of the simulated single-molecule trajectories are generated by programs written in Mathematica.

In a stochastic two-conformation model, dynamic disorder is envisioned as a conformational fluctuation that results in a reaction rate randomly switching between two distinctive values, such as $k_{a,1}$ and $k_{a,2}$. Kinetics of the interconversion between species with different reaction rates (Fig. 1b) is simulated with an algorithm similar to that mentioned above for the simple first-order reaction. A random number is first generated to determine whether the molecule leaves its current reactive state. If the molecule is to leave for another state, a second random number is then generated to determine whether the molecule converts to a different conformational state (with the same fluorescence characteristic) or a different chemical state (with different fluorescence characteristic). For example, if the molecule is currently in state A_1, $\alpha_{a,1} = (1 - \exp(-(k_{a,1} + \gamma_{a,1})\Delta t))$, where $k_{a,1}$ is the reaction rate and $\gamma_{a,1}$ is the rate of conformation interconversion in state A_1, is compared to a random number to determine whether the molecule leaves state A_1. If the result is the molecule leaves state A_1, $\beta_{a,1} = \gamma_{a,1}/(k_{a,1} + \gamma_{a,1})$ is compared to another random number to determine whether the process of a conformational change (molecule goes to state A_2) or a catalytic reaction (molecule goes to B_1) occurs. The waiting times of the on/off-state are approximated by $N \times \Delta t$, where N is the iteration number for the occurrence of a catalytic reaction.

With eight different kinetic rates in the two-conformation model, it is not practical to consider all possible regions of the parameter space using numerical simulations (a distinct advantage of an analytical approach over a Monte Carlo numerical approach). However, for the purpose of illustrating the challenge of interpreting single-molecule data, we consider three limits in terms of the relationship between the conformation interconversion rates and the reaction rates which typify in a general way many common reaction schemes: $\gamma_{a,I}$, $\gamma_{b,i} >> k_{a,i}$, $k_{b,i}$; $\gamma_{a,i}$, $\gamma_{b,i} \sim k_{a,i}$, $k_{b,i}$; $\gamma_{a,i}$, $\gamma_{b,i} << k_{a,i}$, $k_{b,i}$ ($i = 1, 2$). Moreover, in all cases we presume $k_{a,1} = k_{b,1} = k_1$, $k_{a,2} = k_{b,2} = k_2$ (so the forward the backward reaction rates are the same in both conformations) and $\gamma_{a,1} = \gamma_{a,2} = \gamma_{b,1} = \gamma_{b,2}$ (the interconversion between conformations occurs at one rate) for simplicity. While of course this may rarely be the case, the results are not qualitatively sensitive to variations of these parameters in the different limits.

Results

The simplest reaction scheme as shown in Fig. 1a involves a single-step random conversion between the on and off states. Our previous simulation results of such simple first order reactions have demonstrated the single exponential characteristic in the dwell time distribution histogram, and correlation analysis of the on–off blinking trajectory has confirmed the randomness in the reaction kinetics, as assumed in the simulation (Shi et al. 2004). Moreover, a reasonable exponential fit for the dwell time distribution of the single-molecule trajectories can be obtained with a relatively small data set (200 turnovers) to extract the observed rate constant with 10% error.

Dwell time distributions of two-conformation model

The two-conformation model shown in Fig. 1b involves more complex reaction kinetics because the molecule is fluctuating between two conformational states that have distinct reaction rates, which is envisioned as one model of dynamic disorder. Dwell time distribution histograms of the simulated reaction trajectories (3,000 turnovers in total) with different rates of conformational interconversion between reaction rates of $k_{a,1} = k_{b,1} = k_1 = 5\ \text{s}^{-1}$ and $k_{a,2} = k_{b,2} = k_2 = 1\ \text{s}^{-1}$ are plotted in Fig. 2 and fit to a multi-exponential decay, $f(t) = C\sum_i \alpha_i e^{-k_i t}$, where C is the normalization constant, k_i is the observed decay constant of the ith phase and α_i is the corresponding amplitude. When the interconversion rates ($\gamma = 30\ \text{s}^{-1}$) are much faster than the reaction rates ($k_{a,1} = k_{b,1} = k_1 = 5\ \text{s}^{-1}$ and $k_{a,2} = k_{b,2} = k_2 = 1\ \text{s}^{-1}$), the histogram fits to a single exponential decay and the observed rate derived from the dwell time distribution is $3.0 \pm 0.1\ \text{s}^{-1}$, the average of the reaction rates (5 and $1\ \text{s}^{-1}$), as expected (Fig. 2a). When the interconversion rate ($\gamma = 0.3\ \text{s}^{-1}$) is much slower than the reaction rates, the dwell time distribution fits to a bi-exponential decay, and demonstrates the heterogeneity with two phases corresponding to 5.4 ± 0.2 and $1.2 \pm 0.3\ \text{s}^{-1}$ (Fig. 2c, d). The derived rate constants are similar to their theoretical value ($k + \gamma$) of 5.3 and $1.3\ \text{s}^{-1}$. At an intermediate interconversion rate ($\gamma = 3\ \text{s}^{-1}$), results also show two kinetic phases and fit to a bi-exponential decay (Fig. 2b). The rate constants, 9.2 ± 1 and $2.4 \pm 0.2\ \text{s}^{-1}$, derived from the distribution are similar to the theoretical values (Yang and Cao 2001), 9.6 and $2.4\ \text{s}^{-1}$. The error in the fitting results of the on-time histogram

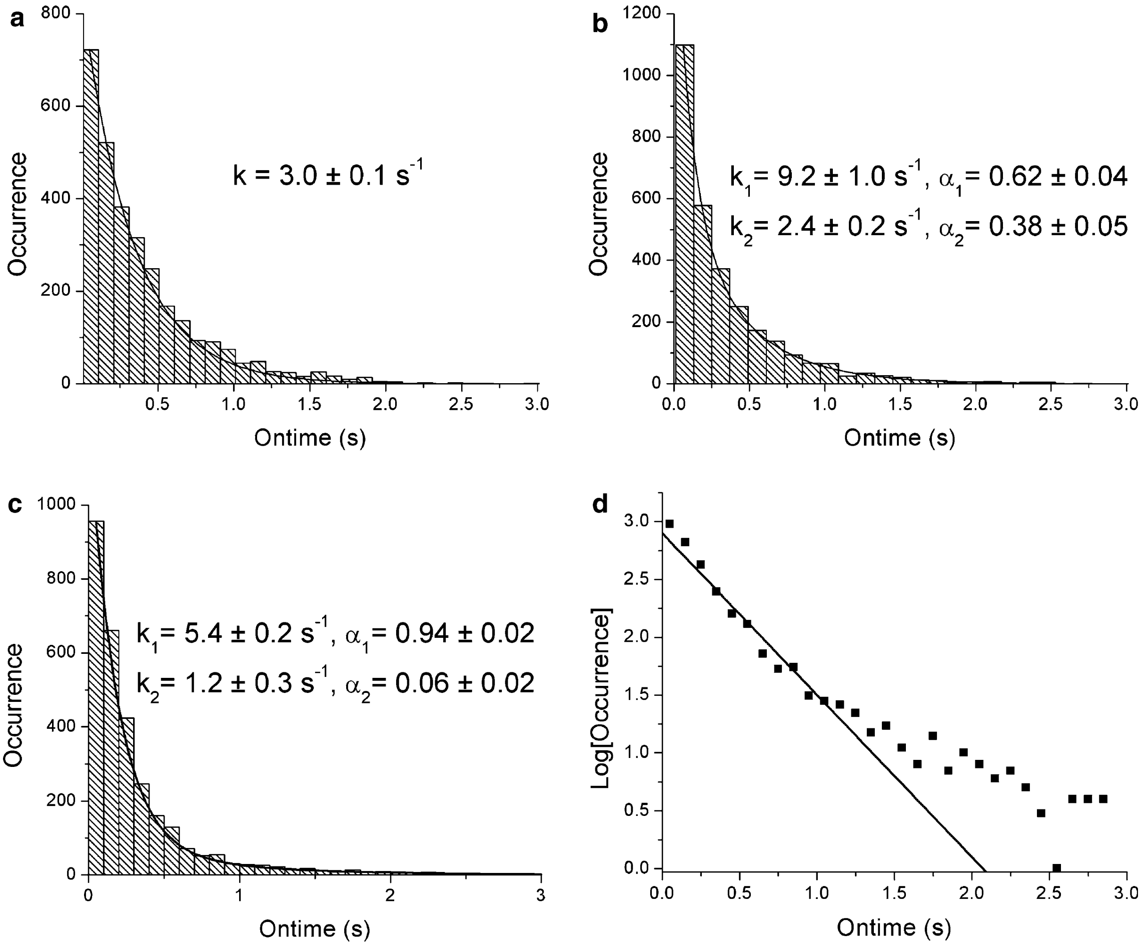

Fig. 2 Histogram of the on-time distributions of a simulated data set of 3,000 turnovers, in which a single molecule interconvertes between reaction rate of 5 and 1 s^{-1}. The histogram is fit to $f(t) = C\sum_i \alpha_i \mathrm{e}^{-k_i t}$, where C is the normalization constant, k_i is the observed decay constant of each phase and α_i is the amplitude. **a** Interconversion rate γ = 30 s^{-1}. As expected, the observed rate constant, 3 s^{-1}, is the average of the two reaction rates. **b** Interconversion rate γ = 3 s^{-1}. **c** Interconversion rate γ = 0.3 s^{-1}. Both **b** and **c** demonstrate heterogeneity with two kinetic phases. **d** On-time distribution of the data shown in (**c**) plotted in log-linear scale reveals the multiexponential nature of the decay more clearly

decreases as the size of the data set increases, approaching the theoretical values.

Our simulation and fitting results show that kinetic parameters of the system can be obtained accurately (within one standard deviation of the theoretical value). Nonetheless, interpreting the extracted kinetic parameters is highly model dependent. Without knowing the reaction model a priori, the bi-exponential feature in the dwell time histogram could be produced by a model based on multiple reaction intermediates or on multiple conformations. Moreover, correlating the observed rate constants to the microscopic kinetic rate constants is in general complex. For our simple two-conformation model, the observed rate constants of the two exponential phases in the dwell time histograms, analytically derived as follow, are already non-linear functions of both the reaction rates and conformation interconversion rates (Yang and Cao 2001).

$$k_{\mathrm{obs}} = \frac{1}{2}(k_1 + k_2) + \gamma \pm \frac{1}{2}\sqrt{(k_1 - k_2)^2 + 4\gamma^2}.$$

It can be anticipated that the observed rate constants for reaction models with more complex conformational dynamics would have more complex dependence on the microscopic kinetic rate constants. Therefore, determination of the microscopic kinetic information and the appropriate reaction model requires evaluation of other statistical parameters to provide additional information, such as calculation of correlation parameters that are discussed below.

In the case of slow conformational modulation (γ = 0.3 s^{-1}) in the two-conformation model, there is a much higher amplitude associated with the conformation with a fast reaction rate (94%) in the dwell time histogram, as compared to the slow rate component (6%). The significant skew towards the fast component is due to the higher order dependence on the rate constants of each conformation as shown by the analytic solution of the dwell time distribution for a two-conformation model. The two exponential components relate to the two distinct conformations and the prefactors of each exponential are determined by complex combinations of the rate constants of both reaction and conformation conversion. When the rate of conformation conversion is much slower than the reaction rates ($\gamma << k_{a,i}$, $k_{b,i}$, $i = 1, 2$, the limiting case is static heterogeneity), the dwell time distribution of on-times can be reduced to a bi-exponential decay given by (Yang and Cao 2001),

$$\begin{aligned} f_a(t) \approx N^{-1}((k_{a,2}+k_{b,2})\cdot k_{a,1}^2 k_{b,1}\mathrm{e}^{-k_{a,1}\cdot t} \\ +(k_{a,1}+k_{b,1})\cdot k_{a,2}^2 k_{b,2}\mathrm{e}^{-k_{a,2}\cdot t}) \end{aligned}$$

where $N = 2\gamma\ (k_{a,1} + k_{b,1})(k_{a,2} + k_{b,2})(k_{a,1}\ k_{b,1}\ (k_{a,2} + k_{b,2}) + k_{a,2}\ \ k_{b,2}\ \ (k_{a,1} + k_{b,1}))$ is the normalization factor.

Therefore, in the limit of small γ, the decay constants derived from the two exponential components approximately equal the reaction rate constants of the two distinct conformations. The prefactors become proportional to $k_{a,i}\ k_{b,i}$ in the case of our simulation when $k_{a,i} = k_{b,i}$. Intuitively, when k_a is large, the probability of the occurrence of an on-to-off reaction event in a given time is large, and when k_b is large, the relative frequency of the molecule residing in the on state is large. Because $k_{a,i}\ k_{b,i}$ is large in the conformation with fast reaction rates, compared to that of the conformation with slow reaction rates, the total number of reaction events observed for the fast reacting conformation would be significantly larger than that observed for the slow conformation, thus accounting for significant amplitude in the on-time distribution. The $k_{a,i}\ k_{b,i}$ dependence of the prefactor in the single-molecule data is clearly different from the linear $k_{a,i}$ (or $k_{b,i}$) dependence derived from ensemble stopped-flow measurements of half-reactions. It significantly skews the dwell time distribution histogram towards the fast components. Plotting the dwell time distribution histogram in log-linear scale (Fig. 2d) improves the visibility of the slow component, but analysis based on the simulation results shows that because of the skewing effect the slow component (1 s^{-1}) can only be detected with a data set larger than 2,500 turnovers.

Auto-correlation and cross-correlation parameters

The effect of dynamic disorder on the reaction kinetics of the two-conformation model can be evaluated by different correlation analyses that were introduced by different research groups. We first calculate a correlation parameter $r(m)$ for on-times or off-times. For single molecule spectroscopy the autocorrelation parameter for pairs of on-times (off-times) has been defined as (Kendall and Ord 1990),

$$r(m) = \frac{\frac{n^2}{n-m}\sum_{i=1}^{n-m} t_i t_{i+m} - \left(\sum_{i=1}^{n} t_i\right)^2}{n\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2}, \tag{1}$$

where t_i is the on-time (off-time) for the ith reaction event, m is the number of reaction events separating the pairs of on-times (off-times) in sequence, and n is the total number of on-times (off-times). Similarly, correlation between consecutive on-times and off-times can be evaluated with a cross-correlation parameter defined as,

$$r_c(m) = \frac{\frac{n^2}{(n-m)}\sum_{i=1}^{n-m} t_i^{\mathrm{on}}\cdot t_{i+m}^{\mathrm{off}} - \left(\sum_{i=1}^{n} t_i^{\mathrm{on}}\right)\left(\sum_{i=1}^{n} t_i^{\mathrm{off}}\right)}{\sqrt{n\sum_{i=1}^{n}\left(t_i^{\mathrm{on}}\right)^2 - \left(\sum_{i=1}^{n} t_i^{\mathrm{on}}\right)^2}\sqrt{n\sum_{i=1}^{n}\left(t_i^{\mathrm{off}}\right)^2 - \left(\sum_{i=1}^{n} t_i^{\mathrm{off}}\right)^2}}, \tag{2}$$

where t_i^{on} and t_i^{off} are on-time and off-time, respectively.

The auto-correlation parameter $r(m)$ for sequential on-times from reaction trajectories simulated with $k_{a,1} = k_{b,1} = k_1 = 5\ \mathrm{s}^{-1}$ and $k_{a,2} = k_{b,2} = k_2 = 1\ \mathrm{s}^{-1}$ is shown in Fig. 3. For interconversion rates $\gamma = 30$ and 3 s^{-1}, there appears to be no observable dynamic correlation in the time scale of the reactions (Fig. 3a, b). For the slower interconversion rate $\gamma = 0.3\ \mathrm{s}^{-1}$ (3.3-fold slower than the slow reaction rate), the apparent decay in the auto-correlation parameter is a clear indicator of the existence of dynamic fluctuations (Fig. 3c). We took the usual approach to fit $r(m)$ to a single exponential and the observed decay constant is 1.1 ± 0.2 turnovers, indicating on-times separated by roughly 1–2 turnovers are correlated. Although there is no analytical form of $r(m)$ to correlate the decay constant

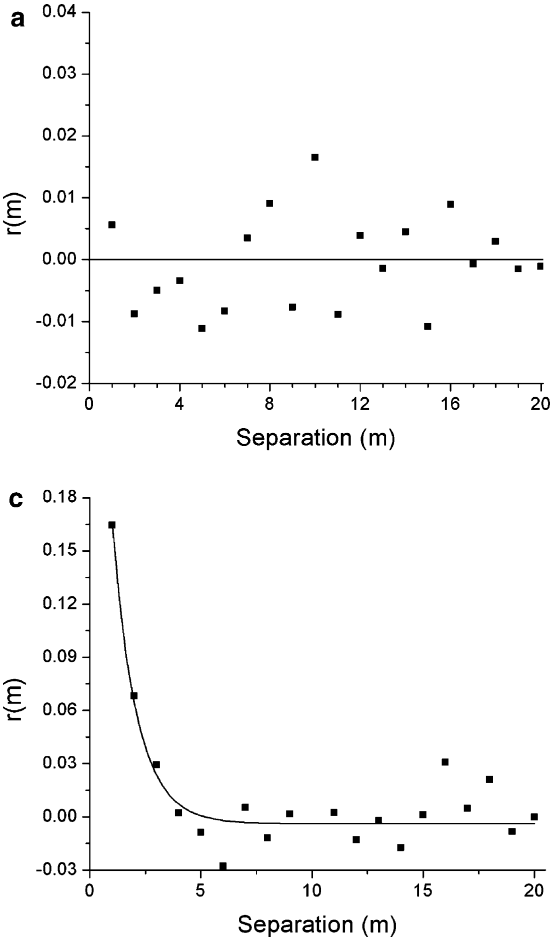

Fig. 3 Auto-correlation function $r(m)$ of on-times with single molecule interconverting between reaction rate of $k_{a,\ 1} = k_{b,\ 1} = 5\ \mathrm{s}^{-1}$ and $k_{a,\ 2} = k_{b,\ 2} = 1\ \mathrm{s}^{-1}$ (3,000 turnovers in total). **a** Interconversion rate $\gamma = 30\ \mathrm{s}^{-1}$. **b** Interconversion rate $\gamma = 3\ \mathrm{s}^{-1}$. **c** Interconversion rate $\gamma = 0.3\ \mathrm{s}^{-1}$. Dynamic correlation is observed in the slowly fluctuating molecules, but not in molecules with fast and intermediate interconversion rates. The decay is fit to an exponential, giving an observed decay half-life of 1.1 ± 0.2 s

to the interconversion rate and reaction rates, it provides insight into time scale of the system retaining its "memory" of the previous reaction steps. $r(m)$ appears to be a better statistical parameter than the dwell time distribution in identifying the existence of multiple reaction channels. This is evident in noting that the correlation feature of decay in $r(m)$ for slow interconverion between two reaction rates is readily observable with a data set as small as 500 turnovers, while the slow component in the dwell time distribution of the same two-conformation model is not detectable for data sets smaller than 2,500 turnovers.

Second-order correlation function of fluorescent states

The dynamic fluctuations can also be quantified by correlating the sequential fluorescent (on) states and non-fluorescent (off) states. The time evolution of the single-molecule fluorescence signal can be described by a state variable, $\xi(t)$, where $\xi(t) = 1$ for fluorescent states and $\xi(t) = 0$ for non-fluorescent states. The second-order correlation function of $\xi(t)$ is given by (Schenter et al. 1999),

$$C_2(\tau) = \frac{\langle \xi(t)\xi(t-\tau)\rangle - \langle \xi(t)\rangle^2}{\langle \xi(t)^2\rangle - \langle \xi(t)\rangle^2}. \tag{3}$$

We calculated $C_2(\tau)$ for the simulated single-molecule data sets shown in Fig. 3. As illustrated in Fig. 4, the distinction between $C_2(\tau)$ calculated for models with dynamic disorder ($\gamma = 0.3\ \mathrm{s}^{-1}$) and without apparent dynamic disorder ($\gamma = 30\ \mathrm{s}^{-1}$) is less evident compared to that of the autocorrelation parameter, $r(m)$. The analytical form of the second-order correlation function of $\xi(t)$ was derived by Schenter et al. (1999), for a two-conformation model with the same kinetic parameters we chose for our simulation,

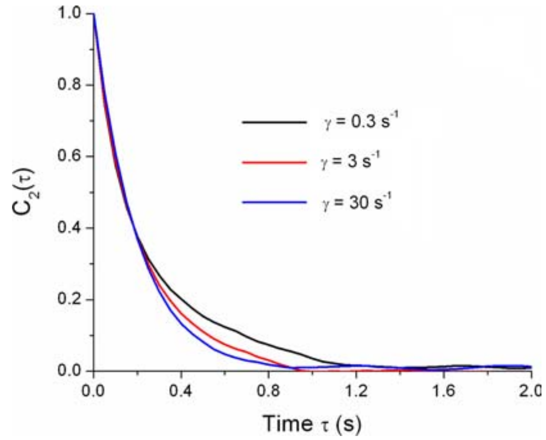

Fig. 4 Second order correlation function, $C_2(\tau)$, of the fluorescent state variable, $\xi(t)$. Single-molecule data sets were obtained with reaction rate of $k_{a,\ 1} = k_{b,\ 1} = 5\ \mathrm{s}^{-1}$, $k_{a,\ 2} = k_{b,\ 2} = 1\ \mathrm{s}^{-1}$ (3,000 turnovers in total), and interconversion rate γ = 30, 3, and 0.3 s^{-1}, respectively (color figure available online)

$$C_2(\tau) = \frac{\mathrm{e}^{-k_+ t}}{2\beta}[\gamma(\mathrm{e}^{2\beta t} - 1) + \beta(\mathrm{e}^{2\beta t} - 1)],$$

where $k_+ = k_1 + k_2 + \gamma + \beta$, and $\beta = \sqrt{\gamma^2 + (k_1 - k_2)^2}$. Therefore, theoretically the conformation interconversion rate can be quantified under the two-conformation model by fitting the experimentally obtained $C_2(\tau)$ to the theoretical function, assuming the reaction rate constants, k_1 and k_2, are derived from other statistical analyses, such as dwell time distribution histogram. However, such quantification is again model specific. Given the subtle difference in $C_2(\tau)$ between systems with and without dynamic disorder as shown in Fig. 4 and its complex dependence on interconversion rates and reaction rates, it is in general difficult to employ $C_2(\tau)$ to identify and quantify dynamic fluctuations in reaction rates under an unknown model.

Additional correlation functions

In the theoretical studies of single-molecule kinetics by Cao et al. (Cao 2000; Yang and Cao 2001, 2002) and Mukamel et al. (Barsegov et al. 2002; Barsegov and Mukamel 2002), three additional correlation functions were employed as signatures to infer dynamic disorder. One is the joint distribution function $f(t_1,\ t_2)$ for sequential dwell times [an on-time (off-time) of t_1 followed by an on-time (off-time) of t_2]. The second one is the difference function $\delta(t_1,\ t_2)$ defined as $\delta(t_1,\ t_2) = f(t_1,\ t_2) - f(t_1)f'(t_2)$, where $f(t)$ and $f'(t)$ are the probability distribution functions of on-times and off-times and $f(t_1,\ t_2)$ is the joint distribution function of pairs of on-times and off-times. The third one is the same-time difference function $\delta(t,\ t)$, which is the diagonal value of $\delta(t_1,\ t_2)$.

Given the single-molecule reaction trajectories, $f(t_1,\ t_2)$ is determined by counting the occurrence of pairs of dwell times that are t_1 and t_2, respectively. If the reaction events are not correlated, the joint distribution is given by $f(t_1,\ t_2) = f(t_1)f'(t_2)$, where $f(t)$ and $f'(t)$ is the probability function of dwell times, i.e. on-times or off-times. The overall characteristic of the joint distribution function $f(t_1,\ t_2)$ is a two-dimensional exponential decay along the coordinates of t_1 and t_2. As predicted by the theoretical calculations, when the rate of conformation interconversion is slow, as resulting from a slow environmental modulation, dynamic correlation in a two-conformation model appears in the 2D distribution plot of $f(t_1,\ t_2)$ as an initial slow decay, relative to that demonstrated in a one-conformation model, and a long tail (Barsegov et al. 2002; Barsegov and Mukamel 2002). Such characteristics disappear as the interconversion rate increases. In addition, increasing the size of the data set does not appear to bring up new correlation patterns.

The difference function, $\delta(t_1,\ t_2) = f(t_1,\ t_2) - f(t_1)f'(t_2)$, is an alternative means to identify the existence of a correlation between the consecutive on-times and off-times because when there is no dynamic correlation, $f(t_1,\ t_2) = f(t_1)f'(t_2)$ so $\delta(t_1,\ t_2) = 0$. The difference function is conceivably more sensitive in detecting dynamic correlation than the joint distribution function because it subtracts the contribution of the uncorrelated response, $f(t_1)f'(t_2)$, from the joint distribution function.

Indeed the 2D plot of the difference function $\delta(t_1, t_2)$ for sequential on-times derived from the simulated data of the two-conformation model with slow interconversion rate γ = 0.3 s^{-1} shows an interesting diagonal feature for data set larger than 20,000 turnovers, signifying the existence of dynamic fluctuation in the reaction rates, where the on-times (off-times) show a dependence on the previous on-times (off-times) (Fig. 5). Unfortunately, the diagonal correlation signature is not readily identifiable with small data sets (< 20,000 turnovers), which are typical in most of the single-molecule experiments due to photodestruction of the fluorophore. A similar diagonal feature was reported for cholesterol oxidase (Lu et al. 1998), though in that case it was observed in the joint distribution function, $f(t_1,\ t_2)$. Interestingly, the 2D plot of just the joint distribution function for the data plotted in Fig. 5 reveals no obvious diagonal structure, even when the effect of static heterogeneity is included in the reaction model. This suggests the model for cholesterol oxidase

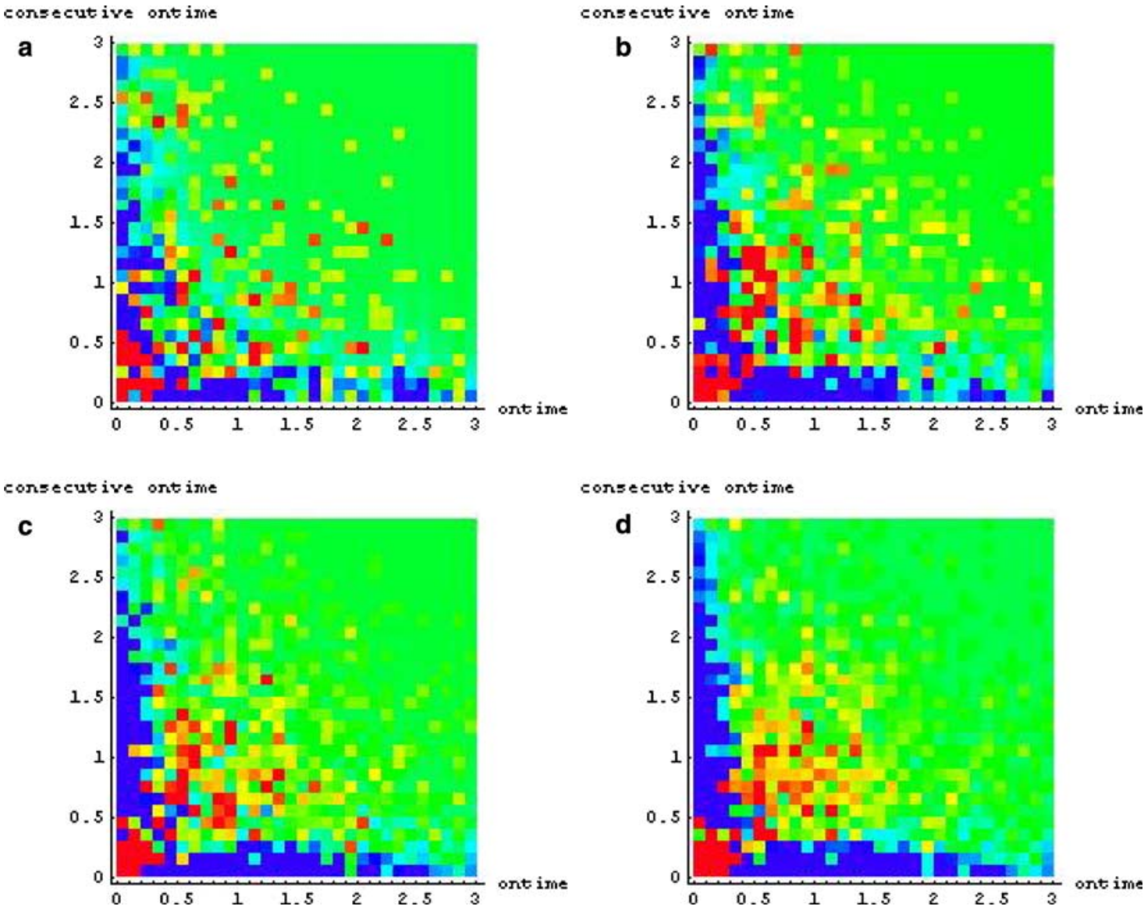

Fig. 5 2D plot of the difference function $\delta(t_1, t_2)$ for pairs of sequential on-times for a molecule with interconversion rate of 0.3 s^{-1}. The value of $\delta(t_1, t_2)$ is shown in a color scale utilizing the Hue function in Mathematica. The value of $\delta(t_1, t_2)$ decreases from color *red* to *blue*. The diagonal feature is more readily identifiable when the size of the data set increases: **a** 3,000 turnovers, **b** 10,000 turnovers, **c** 20,000 turnovers, **d** 40,000 turnovers (color figure available online)

catalysis is more complex than that considered in Fig. 1b.

Cao (Cao 2000; Yang and Cao 2001) suggested that the time evolution of the same-time difference function $\delta(t, t)$, which is the diagonal value of $\delta(t_1, t_2)$ when $t_1 = t_2 = t$, provides another signature of dynamic conformational fluctuation. Theoretical analysis of $\delta(t, t)$ showed that in addition to the initial maximum at $t = 0$, $\delta(t_1, t_2)$ has a small echo peak at time t_e when $\mathrm{d}\delta(\mathrm{t}, \mathrm{t})/\mathrm{d}t = 0$. Cao formulated that the echo time t_e in the same-time difference function directly measures the conformational fluctuation rate and the amplitude of the echo probes the variance of the reaction rates (Yang and Cao 2001). To compare the simulation results with the theoretical results explicitly given by Cao (2000), we calculate the same-time difference function $\delta(t, t)$ for pairs of on-times and consecutive off-times with the simulation data set. As shown in Fig. 6a, the fluctuation of $\delta(t, t)$ as indicated by the error bar is large compared to its theoretical value, thus, within the context of typical experimental data, it would be challenging to infer the echo peak. The correlation becomes more distinct when the size of the single-molecule data set is significantly increased (Fig. 6). However, with a data set consisting of 40,000 turnovers, our results still do not show a clear signature of correlation echo, The reason for the need of a very large data set to infer the correlation features from $\delta(t, t)$ is probably due to fact that $\delta(t, t)$ only involves the diagonal values of $\delta(t_1, t_2)$, neglecting most of the available data. In practice it is difficult to obtain such large data sets because of photobleaching or fluorophore dissociation. Therefore, it might be challenging to quantify conformational fluctuations with the same-time difference function $\delta(t, t)$ derived from experimental data, though in theory it provides important information about the process.

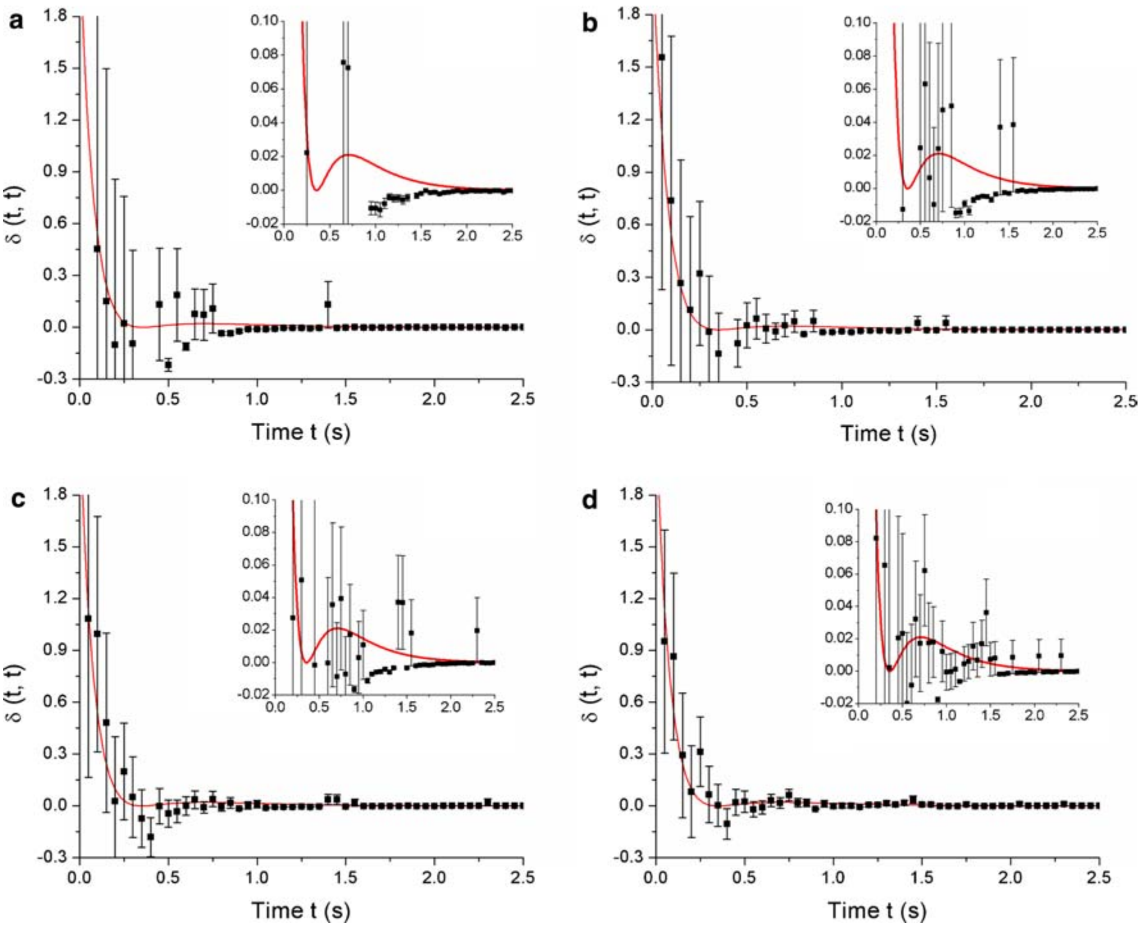

Fig. 6 Same-time difference function $\delta(t, t)$ for pairs of on–off events for a molecule with interconversion rate of 0.3 s^{-1}. The *red lines* denote the theoretical value of $\delta(t, t)$ derived according to Reference (Sakmann and Neher 1995). A smaller scale (between −0.02 and 0.1 s) is used in the insets in order to show the small echo peak. Significant fluctuation of the data makes it difficult to resolve the echo peak in data set with **a** 3,000 turnovers and **b** 10,000 turnovers. The echo peak is more readily identifiable in data set with **c** 20,000 turnovers and **d** 40,000 turnovers (color figure available online)

Discussion and conclusion

Our simulation analysis examines some of the key questions in the laboratory regarding extracting kinetic information from single-molecule data, such as how big a data set is needed and the level of confidence in the statistical analyses. The ability to resolve among specific reaction models depends on the amount of available data, the noise in the system (not addressed in this paper), and the differences between various fitting parameters. For example, we have shown that in order to distinguish the two-conformation model shown in Fig. 1b from the single conformation model (Fig. 1a) by dwell time distribution, the data set has to consist of over 2,500 turnovers under the parameters we simulated. In addition, if the difference between the reaction rates of the distinct conformations as well as the difference between the reaction rate and conformation interconversion rate are not significant, the presence of multiple conformations would not be identifiable through statistical analysis of typical single-molecule data. Simulation results of various reaction rates and conformation interconversion rates (data not shown) revealed that $r(m)$, the correlation parameter of dwell times that is shown to be most sensitive in detecting dynamic disorder, only demonstrates a pronounced decay feature when $k_{a,1} \geq 3k_{a,2}$ and $k_{a,2} > 2\gamma$ in our two-conformation model. Therefore, at least a threefold difference between the kinetic parameters is necessary to characterize the existence of multiple conformations with different reactivity.

In practice, the amount of available data derived from a single-molecule experiment, i.e. the number of turnovers observed for a single fluorophore, is limited by photobleaching or by fluorophore dissociation from the protein, as in our study of flavoprotein. Given a fixed number of total available photons (accounting for detection efficiency and photobleaching), N_0, the number of fluorescence on–off cycles of a single-fluorophore can be roughly estimated as follow. For the simple two-state model defined in Fig. 1a with forward and backward reaction rates of k_a and k_b, respectively, let k_f be the faster of the two. The size of the time bin (timing resolution) can be selected as $\Delta t = \chi \, k_f^{-1}$, where $0 \leq \chi \leq 1$ is the fraction of the fastest transition time, k_f^{-1}. The photon detection rate is then $\varepsilon = N/\Delta t = Nk_f/\chi$, where N is the number of photons detected in a time bin and is determined by the desired signal to noise ration $\eta = N/\sqrt{N} = \sqrt{N}$. On average, the typical duration of the fluorescence on-state is k_a^{-1}. Hence, the average number of photons detected in a cycle between the fluorescence on- and off-state is given by $\bar{N}_{\text{det}} = k_a^{-1}\varepsilon = \chi^{-1}N\left(k_f/k_a\right)$. This then gives the total number of cycles observed in the single-molecule measurement, $N_{\text{tot.cycles}} = \frac{N_0}{\bar{N}_{\text{det}}} = \chi\left(\frac{k_a}{k_f}\right)\left(\frac{N_0}{\eta^2}\right)$. For a simple case, assuming $k_f = k_a$, $\chi = 0.1$ for a time bin of 10% that of the inverse forward rate constant, a signal-to-noise ratio of $\eta = 3.3$ ($N = 10$), and a total number of $N_0 = 10^5$ detected photons (limited for example by photobleaching, detection efficiency, etc), it can be estimated that $N_{\text{tot.cycles}} = 10^3$. One can optimize the experimental setup and the rate of data acquisition to maximize the number of observed data, as discussed by Shi et al. (2006).

The above estimation indicates that it may be possible to obtain a relatively large single-molecule data set for photostable chromophore covalently attached to biomolecules in study (Ha et al. 1999; Zhuang et al. 2002). However, for systems with a non-covalently bound chromophore like the flavoproteins, the dissociation rate is too fast to allow observation of the chromophore for more than a relatively small number of turnovers (Shi et al. 2004), resulting in single-molecule data set of limited size.

Static heterogeneity is another interesting feature that can be probed by single-molecule experiment, although this simulation study has focused on the practicality and viability of obtaining information of dynamic disorder involved in the system by different statistical analyses. Note that static heterogeneity is not visible from analysis of individual reaction traces because such heterogeneity exemplifies in the difference between molecules. Therefore, one needs to compile a data set from a number of different molecules and then compare the results for each molecule to detect variation in the population. In our experimental study of DHOD, we employed distribution of averaged reaction rate of each molecule and distribution of individual on-times to display and quantify static heterogeneity in DHOD catalysis (Shi et al. 2004). It is possible that dynamic heterogeneity would show up as static heterogeneity in single-molecule analysis if the time scale for conversion from one conformation to another is long compared to the time duration of the individual data set. In such case the individual reaction trajectory is relatively short so it may not comprise turnover events in all of the conformational states. Hence, it is possible to confuse static and dynamic heterogeneity with a small single-molecule data set.

Dynamic disorder can be potentially quantified by the characteristic decay time of the correlation parameter $r(m)$, the correlation function $C_2(\tau)$, the difference function $\delta(t_1, t_2)$, and the echo peak of the same time difference function $\delta(t, t)$. From our simulation results, it can be inferred that under the limit of the parameters we simulated, if no kinetic information is known a priori, $r(m)$ is the most sensitive parameter to identify the existence of dynamic disorder and also quantify the relaxation time scale of such dynamic disorder, although the exact relation between the observed decay rate of $r(m)$ and the interconversion rate between conformations is complex and not easily derivable in general. The small difference in $C_2(\tau)$ between a random system and systems with dynamic disorder makes it difficult to use $C_2(\tau)$ to characterize dynamic disorder. The small amplitude of the echo peak of $\delta(t, t)$ poses similar difficulty in practice. We have also used a 2D plot in the paper (Fig. 5) to display dynamic correlation revealed by the difference function. Although the 2D plot conveniently demonstrates correlation features visually, it is difficult to quantify the diagonal pattern because of the complexity in deriving the analytic form for the difference function. So in general the 2D plot is employed only to display and visualize dynamic correlation qualitatively, instead of displaying quantitative information.

A variety of chemical reactions and biological processes exhibit the characteristics of first-order kinetics assumed in our simulation. Therefore, it is reasonable to apply our simulation results to a wide range of single-molecule experimental data sets and this approach should help improve the understanding of more complex reaction kinetics. Moreover, we anticipate that our results and conclusions concerning the characteristics of the correlation dynamics remain valid for other more complex reaction models with multiple reaction steps, such as enzymatic reaction involving substrate binding and product dissociation, as long as all the individual steps are stochastic. For reactions that involve multiple intermediate steps, the dwell time distribution of the single-molecule trajectories would exhibit a multi-exponential characteristic. And the decay rate constants obtained from the distribution are a function of the kinetic rates of all individual steps. If all the reaction steps are stochastic, the waiting times that the molecule resides in each intermediate state are still random variables so the overall dwell time as the sum of random variables remains random.

In summary, a simulation study that generates single-molecule data sets based on different hypothesized models under experimental limitations is an effective measure to parallel statistical analysis of the experimental data to determine the appropriate reaction model. In addition, in the presence of dynamic disorder our simulation and analytical results show that the dwell time distribution histogram of single-molecule reaction trajectories is significantly skewed towards the fast reaction states due to the higher order dependence of the exponential prefactors on the reaction rates of the different conformational states. Moreover, our correlation analyses suggest that the auto-correlation parameter $r(m)$ is the most straightforward function for quantifying dynamic disorder arising from fluctuations between multiple conformational states and can identify multiple reaction channels with a relatively small data set. The correlation function of fluorescent state, joint distribution function and difference distribution function are theoretically more informative; however, considering their subtle characteristics and relatively small amplitudes derived from even data simulated in ideal conditions with a large number of turnovers and no noise, it appears difficult to employ them for probing dynamic disorder with experimental single-molecule data of limited size.

Acknowledgements This work is supported by the National Institutes on Aging (5 R01 AG018017-04), and Michigan Life Science Corridor (MLSC, proposal#1537) to A.G. and D.S. The authors also would like to acknowledge helpful discussion with X. Sunney Xie. The simulation source code written in Mathematica is available for download in our group website at http://www.umich.edu/~protein/monte.html.

References

Ariga T, Masaike T, Noji H, Yoshida M (2002) Stepping rotation of F(1)-ATPase with one, two, or three altered catalytic sites that bind ATP only slowly. J Biol Chem 277(28):24870–24874

Barsegov V, Mukamel S (2002) Probing single molecule kinetics by photon arrival trajectories. J Chem Phys 116(22):9802–9810

Barsegov V, Chernyak V, Mukamel S (2002) Multitime correlation function for single molecule kinetics with fluctuating bottlenecks. J Chem Phys 116(10):4240–4251

Bianco PR, Brewer LR, Corzett M, Balhorn R, Yeh Y, Kowalczykowski SC, Baskin RJ (2001) Processive translocation and DNA unwinding by individual RecBCD enzyme molecules. Nature 409:374–378

Boguna M, Berezhkovskii AM, Weiss GH (2000) Residue time densities for non-Markovian systems.1. The two-state system. Physica A 282:475–485

Brown FL (2003) Single-molecule kinetics with time-dependent rates: a generating function approach. Phys Rev Lett 90(2):028302

Cao J (2000) Event-averaged measurement of single-molecule kinetics. Chem Phys Lett 327:38–44

Chakrabarti D, Bagchi B (2003) Waiting time distribution and nonexponential relaxation in single molecule spectroscopic studies: realization of entropic bottleneck in a simple model. J Chem Phys 118(17):7965–7972

Edman L, Rigler R (2000) Memory landscapes of single-enzyme molecules. Proc Natl Acad Sci USA 97(15):8266–8271

Forkey JN, Quinlan ME, Shaw MA, Corrie JE, Goldman YE (2003) Three-dimensional structural dynamics of myosin V by single-molecule fluorescence polarization. Nature 422:399–404

Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J Phys Chem 81(25):2340–2361

Ha T, Zhuang X, Kim H, Orr JW, Williamson JR, Chu S (1999) Ligand-induced conformational changes observed in single RNA molecules. Proc Natl Acad Sci USA 96:9077–9082

Ha T, Rasnik I, Cheng W, Babcock HP, Gauss GH, Lohman TM, Chu S (2002) Initiation and re-initiation of DNA unwinding by the *Escherichia coli* Rep helicase. Nature 419:638–641

Ishijima A, Doi T, Sakurada K, Yanagida T (1991) Sub-piconewton force fluctuations of actomyosin in vitro. Nature 352:301–306

Kendall M, Ord JK (1990) Chap. 6 in time series. Hodder and Stoughton Educational, Kent

Kleinekathofer U, Barvik I, Herman P, Kondov I, Schreiber M (2003) Memory effects in the fluorescence depolarization dynamics studied within the B850 ring of purple bacteria. J Phys Chem B 107(50):14094–14102

Lerch H-P, Mikhailov AS, Hess B (2002) Conformational-relaxation models of single enzyme kinetics. Proc Natl Acad Sci USA 99(24):15410–15415

Lu HP, Xun L, Xie XS (1998) Single-molecule enzymatic dynamics. Science 282:1877–1882

Oijen AM, Blainey PC, Crampton DJ, Richardson CC, Ellenberger T, Xie XS (2003) Single-molecule kinetics of λ exonuclease reveal base dependence and dynamic disorder. Science 301:1235–1238

Pease PJ, Levy O, Cost GJ, Gore J, Ptacin JL, Sherratt D, Bustamante C, Cozzarelli NR (2005) Sequence-directed DNA translocation by purified FtsK. Science 307:586–590

Sakmann B, Neher E (eds) (1995) Single-channel recording. Plenum Press, New York

Schenter GK, Lu HP, Xie XS (1999) Statistical analyses and theoretical models of single-molecule enzymatic dynamics. J Phys Chem A 103:10477–10488

Schuler B, Lipman EA, Eaton WA (2002) Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. Nature 419:743–747

Shi J, Palfey B, Dertouzos J, Jensen KF, Gafni A, Steel D (2004) Multiple states of the Tyr318Leu mutant of dihydroorotate dehydrogenase revealed by single molecule kinetic. J Am Chem Soc 126(22):6914–6922

Shi J, Gafni A, Steel D (2006) Application of single molecule spectroscopy in studying enzyme kinetics and mechanism. Methods Enzymol (in press)

Vlad MO, Moran F, Schneider FM, Ross J (2002) Memory effects and oscillations in single-molecule kinetics. Proc Natl Acad Sci USA 99(20):12548–12555

Wang J, Wolynes P (1995) Intermittency of single molecule reaction dynamics in fluctuating environments. Phys Rev Lett 74(21):4317–4320

Witkoskie JB, Cao J (2004) Single molecule kinetics I theoretical analysis of indicators. J Chem Phys 121(33):6361–6372

Yang S, Cao J (2001) Two-event echos in single-molecule kinetics: a signature of conformational fluctuations. J Phys Chem 105:6536–6549

Yang S, Cao J (2002) Direct measurement of memory effects in single-molecule kinetics. J Chem Phys 117(24):10996–11009

Yang H, Luo G, Karnchanaphanurach P, Louie T, Rech I, Cova S, Xun L, Xie XS (2003) Protein conformational dynamics probed by single-molecule electron transfer. Science 302:262–266

Zhuang X, Kim H, Pereira M, Babcock H, Walter N, Chu S (2002) Correlating structural dynamics and function in single ribozyme molecules. Science 296:1473–1476

Zwanzig R (1990) Rate processes with dynamical disorder. Acc Chem Res 23:148–152