Testing for Renewal and Detailed Balance Violations in Single-Molecule Blinking Processes†

James B. Witkoskie and Jianshu Cao*

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

Received: March 9, 2006; In Final Form: July 28, 2006

This paper examines methods to test one- and two-dimensional histograms for several features including the renewal properties, detailed balance violations, and experimental condition dependences. The tests are simple to implement and allow rigorous statistical determination of the existence of these kinetic features. The tests are used to determine the lower bound on the number of measurements necessary to differentiate underlying kinetic models.

I. Introduction

Single-molecule experiments offer a window into the microscopic world of chemical systems that allows the observation of individual chemical *events*.[1] These techniques have been used to explore systems from the oxidation−reduction of a single cofactor in an enzyme to the folding of ribozymes.[2] Three of the most popular single-molecule experiments, fluorescence resonance energy transfer (FRET), simple fluorescence blinking, and fluorescence quenching often show the molecules hopping between discrete values for either the FRET efficiency or the fluorescence yield.[3,4] We define the collection of configurations that have a degenerate single-molecule value as a manifold. The hopping processes give rise to the event statistics that examine the sojourn time within a certain manifold before making a transition to another manifold.[5,6] Many models for event statistics contain hidden substates that are degenerate with respect to the probe (FRET, fluorescence, etc.) but have different sojourn time distributions.[7] Determining both the topology and the parameters of a model with hidden substates becomes cumbersome.[8,9] As an example, there are over 20 000 topologies for connecting eight substates in a linear chain. Although much effort has been dedicated to efficient search algorithms and reducing redundancies, gaining physical insight by examining features that appear in coarse grained measures can restrict possible models and even identify properties of the system without model optimization.

In this paper, we examine statistical tests of properties that appear in the one- and two-dimensional histograms of events but do not require the construction of an underlying model. These tests are derived from the Bayesian analysis and information theory, which has been implemented by several groups on single photon counting experiments.[10−12] While earlier approaches mainly concerned determination of the transition times between different states and the underlying models to describe the transitions,[13,14] our work concerns inferring properties of the system based on the transition times. The tests capture important aspects of the system, including the non-Markovian/nonrenewal nature of the system, violations of detailed balance, and similarity in behaviors of molecules in different experimental conditions.[15] The renewal behavior determines the existence of parallel paths, while the detailed balance violations give insight into circulation in the underlying topology, including the topology of the circulation loop. Concentration dependence indicates the role of cofactors in a single-molecule function, such as metal ions in ribozyme folding and the energy transfer between the substrate and the macromolecule.[15,16]

The major difficulty with the two-event waiting time distribution is the need to create a histogram for the events (resulting in a histogram h_{ij}). In the infinite data and infinitesimal bin size limit, the traditional analysis discussed in the single-molecule literature will suffice. However, these methods are sensitive to noise and cannot be easily implemented for finite sets of data. The implementation difficulties have several sources. One source is the binning methods, which are linear in many proposed applications but should be logarithmic for examining exponentially distributed data (Appendix).[17] In this paper, data will be presented in logarithmic histograms, but the discussion will present linear probability distributions. Another issue is the scatter that is present in the histogram. The scatter in a histogram should be approximately Poisson-distributed with the variance in the number of events in a bin being approximately equal to the number of events in the bin. As a result, the scatter will not be uniformly distributed throughout the histogram, which can cause misinterpretation in features that are measured from the differences between histograms. The nearly Poisson nature of the data scatter causes an additional problem for low-count bins, where the deviations are not Gaussian. Avoiding these difficulties is the motivation behind introducing more rigorous statistical methods of assessing the relevance of apparent features. An important result is the establishment of the number of measurements necessary to elucidate the existence of features in the histogram properties and distinguish different models. These numbers should be used as a guideline for the types of models to compare to the data.

II. Poisson Kinetic Models

Although generalizations are simple, we concentrate on the two-manifold model, where the system exhibits hops between two intensities, labeled + for *bright* and − for *dark*, resulting in a blinking sequence. The durations of the + or − intensity have been referred to as events in the literature.[7] These hopping events are generally modeled by stochastic waiting time processes (semi-Markov).[18] If the waiting times show correlations between adjacent transition times, then additional hidden states are added. Including only a few states can greatly increase

† Part of the special issue "Robert J. Silbey Festschrift".

* Author to whom correspondence should be addressed. E-mail: jianshu@mit.edu.

10.1021/jp061471w CCC: $33.50
Published on Web 08/30/2006

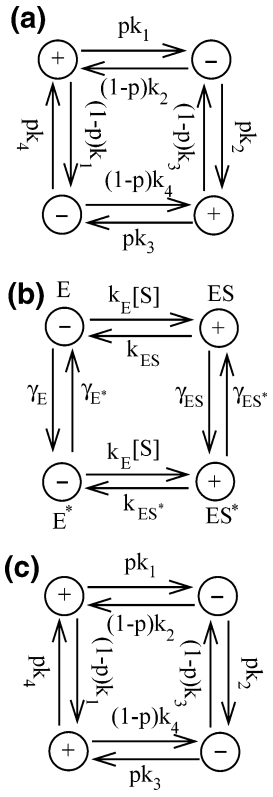

Figure 1. Kinetic schemes that violate detailed balance. (a) A scheme that has a circulation loop passing through both manifolds twice, which gives time reversibility and diagonal dominance violations. (b) A concentration-dependent Michaelis−Menten scheme, where the substrate pumps the conformational coordinates of the system. (c) A kinetic scheme that has a circulation loop resulting in a peak in the single waiting time distributions.

the space complexity of possible models, so it is desirable to develop methods to obtain insights that do not require model optimization.[8] The kinetics for this model are described by the rate equation

$$\partial_t \begin{bmatrix} \rho_+ \\ \rho_- \end{bmatrix} = - \begin{bmatrix} K_{++} + \Gamma_{++} & -K_{+-} \\ -K_{-+} & K_{--} + \Gamma_{--} \end{bmatrix} \begin{bmatrix} \rho_+ \\ \rho_- \end{bmatrix} \quad (1)$$

An event is the sojourn time in a particular manifold. For a given kinetic scheme the probability of n events starting from $\pm$ is is given by

$$P_{\pm\ldots\pm}(t_1, \ldots, t_n) = \sum \left[\prod_i K_{\pm\mp} \, \mathrm{e}^{-(K_{\mp\mp}+\Gamma_{\mp\mp})t_i} \right] \frac{K_{\mp}(\rho_{\mathrm{eq}})_{\mp}}{\sum K_{\mp}(\rho_{\mathrm{eq}})_{\mp}} \quad (2)$$

where the sum is over the final state. The distribution of the log times that is shown in the figures is derived by replacing t_i with e^{lnt_i} and multiplying by e^{lnt_i}. To obtain reduced probability matrices one integrates out the unwanted degrees of freedom. The results are general, but we will restrict our discussion to the four substate models shown in Figure 1. These models have been studied in the literature and applied to several single-molecule experiments, including fluorescence blinking and ion channel experiments.[19,20]

III. Testing Models for the Histogram Data

To avoid the combinatorially complex problem of finding the best model for the data, we construct histograms of events from the data and then test generic features of the histograms instead of the underlying model that produced these features. The tests are performed within a Bayesian framework, by determining the likelihood that the data are produced by a model of the histogram that contains certain properties. The Bayesian methods outlined in this section are standard, but many of the tested features have not been explored within the Bayesian framework or at all. The non-Bayesian tests applied to determine these features are known to be sensitive to fluctuations in the data and to be reliable.

The features of a histogram are tested by constructing models of the histograms and comparing these models to the data. The original histograms, h_i or h_{ij}, include the single events (single sojourn time) and pairs of events, such as adjacent + and − sojourn times or two + sojourn times separated by one − sojourn. The objective is to test possible models for the histograms. The best fitting model will always be the histogram itself $P_{ij} = (1/N)h_{ij}$ with $N = \sum_{ij} h_{ij}$. The main issue is the establishment of other simpler models $\tilde{P}_{ij}$ that are consistent with the data. For a model $\tilde{P}_{ij}$, the probability of obtaining the histogram h_{ij} is $P(h_{ij}|\,\tilde{P}_{ij}) = \prod_{ij}(\tilde{P}_{ij})^{h_{ij}}$. Taking the log of this probability gives

$$\ln P(h_{ij}|\tilde{P}_{ij}) = \sum_{ij} h_{ij} \ln(\tilde{P}_{ij}) \quad (3)$$

Since each event is a random variable, $\sum_{ij} h_{ij} \ln(\tilde{P}_{ij})$ is the result of a sum of random variables that converges to a Gaussian distribution in the large N limit. As a result, the difference in the log probability of the histogram, $P_{ij} = (1/N)h_{ij}$, and $\tilde{P}_{ij}$ is a natural method of comparing the model to the data and assessing if the model is adequate.[21] For the histogram h_{ij}, the difference in the logarithms is

$$\sum_{ij} h_{ij}[\ln(P_{ij}) - \ln(\tilde{P}_{ij})] = \sum_{ij} h_{ij} \ln(P_{ij}/\tilde{P}_{ij}) \quad (4)$$

Since $P_{ij} = (1/N)h_{ij}$, we are left with N times the Kullback−Liebler metric (KL)[22,23]

$$NI_{P|\tilde{P}} = N \sum_{ij} (\delta I_{P|\tilde{P}})_{ij} = N \sum_{ij} P_{ij} \ln(P_{ij}/\tilde{P}_{ij}) \quad (5)$$

The log likelihoods should be compared against the expected variances

$$N\delta I_{P|\tilde{P}}{}^2 = N\left[\sum_{ij} P_{ij} \ln(P_{ij})^2 - \left(\sum_{ij} P_{ij} \ln(P_{ij}) \right)^2 + \sum_{ij} P_{ij} \ln(\tilde{P}_{ij})^2 - \left(\sum_{ij} P_{ij} \ln(\tilde{P}_{ij}) \right)^2 \right] \quad (6)$$

If the KL metric is small compared to the variance estimate, then $\tilde{P}$ is an adequate model for the data. It is simple to account for correlations in the data by modifying the variance. This testing method penalizes using the histogram P_{ij} itself by a factor that scales as $\sqrt{N}$ without regard to the number of parameters, whereas other methods penalize by factors of $\ln(N)$ or unity with a parameter-dependent prefactor.[24] The preferred method should depend on both the number of data points and the number of parameters.

If $P_{ij} \approx \tilde{P}_{ij}$, then a Taylor expansion gives

$$I_{ij} \approx \sum_{ij} \frac{1}{2} \frac{(P_{ij} - \tilde{P}_{ij})^2}{P_{ij}} = \sum_{ij} \frac{1}{2} \frac{\Delta P_{ij}{}^2}{P_{ij}} \quad (7)$$

since the linear term averages to zero. This result follows from the previous discussion about the approximately Poisson nature

of the variations in the data. The variance in the number of events in each bin is equal to the number of events in a bin ($\sigma_{ij}^2 = NP_{ij}$), and this procedure reduces to least-squares analysis in the large data limit.[17] The expression demonstrates that the relative instead of absolute values of deviations of $\tilde{P}_{ij}$ from P_{ij} are the important quantities. Large absolute deviations appear in regions with larger numbers of events, and reweighting the residual deviations is necessary.

IV. Testing Renewal Behavior

As the simplest example of testing models, consider trying to determine if the histogram h_{ij} corresponds to a simple renewal or alternating renewal process. Renewal or alternating renewal processes assume that the sojourn times of events are independent, which implies that the process does not exhibit multiple paths connecting the + and − manifolds.[25,26] To test the renewal property, the data, $P_{ij} = (1/N)h_{ij}$, must be compared with the best fitting model for independent events

$$\tilde{P}_{ij} = P_i P_j = \frac{1}{N^2}\left[\sum_j h_{ij}\right]\left[\sum_i h_{ij}\right] \tag{8}$$

The resulting KL metric for comparing P and $\tilde{P}$ is $I_{P|\tilde{P}} = \sum_{ij} P_{ij} \ln(P_{ij}/(P_i P_j))$, which is sometimes called the mutual information between the variable i and j.[27] The mutual information is always positive since P_{ij} is a better fit. To determine if the difference in fits is significant, we compare $NI_{P|\tilde{P}}$ with $N\delta I^2_{P|\tilde{P}}$. If zero falls within the 95% confidence interval (2 standard deviations), then the events are considered independent, and we adopt $\tilde{P}_{ij} = P_i P_j$.

We explore application of this test to the model in Figure 1a as a function of p and K for $k_1 = k_2 = K^{-1}k_3 = K^{-1}k_4 = 1$. In this case the two-event function for a + sojourn followed by a − sojourn is

$$P_{+-}(t_1, t_2) = \frac{1}{2} e^{-t_1}[p\, e^{-t_2} + (1-p)K\, e^{-Kt_2}] + \frac{1}{2} K\, e^{-Kt_1}[pK\, e^{-Kt_2} + (1-p)e^{-t_2}] \tag{9}$$

The best fitting renewal process for P_{+-} has $p = 1/2$. A comparison of the P_{+-} (in log time form) to the renewal prediction in Figure 2a for $p = 3/4$ and $K = 4$ shows that the true distribution is stretched along the diagonal compared to the renewal process.

In real experiments, the data is binned for the comparison, but the KL metric can be defined in the continuum limit, which maximizes the KL metric for comparing the true distribution P with the model $\tilde{P}$.[21] Binning the data corresponds to coarse graining the distributions, which reduces the KL metric. The extreme example is the spacing being divided into a single bin, where all data points fall into the bin and the log likelihood of P and $\tilde{P}$ is zero. For any binning, the likelihood calculation determines the probability that the model histogram $\tilde{P}$ is an adequate representation of the data histogram P. In the continuum limit, the KL metric comes from an integration over a two-dimensional contour

$$\int d\ln t_1\, d\ln t_2\, \delta I_{P|\tilde{P}}(\ln t_1, \ln t_2) = \frac{\int d\ln t_1\, d\ln t_2\, P(\ln t_1, \ln t_2) \ln(P(\ln t_1, \ln t_2)}{\tilde{P}(\ln t_1, \ln t_2)} \tag{10}$$

that is shown in Figure 2b. We call $\delta I_{P|\tilde{P}}(\ln t_1, \ln t_2)$ the KL difference function. To the first order in ΔP the KL difference function resembles the traditional difference function, $\delta P(\ln t_1, \ln t_2) = P(\ln t_1, \ln t_2) - P(\ln t_1)P(\ln t_2)$, but this term does not

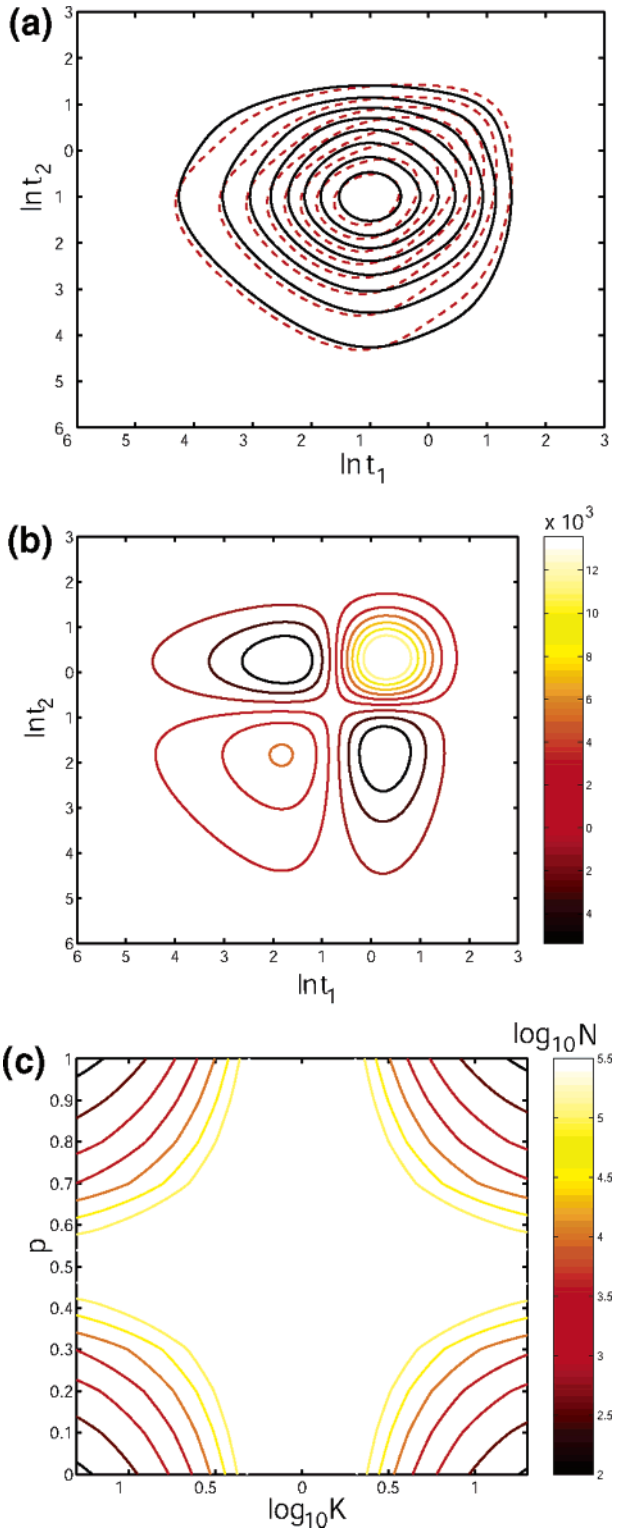

Figure 2. Testing the renewal hypothesis for the scheme in Figure 1a with $p = 3/4$ and $2k_1 = k_2 = K^{-1}k_3 = K^{-1}k_4 = 1$ for $K = 4$. (a) The two-event distribution, $P_{+-}(\ln t_1, \ln t_2)$ (dashed line) is compared with the best fitting renewal process $\tilde{P}(\ln t_1, \ln t_2) = P_+(\ln t_1)P_-(\ln t_2)$ (solid line). Note the log scales. (b) The KL difference, $\delta I_{P|\tilde{P}} = P_{+-} \ln(P_{+-}/P_+P_-)$ for comparing these two models. (c) The expected number of measurements required to distinguish the data from the renewal model at the 95% confidence level. As the model becomes closer to being a renewal process $p = 1/2$ or $K = 1$, more measurements are required to distinguish the two models.

contribute to the test of $\tilde{P}$.[7,25] The expected minimum number of measurements necessary to distinguish the renewal process from the nonrenewal process at a 95% confidence interval is plotted in Figure 2c.

The ability to detect a renewal violation depends on the magnitude of $p - 1/2$ and $\log(K)$. The degree of the renewal violation in the underlying scheme is captured by the magnitude of $p - 1/2$, while the magnitude of $\log(K)$ has the ability to distinguish substates in the underlying manifold, which is necessary to detect the violation. The waiting time is a signature of the substate that was entered. If the waiting times are identical, $\log(K) = 0$, then the signatures cannot distinguish the states, so the nonrenewal nature defined by $p - 1/2$ cannot be detected, while for cases where $|\log(K)|$ is large, the two states are easily identifiable and the nonrenewal nature is detectable. For systems that strongly violate the renewal property $p \approx 0$, 1, the number of measurements is reasonable (a couple thousand data points). Following our intuition, as the degree of the renewal violation decreases $p \rightarrow 1/2$, the number of measurements necessary to detect the violation increases and diverges for $p \approx 1/2$. Similar behavior is observed for the number of measurements necessary to distinguish different K values, since the system is a renewal process for $K = 1$.

The p parameter dominates the mixing of the process. For $p \approx 1/2$ the process mixes quickly, and even a full sequence analysis cannot distinguish the renewal and nonrenewal models. If the mixing is slow, but the kinetic rates are similar, $K \approx 1$, then a full sequence analysis can distinguish the data from a renewal model by detecting the weak but long-lived correlations in the waiting times, although the proposed two-dimensional test may be weak. One may be able to use two-dimensional analysis to overcome these difficulties by examining the probability distributions of sums of events, such as $P(t_1 + t_2, t_3 + t_4)$ or $P(t_1 + t_3, t_2 + t_4)$ (or the log equivalent). The first test would be sensitive to positive correlations, while the second test would be sensitive to negative correlations.

It is important to emphasize the difference between the mutual information and the correlation analysis. Symmetry may make the first few correlations zero even if the measured quantities are correlated at higher moments, whereas the mutual information is only zero if the two quantities are independent. The properties that one measures with correlations also need to be characterized by a numerical value, but mutual information only requires binning of the data, which can be performed on data with qualitative labels or multidimensional data. An example is a traditional photon counting experiment, where the arrival time and fluorescence lifetimes of the photons are recorded. In these experiments, the number of photons that arrive in a small time window and the average fluorescence lifetime in each bin may be recorded. To perform correlation analysis for the number and fluorescence lifetimes in two time windows separated by a fixed time t requires calculation of all possible correlations between the number of photons in different bins and the average fluorescence lifetime. For the mutual information, a two-dimensional histogram of the number of photons and average lifetime can be constructed, and then the mutual information for all of these inputs results in a single number. If two variables are Gaussian-distributed, then the correlation function and mutual information are related since the correlation function defines the probability distribution.

V. Comparing Experimental Conditions

The simple test can be extended to the determination of the existence of concentration dependence in experiments.[15] Many single-molecule experiments attempt to ascertain the mechanism of an enzyme's or ribozyme's reactivity by attaching a probe to the single molecule of interest by chemical modification. In this scenario, it becomes important to establish that the probe's motion is coupled to the reaction center of the single molecule by examining a substrate concentration dependence.[15]

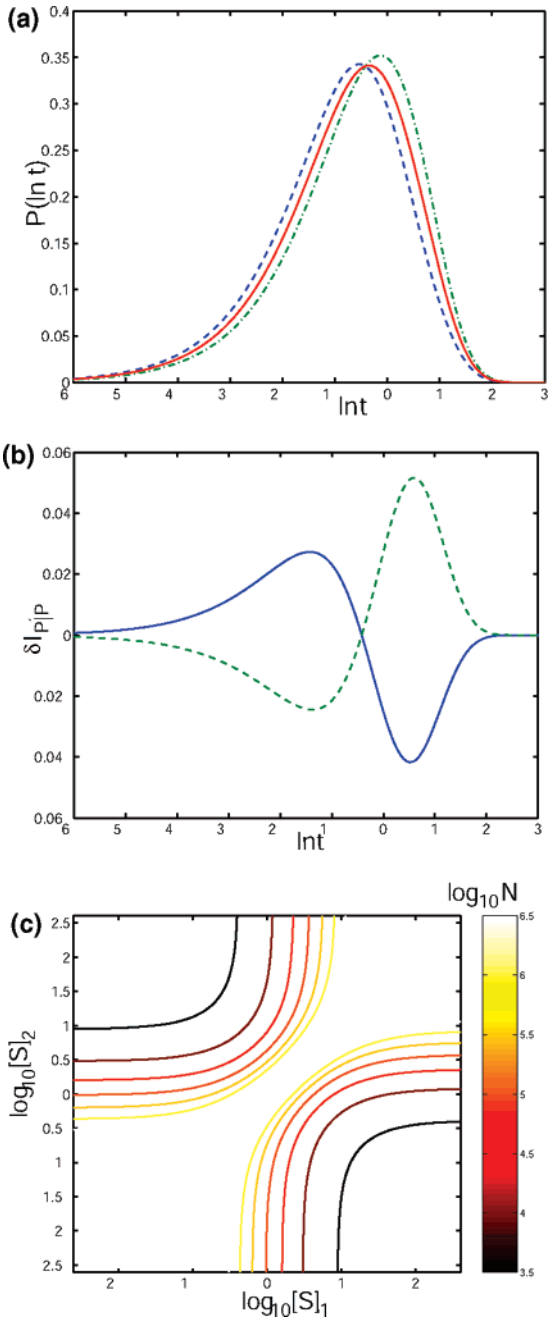

Figure 3. Ability to determine concentration-dependent behaviors in $P_+(t_1)$ in the model depicted in Figure 1b for $K_E = K_{ES^*} = 1$, $K_{ES} = 2$, $\gamma_{ES^*} = \gamma_S = 0$, and $\gamma_{ES} = \gamma_S^* = 1/5$. (a) $P_+(\ln t)$ for [S] = 1 (dot−dashed line) and [S] = 10 (dashed line) are compared with with $\tilde{P} = (1/2)(P_+([S] = 1) + P_+([S] = 10)$ (solid line). These probability distributions can be used to calculate $\delta I_{P|\tilde{P}}$, shown in part b for $P_+([S] = 1)$ (solid line) and $P_+([S] = 10)$ (dashed line). (c) The expected number of measurements needed to discriminate behaviors between $[S]_1$ and $[S]_2$ at the 95% confidence level.

We analyze the sensitivity in detecting changes for the model in Figure 1b, which is the reduced model that corresponds to Michaelis−Menten kinetics with an extremely fast product release step, a diffusion-limited substrate binding step, and a fluctuating kinetic rate for the enzyme−substrate to enzyme−

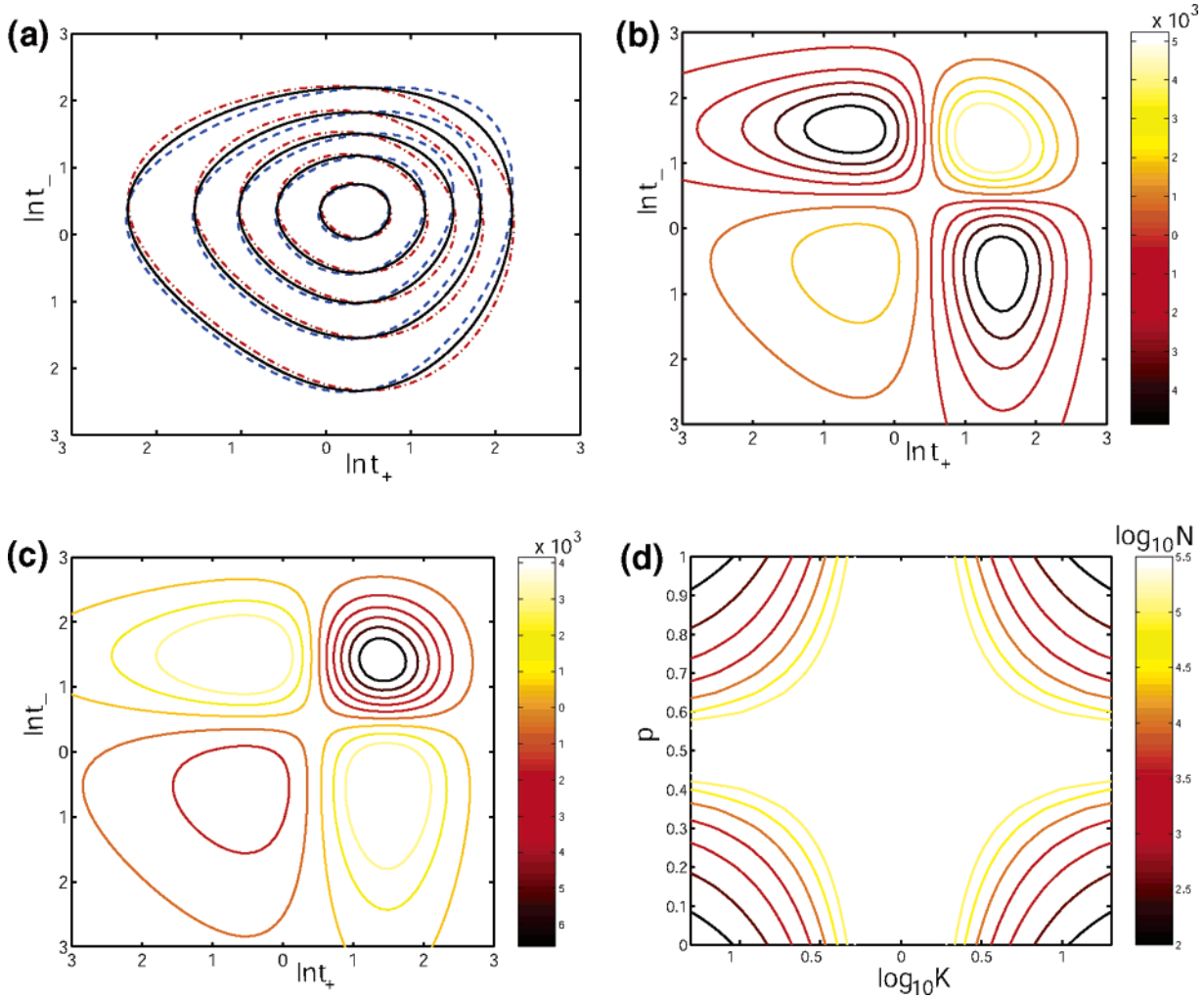

Figure 4. Determination of time reversibility in the model depicted in Figure 1a for $k_1 = k_2 = 1$, $k_3 = k_4 = K = 2.7$, and $p = {}^3/_4$. (a) $P_{+-}(\ln t_1, \ln t_2)$ (dashed line) and $P_{-+}(\ln t_2, \ln t_1)$ (dot−dashed line) are compared against the time reversible model, $\tilde{P} = ({}^1/_2)(P_{+-} + P_{-+})$ (solid line). (b) A contour of $\delta I^{(+)} = P_{+-}(\ln t_1, \ln t_2)\ln(P_{+-}(\ln t_1, \ln t_2)/\tilde{P}(\ln t_1, \ln t_2))$ (c) A contour of $\delta I^{(-)} = P_{-+}(\ln t_2, \ln t_1)\ln(P_{-+}(\ln t_2, \ln t_1)/\tilde{P}(\ln t_1, \ln t_2))$. (d) The expected number of measurements needed to discriminate P_{+-} and P_{-+} from $\tilde{P}$ at the 95% confidence level as a function of p and K.

product reaction. For simplicity we choose $K_E = K_{ES^*} = 1$, $K_{ES} = 2$, $\gamma_{ES^*} = \gamma_S = 0$, and $\gamma_{ES} = \gamma_{S^*} = {}^1/_5$. The − waiting time is a simple exponential process that depends linearly on the substrate concentration, [S], i.e., $P_-(t, [S]) = [S]\, e^{-[S]t}$. The + waiting time has a more complex substrate dependence

$$P_+(t,[S]) = \frac{1}{66 + 30[S]}(121\, e^{-(11/5)t} + (11 + 30[S])\, e^{-t}) \quad (11)$$

The first step in comparing two histograms $h_i^{(1)}$ and $h_i^{(2)}$ for different sets of experimental conditions is to construct the model for the two histograms being produced by the same underlying process, $\tilde{P}_i = 1/(N^{(1)} + N^{(2)})(h_i^{(1)} + h_i^{(2)})$. This model should be compared against the data to determine if a substrate dependence in the various measurements exists. The test is very strong for the linearly dependent rates of the − waiting time. Even for concentration differences of 5%, the linear dependence in the waiting times can be detected with less than 500 measurements. Of greater interest is the ability to detect the more subtle dependence in the + waiting time distribution. Figure 3a shows that even for a factor of 10 difference in the concentration, [S] = 1, 10, the waiting time distributions are similar and the composite model $\tilde{P}_+(t) = ({}^1/_2)(P_+(t,[S] = 1) + P_+(t,[S] = 10)$ can be a very good fit. The difference measure $\delta I^{(i)}{}_{P|\tilde{P}} = P_+^{(i)}(\ln t)\ln(P_+^{(i)}(\ln t)/\tilde{P}_+(\ln t))$ is shown in Figure 3b, with [S] = 1 ($i = 1$ (solid line)) and [S] = 10 ($i = 2$ (dashed line)). The expected number of measurements needed to discriminate $\tilde{P}$ from the two true probability distributions at the 95% confidence interval is presented in Figure 3c.

The ability to distinguish the two waiting time distributions depends on the difference in the concentrations. This waiting time distribution is a weighted average of two exponentials with concentration-independent decay constants but concentration-dependent weights. These changes in the weights saturate at high and low concentrations, which results in the plateau in the ability to detect the concentration dependence at high and low concentrations. This comparison of two distributions can be used to test different single molecules in the same experimental conditions or segments of a single trajectory to determine if the experiment is ergodic. As will be discussed elsewhere, this idea can be extended to examining collections of single molecules to classify their behaviors.[15]

VI. Time Reversibility

The existence of detailed balance is an important property to establish for various protein systems since detailed balance violations imply that the conformational kinetics being probed are also pumped by an external source of energy, such as the substrate in an enzymatic turnover process or the ionic potential

across a membrane that is often explored in ion channel experiments.[15,17] The easiest test of the probed coordinate's motion violating detailed balance is a substrate concentration dependence, as discussed above. These concentration dependences are not always easy to detect[17] and do not always give insight into the topology of the system that leads to the detailed balance violation.

As discussed previously,[15] there are several manifestations of detailed balance violations that can be seen in the two-dimensional event probability contours. These manifestations include a violation of time reversibility $P_{+-}(t_1, t_2) = P_{-+}(t_2, t_1)$, a violation of the triangle inequality of same event measurements, $P_{++}(t_1, t_2)^2 \leq P_{++}(t_1, t_1)P_{++}(t_2, t_2)$, and a peak in the single waiting time distribution.[15] Most previous analyses concentrated on the time reversibility.[17] Here we will use one- and two-dimensional analyses to explore all three of these possible manifestations of detailed balance violation without resorting to examining the underlying model.

If detailed balance holds for a system, then the system is time reversible, and the statistics of the forward and backward processes are identical[15]

$$P_{+-}(t_1, t_2) = P_{-+}(t_2, t_1) \tag{12}$$

A typical realization of a time reversibility violation occurs when there is a circulation loop in the underlying topology of the kinetic scheme that enters both manifolds at least twice. The simplest realization of this has only four substates, $+_1$, $+_2$, $-_1$, and $-_2$, with the conformational dynamics preferring to proceed in a circular sequence, $+_1 \rightarrow -_1 \rightarrow +_2 \rightarrow -_2 \rightarrow +_1 \rightarrow \ldots$ This situation is depicted in Figure 1a.

The time reversibility is easily tested within the framework applied to test for concentration dependence since time reversibility reduces to determining if two probability distributions are identical. Comparing P_{+-} and P_{-+} in the model in Figure 1a, with $k_1 = k_2 = K^{-1}k_3 = K^{-1}k_4 = 1$ as a function of p and K results in Figure 4. As shown in Figure 4a, for $p = {}^3/_4$ and $K = 2$, $P_{+-}(t_1, t_2)$ is elongated along the $t_1 = t_2$ line compared to $P_{-+}(t_2, t_1)$ (remember logarithmic binning). The alternative hypothesis that $P_{+-}(t_1, t_2) = P_{-+}(t_2, t_1) = \tilde{P}(t_1, t_2)$ with $\tilde{P} = ({}^1/_2)(P_{+-}(t_1, t_2) + P_{-+}(t_2, t_1))$ is similar to both distributions. The KL differences, $\delta I^{(+)}_{P|\tilde{P}} = P_{+-}\ln(P_{+-}/\tilde{P})$ and $\delta I^{(-)} = P_{-+}\ln(P_{-+}/\tilde{P})$, are plotted in Figures 4b and 4c, respectively. Since the P_{+-} distribution is elongated along the diagonal compared to P_{-+}, $\delta I^{(+)}$ is positive along the diagonal and negative on the off diagonal. The other KL difference, $\delta I^{(-)}$, shows the opposite behavior with a negative diagonal and a positive off diagonal.

The necessary number of measurements to distinguish P_{+-} and P_{-+} from $\tilde{P}$ at the 95% confidence interval is plotted in Figure 4d. Similar to the renewal indicator, the ability to discriminate depends on the magnitudes of $p - {}^1/_2$ and $\log(K)$. The magnitude of $p - {}^1/_2$ is a measure of the detailed balance violation, while the magnitude of $\log(K)$ is a measure of our ability to distinguish the two states. If two states are distinguishable ($|\log(K)|$ is large), then it is easier to detect the detailed balance violation. If p is very different from ${}^1/_2$, ($p \rightarrow 0, 1$) but K is near unity, then the time reversible and irreversible models may still be discriminated by either complete sequence analysis or comparing sums of events in another two-dimensional analysis as discussed in section 4.

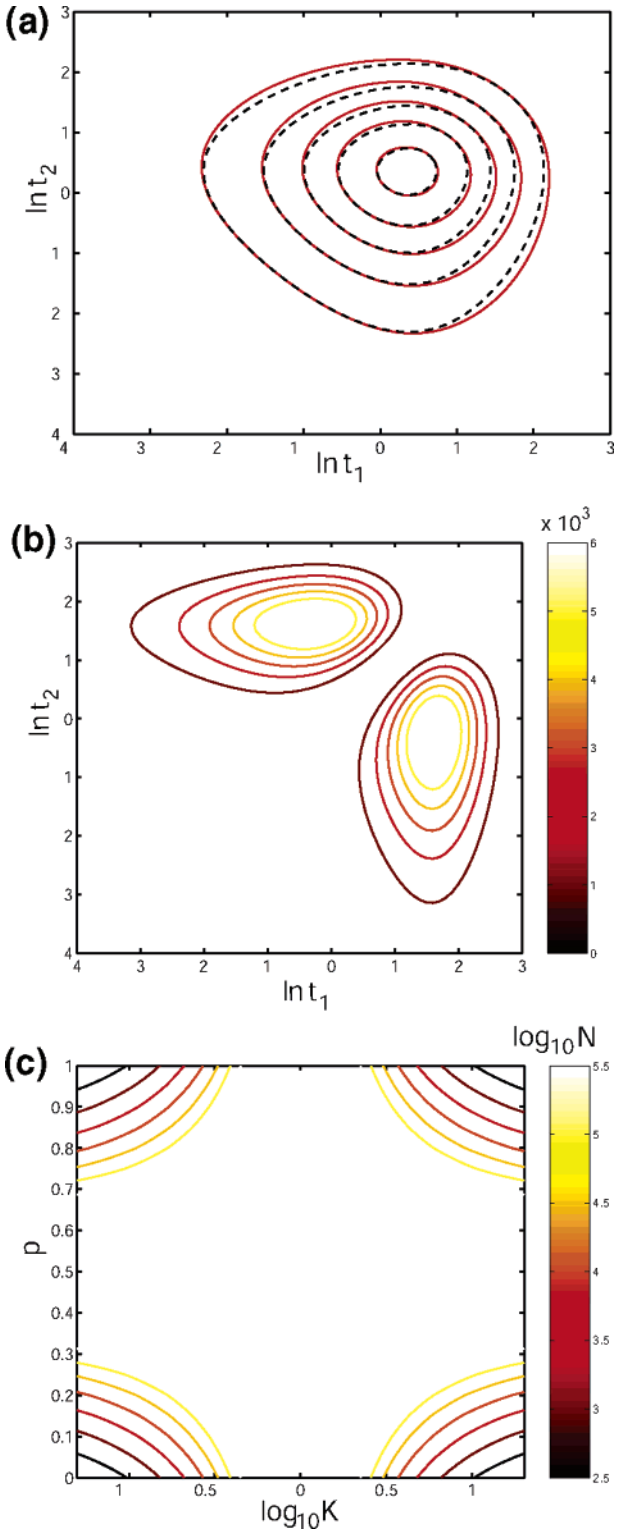

Figure 5. Ability to discriminate a diagonal dominance violation in the model depicted in Figure 1a for $k_1 = k_2 = k_4 = 1$, $k_3 = K = 2.7$, and $p = {}^3/_4$. (a) $P_{++}(\ln t_1, \ln t_2)$ (solid line) is compared against $\sqrt{P_{++}(\ln t_1, \ln t_1)P_{++}(\ln t_2, \ln t_2)}$ (dashed line). (b) $\Delta P_{++}(\ln t_1, \ln t_2) = P_{++}(\ln t_1, \ln t_2) - \sqrt{P_{++}(\ln t_1, \ln t_1)P_{++}(\ln t_2, \ln t_2)}$, which shows two positive off-diagonal peaks indicating a detailed balance violation. (c) The expected number of measurements needed to determine the existence of a diagonal dominance violation.

VII. Diagonal Features

If the waiting time in one of the manifolds has a rate-limiting step that is the same for all possible paths, such as substrate transport, then we expect a degeneracy in the eigenspectrum of the waiting time distribution.[15] This degeneracy can make the sequence time reversible even if it violates detailed balance. In these cases, the distribution of two similar events such as two

+ events separated by a − event must be used to test for detailed balance violations. As an example, setting $k_2 = k_4$ for the model in Figure 1a makes the − waiting time a simple exponential, $P_-(t) = k\ e^{-kt}$, and the system time reversible, but there is a detailed balance violation, and two adjacent + events can show features of this violation.

One feature of a detailed balance obeying P_{++} distribution is diagonal dominance, where $P_{++}(t_1, t_2)^2 \leq \sqrt{P_{++}(t_1,t_1)P_{++}(t_2,t_2)}$.[15] The violation of this diagonal dominance for the model in Figure 1a with $k_1 = k_2 = k_4 = 1$, $k_3 = K = 2.7$, and $p = {}^3/_4$ is shown in Figures 5a and 5b. Figure 5a shows that the isocontours of $\sqrt{P_{++}(t_1,t_1)P_{++}(t_2,t_2)}$ are narrower than those of the true distribution, so the distribution becomes greater than the theoretical detailed balance limit, resulting in the positive difference between $P_{++}(t_1, t_2)$ and $\sqrt{P_{++}(t_1,t_1)P_{++}(t_2,t_2)}$. In other words

$$\Delta P_{++} = P_{++}(t_1, t_2) - \sqrt{P_{++}(t_1, t1)P_{++}(t_2, t_2)} > 0 \tag{13}$$

indicates a detailed balance violation.

Diagonal dominance holds for integration over t or ln t, so

$$\left[\int_{t_1-\Delta/2}^{t_1+\Delta/2} dt' \int_{t_1-\Delta/2}^{t_1+\Delta/2} dt''\, P_{++}(t', t'')\right]^2 > \left[\int_{t_1-\Delta/2}^{t_1+\Delta/2} dt' \int_{t_1-\Delta/2}^{t_1+\Delta/2} dt''\, P_{++}(t', t'')\right] \times \left[\int_{t_1-\Delta/2}^{t_1+\Delta/2} dt' \int_{t_1-\Delta/2}^{t_1+\Delta/2} dt''\, P_{++}(t', t'')\right] \tag{14}$$

also indicates a detailed balance violation, and the diagonal dominance test is also valid for a histogram $P_{ij} = a_{ij}P_i^{1/2}P_j^{1/2}$.[15] Unlike previous tests, histograms may improve the KL measure of the detailed balance violation by allowing comparison of a narrow diagonal feature with broad off-diagonal features, which can also violate detailed balance and has appeared in some models that violate detailed balance.[28] Similar to the time reversibility test, a diagonal dominance violation occurs when there is a circulation loop in the underlying topology of the kinetic scheme that enters both manifolds at least twice, as depicted in Figure 1a.

To demonstrate a diagonal dominance test, we examine the model in Figure 1a. The two large off-diagonal peaks in the ΔP_{++} distribution in Figure 5a indicate that splitting the distribution into four quadrants along the diagonal, $t_1 = t_2$, should be sufficient to test for diagonal dominance. The position of the split depends on p and K. For a histogram with only four quadrants, diagonal dominance implies that $\tilde{P}_{12}{}^2, \tilde{P}_{21}{}^2 \leq \tilde{P}_{11}\tilde{P}_{22}$. Assuming that we are testing detailed balance, time reversibility is also required, and the optimal diagonally dominant time reversible distribution is given by $\tilde{P}_{12} = \tilde{P}_{21} = (1/2N)(h_{12} + h_{21})$ and $\tilde{P}_{ii} = (1/N)h_{ii}$, for $({}^1/_4)(h_{12} + h_{21})^2 \leq h_{11}h_{22}$. If this inequality is not satisfied, then we must modify the probability to

$$\tilde{P}_{11} = \frac{\left(h_{11} + \frac{1}{2}(h_{12} + h_{21})\right)^2}{N^2}$$

$$\tilde{P}_{22} = \frac{\left(h_{22} + \frac{1}{2}(h_{12} + h_{21})\right)^2}{N^2}$$

$$\tilde{P}_{21} = \tilde{P}_{12} = \frac{\left(h_{11} + \frac{1}{2}(h_{12} + h_{21})\right)\left(h_{22} + \frac{1}{2}(h_{12} + h_{21})\right)}{N^2} \tag{15}$$

Following previous analyses, we compare this model to the data, $P_{ij} = h_{ij}/N$ to determine the probability of a diagonal dominance violation. The number of measurements necessary to discriminate a diagonal dominance violation at the 95% confidence interval using the four quadrant test on the model in Figure 1a is plotted in Figure 5c as a function of p and K. The features are similar to those in the previous tests, with the discriminating power of the test depending on the magnitude of $p - {}^1/_2$ and log(K) since these measure the detailed balance violation and distinguishability, respectively.

VIII. Single Waiting Time Test

A one-dimensional feature that indicates detailed balance violations is the existence of a peak in the single waiting time distribution.[15] This detailed balance violation has a different origin than the previously discussed time reversibility violation and diagonal dominance violations that are usually associated with the circulation loop passing though the + and − manifolds at least twice. A peak in the single waiting time distribution results from a flow within a single manifold so that the system has a tendency to enter the + or − manifolds through one substate and exit through another. This indicates a microscopic time reversibility violation between states in the same manifold even though the mesoscopic time reversibility may hold. If detailed balance holds, then the single waiting time distribution can be expressed as a sum of exponentials $P(t) = \int dk\ P(k)k\ e^{-kt}$, where $P(k)$ is a proper probability density, $P(k) \geq 0$, $\int dk\ P(k) = 1$. A rigorous method of testing for a peak is to determine $P(k)$ from maximum entropy fits or another method and compare this probability distribution to the data.

As a simple example, we examine the + waiting time distribution, $P_+(t)$, in the model depicted in Figure 1c with $k_1 = k_2 = k_3 = K^{-1}k_4 = 1$. This waiting time distribution with $K = 2.7$ and $p = {}^1/_{10}$ is compared against the best-fitting detailed balance obeying distribution, $\tilde{P} = \int dk\ P(k)k\ e^{-kt}$, in Figure 6a. The detailed balance distribution is wider than the detailed balance violating scheme. Figure 6b shows the KL difference, $\delta I_{P|P'}$. Similar to the previous tests, the ability to detect this detailed balance violation increases with increasing magnitude of $p - {}^1/_2$, but the ability to distinguish the peak varies inversely with the magnitude of log(K). Taking $p = 1$, the waiting time distribution is given by

$$P_+(\ln t) = \frac{k_4}{k_4 - k_1}k_1\ e^{-k_1 t} - \frac{k_1}{k_4 - k_1}k_4\ e^{-k_4 t} \tag{16}$$

which has a zero at $t = 0$. If k_1 is much smaller than k_4 ($K \gg 1$), then the waiting time of the system is nearly monoexponential, $P_+(t) \approx k_1\ e^{-k_1 t}$, with only a brief deviation at short times, so it is difficult to detect the detailed balance violation. Similar results hold for $K \ll 1$, and it is only when $k_4 \approx k_1$ that the deviation from simple exponential behavior can be detected.

IX. Conclusion

As demonstrated above, one- and two-dimensional histograms can elucidate many properties of a system without solving the combinatorial complex problems of determining the exact underlying model. These methods allow detection of correlations in events through the renewal test, similarities in behaviors under different experimental conditions, and detailed balance violations. The renewal test determines if the transitions between the two manifolds correspond to multiple paths, and the experimental condition dependence indicates that the probe is coupled to the reaction coordinate. The detailed balance

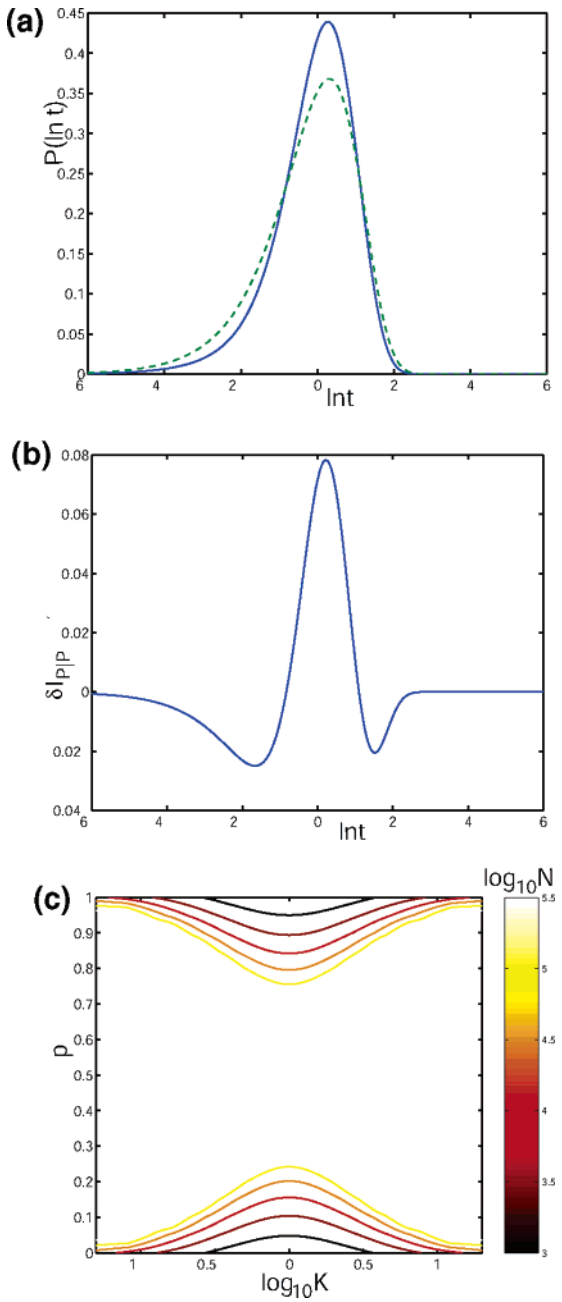

Figure 6. Ability to determine if the waiting-time distribution is not consistent with a detailed balance scheme. (a) The single + waiting time distribution, $P_+(\ln t)$ in the model depicted in Figure 1c for $k_1 = k_2 = k_3 = 1$, $k_4 = K = 2.7$, and $p = 1/10$ (solid line) is compared to the best fit of a detailed balance obeying scheme, $\tilde{P}(t) = \int dk\, P(k) k\, e^{-kt}$, $P(k) > 0$ (dashed). (b) $\delta I_{P|\tilde{P}}$ for the distributions in part a. (c) The expected number of measurements needed to discern the detailed balance violation as a function of K and p. Unlike other tests, the ability to determine the existence of a peak in the waiting time distribution depends on $K \approx 1$.

violations can result in time reversibility violations or a lack of diagonal dominance that indicates a circulation loop that goes through both manifolds as least twice and peaks in the single waiting time distribution, which indicates a multiple-step circulation through a single manifold. Knowledge of these properties can give insight into the underlying topology of Markovian models without determining the specific model parameters.

The proposed tests not only allow determination of the existence of these properties but also create rigorous bounds on the ability to discriminate models with one- and two-dimensional data. The number of measurements necessary to distinguish these models at the 95% confidence interval cannot be reduced by introducing another measure of these data. The only major assumption that may need to be corrected is the independence of events, which may not be true if the histogram is constructed from parsing a long single trajectory, but this will only modify the variance estimate.

The analysis can be extended to higher-dimensional binned data, but creating histograms will not be practical. Instead, one needs to fit the data to flexible functional forms that continue to obey the restrictions on the properties, such as time reversibility. These reduced information tests will never be as powerful as full sequence analysis, but even using flexible functional forms is orders of magnitude less computationally intensive than a full sequence analysis and does not require one to propose underlying topologies. These reduced information methods can also be expanded to classify different molecular trajectories, to test ergodicity, and to determine properties of the transition-state ensemble. The computational simplicity, along with the rigorous bounds in the ability to discriminate models, makes the information theoretical approach to reduced data representations an advantageous first step in performing single-molecule analysis.

Acknowledgment. This research is supported by the National Science Foundation Career Award (Che-0093210) and the Camille Dreyfus Teacher−Scholar Award.

Appendix

To simplify the presentation our focus has been on $P(t_1, ..., t_n)$, but the results are independent of binning methods and can be easily applied to $P(\ln t_1, ..., \ln t_n)$. In fact, logarithmic binning gives a nice interpretation of one- and two-dimensional histograms of events. If the kinetics of the system corresponds to simple first-order kinetics with a complete basis set of eigenvectors, such as for systems where detailed balance holds, then the n event probability distribution (histogram) is

$$P_{\pm...\pm}(t_1, ..., t_n) = \int \left[\prod dk_i\, k_i\, e^{-k_i t_i} \right] P(k_1, ... k_n) \quad (17)$$

Logarithmic binning allows plotting of data with multiple time scales and allows us to write the waiting time distribution as

$$P_{\pm...\pm}(\ln t_1, ..., \ln t_n) = \int \prod d \ln \tau_i\, f(\ln t_i - \ln \tau_i) P(\ln \tau_1, ..., \ln \tau_n) \quad (18)$$

where $f(x) = \exp(x - e^x)$ is a Gumbel distribution and $\tau_i = k_i^{-1}$. From this expression, it becomes apparent that $P_{\pm...\pm}(\ln t_1, ..., \ln t_n)$ is a convolution of $P_{\pm...\pm}(\ln \tau_1, ..., \ln \tau_n)$ with Gumbel distributions. Since the Gumbel distribution is a probability distribution, it is a simple smear factor. As a result, in the limit of large data, one can perform density functional theory (DFT) calculations of the log distribution and filter out the Gumbel smear factor to obtain the Fourier transform of the spectrum of kinetic rates. Inverse DFT will yield a spectrum of peaks that correspond to the logarithmic time scales of the system and can be interpreted similar to two-dimensional spectroscopy. The peaks in the two-dimensional spectrum correspond to coupling of different time scales of motions. These couplings can be positive or negative depending on the connectivity of the system.

References and Notes

(1) Moerner, W. E.; Orrit M. *Science* **1993**, *283*, 1670.
(2) Xie, X. S.; Trautman, J. K. *Annu. Rev. Phys. Chem.* **1998**, *49*, 441.

(3) Weiss, S. *Nat. Struct. Biol.* **2000**, *7*, 724.
(4) Barkai, E.; Jung, Y.; Silbey, R. J. *Annu. Rev. Phys. Chem.* **2004**, *55*, 457.
(5) Gopich, I. V.; Szabo, A. *J. Chem. Phys.* **2003**, *118*, 454.
(6) Lippitz, M.; Kulzer, F.; Orrit, M. *ChemPhysChem* **2005**, *6*, 770.
(7) Cao J. S. *Chem. Phys. Lett.* **2000**, *327*, 38.
(8) Bruno, W. J.; Yang, J.; Pearson, J. E. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6326.
(9) Ball, F. G.; Rice J. A. *Math. Biosci.* **1992**, *112*, 189 and notes and references within.
(10) Yang, H.; and Xie, X. S. *Chem. Phys.* **2002**, *284*, 423.
(11) Watkins, L. P.; Yang, H. *Biophys. J.* **2004**, *86*, 4015.
(12) Andrec, M.; Levy, R. M.; Talaga, D. S. *J. Phys. Chem. A* **2003**, *107*, 7454.
(13) Kou, S. C.; Xie, X. S.; Liu, J. S. *J. R. Stat. Soc. C* **2005**, *54*, 469.
(14) Witkoskie, J. B.; Cao, J. S. *J. Chem. Phys* **2004**, *121*, 6361.
(15) Witkoskie, J. B.; Cao, J. S., to be submitted for publication.
(16) Qian, M.; Qian. H. *Phys. Rev. Lett.* **2000**, *84*, 2271.
(17) Song, L.; Magleby, K. L. *Biophys. J.* **1994**, *67*, 91.
(18) Feller, W. *An Introduction to Probability Theory and its Applications*; Wiley: New York, 1970; Vol. 2.
(19) Rice, S. A. In *Comprehensive Chemical Kinetics*; Bamford, C. H., Tipper, C. F. H., Compton, R. G., Eds.; Elsevier: New York, 1985; Vol. 25.
(20) Hodgson, M. E. A.; Green, P. J. *Proc. R. Soc. London, Ser. A* **1999**, *455*, 3425.
(21) Legeza, O.; Solyom, J. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 7927.
(22) Kullback, S.; Leibler, R. A. *Ann. Math. Stat.* **1951**, *22*, 79.
(23) Kullback, S. *IEEE Trans. Inf. Theory* **1967**, *IT13*, 126.
(24) Kuha, J. *Sociol. Methods Res.* **2004**, *33*, 188.
(25) Yang, S. L.; Cao, J. S. *J. Chem. Phys.* **2002**, *117*, 10996.
(26) Flomenbom, O.; Klafter, J.; Szabo, A. *Biophys. J.* **2005**, *88*, 3780.
(27) Nalewajski, R. F. *J. Phys. Chem. A* **2000**, *104*, 11940.
(28) Lerch, H. P.; Rigler, R.; Mikhailov, A. S. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 10807.