Detection of Subtle Dynamical Changes Induced by Unresolved "Conformational Coordinates" in Single-Molecule Trajectories via Goodness-of-Fit Tests

Christopher P. Calderon*

High Performance Computing Research Department, Lawrence Berkeley National Laboratory, Berkeley, California 94720

Received: November 23, 2009; Revised Manuscript Received: January 12, 2010

Single-molecule experiments are allowing researchers to track the evolution of a few order parameters characterizing complex biomolecules. At fine temporal resolution, artifacts of unresolved degrees of freedom, for example, those induced by collective molecular motion, often influence the dynamics. Reliably detecting subtle changes in dynamics at the nanoscale can be difficult due to the inherent stochasticity, but such changes can have relevance to understanding complex enzyme kinetics. Surrogate models can be used to summarize the information content in single-molecule time series (containing fluctuations occurring over multiple time scales). The focus in this article is on detecting slow time scale changes through the use of the surrogates. The conditional density, associated with the surrogates, allows one to formulate quantitative hypothesis tests which can detect the influence of unresolved coordinates in cases where the dynamics are modulated subtly. The relevance of quantitative (and appropriate) testing methods to analyze single-molecule time series is discussed and demonstrated. A brief discussion on some merits of using frequentist (versus Bayesian) time series methods to analyze single-molecule data is also presented. Idealized simulations mimicking features relevant to some enzyme systems where an "unresolved conformational coordinate" slowly evolves (1) with inertia and (2) diffusively are studied in the nonstationary (nonergodic) setting; however, the findings are also relevant to experimentally measured time series and stationary signals.

I. Introduction

When small-scale biological systems are tracked at the single-molecule level, the time-ordered observation sequence contains substantial noise, and the fluctuations contain contributions from multiple scales. There is great potential for gaining both fundamental and technologically relevant insights into complex biological systems from single-molecule data. Many early studies have already demonstrated such findings.[1−6] However, traditional analysis methods applied to single-molecule signals encounter several technical difficulties when they attempt to extract the information from single-molecule signals. One difficulty comes from unresolved degrees of freedom. These "lurking" coordinates can substantially modulate the stochastic dynamics of the observables monitored.[5,7−10] Given that the resolution of single-molecule experiments is ever increasing, there is interest in methods for quantitatively capturing signatures of such unresolved degrees of freedom and making more precise (less coarse-grained) statements about various mechanisms; see, for example, refs 10−13. Modern computational power allows researchers to entertain new methods for summarizing the information content in these rich data sets; surrogate models[8,9,14,15] and functional data analysis[16] are some examples.

In this article, stochastic differential equations (SDEs) are considered as surrogates of the (more complex) true data-generating process. The models are data-driven in the sense that a loose structure is imposed, but observational data is used to find parameters specifying the surrogates. The Markovian surrogate models can be fit by maximum likelihood type techniques.[14,17] In order to fit the model, a means for numerically evaluating a transition density (or conditional density) estimate is required.[18] The use of the transition density facilitates inference tasks. For example, inference methods can be used to quantitatively determine if one is statistically justified in ignoring the velocity (or inertia) of the resolved coordinate, given the proposed model and observed data; for examples, see ref 8.

The questions of interest here are associated with how the unresolved coordinate influences the dynamics of the resolved coordinate. If an unresolved coordinate is coupled (either "kinetically" or "thermodynamically") to the resolved coordinate and the former changes appreciably over the time scale over which observations are made, a single Markovian model will be rejected given enough data.[19−21] It is natural to ask, "How accurately can one determine (using a finite amount of discretely observed temporal data) if unresolved coordinates introduce statistically significant changes in the stochastic dynamics?" and "How much does the unresolved coordinate need to change in its value before the (subtle or dramatic) changes become detectable?" It is demonstrated how recent hypothesis tests, using the transition density structure provided by the data-driven surrogate models, can be used to detect subtle changes induced by coordinates not explicitly modeled by the surrogate. More importantly, it is demonstrated how an omnibus test[19] has substantial power (even with moderate sized times series samples) in detecting subtle changes induced by an unresolved conformational coordinate. The unresolved conformational coordinate is constructed to evolve on time scales that are slower than those associated with the resolved coordinate. In this article, several toy systems are studied where the "unresolved conformational coordinate" evolves via dynamical rules in which inertia is non-negligible and in cases where the unresolved coordinate evolves diffusively (or in an "overdamped" fashion).

* To whom correspondence should be addressed. E-mail: CPCalderon@lbl.gov.

10.1021/jp911124z © 2010 American Chemical Society
Published on Web 02/11/2010

In all cases, the changes in this unresolved coordinate alter the dynamics of the resolved coordinate. The connection of the model features studied to real enzymes is briefly sketched. In ref 22, some consequences of ignoring the dynamics of the unresolved coordinates are discussed more extensively; for example, it is shown how the autocorrelation function can be substantially changed by the types of (seemingly innocuous) subtle changes studied here.

The remainder of the article is organized as follows. The motivation, candidate surrogates, procedure used to estimate and assess the models, and the data-generating models considered are presented in section II. This section also contains a discussion comparing/contrasting this approach to previous efforts in chemical physics. Section III follows with the results and discussion, and section IV summarizes and highlights the relevance of the findings to several different types of single-molecule studies.

II. Background and Methods

A. Basic Setup. The basic situation of interest here is the following: An unresolved (assumed experimentally unobservable) coordinate, x, modulates the dynamics of a resolved (assumed experimentally observable) coordinate y. The characteristic time scales associated with the x dynamics are "slow"[23] relative to those of y. Furthermore, the x coordinate is nonstationary (the moments of the process change with time). However, the coupling is such that the nonstationarity of y is not as pronounced as it is in x, that is, the former appears to be roughly a stationary stochastic process over the time scale of observations. The details of the evolution equation (for either coordinate) are assumed unknown a priori to the investigator. It is also assumed that a single marginal distribution in the resolved coordinate, y, is not adequate for characterizing the observed distribution of y (i.e., one unwilling to assume stationarity) and/or the researcher cannot reliably determine the initial distribution associated with y using the stationary distribution predicted using a single stochastic model calibrated from observing y alone (these types of situations are particularly relevant to atomistic modeling[8]). In the cases presented, the nonlinear equations are of course known, but before displaying these equations, the physical motivation is first described, and two different situations of interest are then described qualitatively.

Regarding the physical motivation, it has been suggested that the "floppiness" of molecules and/or local unfolding can be important to a protein's functionality.[11,24−26] Motions associated with these features can potentially be detected by measuring the local diffusion coefficient and/or estimating statistically significant changes in the effective molecular stiffness.[25] Both of these features can be difficult to quantitatively measure/detect in single-molecule trajectories using standard statistical physics methods; furthermore, aggregating or combining paths (i.e., trajectories) from different experiments may obscure kinetic or thermodynamic details of interest. Previous works have demonstrated this problem in adenylate kinase[8] and dihydrofolate reductase.[22] In this article, we demonstrate how subtle changes induced by unresolved coordinates can be detected on a pathwise basis in highly controlled examples. It should be noted that minor variants of the methods presented here are also applicable to experimental time series where measurement noise is also present and must be dealt with to detect the physical signal of interest (for examples, see refs 9 and 15), but here, we focus exclusively on cases without measurement noise.

As for the qualitative descriptions of the idealized models, in Case 1 situations, the unresolved x coordinate warps the shape of the underlying potential energy surface, $U(x, y)$. The force acting on y comes from taking the gradient of this potential, that is, $\partial U(x, y)/\partial y \equiv \nabla yU(x, y)$. Figure 1A displays U as a function of y for a various fixed values of x (this situation closely mimics a study in ref 23). Note that as x changes, the well bottom changes as does the barrier height. In Case 2, the potential energy surface governing the force experienced by y is not affected by x, but there is a kinetic or dynamic coupling. The magnitude of the instantaneous Brownian noise $\sigma^y(x, y)$ experienced by y is coupled to x. In our controlled example, the function $\sigma^y(\cdot,\cdot)$ is related to the position-dependent effective friction coefficient, $\gamma^y(x, y)$ experienced by y and is given by the Einstein relation, that is, $\sigma^y(x, y)^2 = k_BT/\gamma^y(x, y)$, where k_BT denotes Boltzmann's constant multiplied by the system temperature. The coupling is demonstrated in Figure 1B, where one can see that the range of x studied makes fairly small changes to σ^y. The changes induced in the dynamics are even less pronounced.

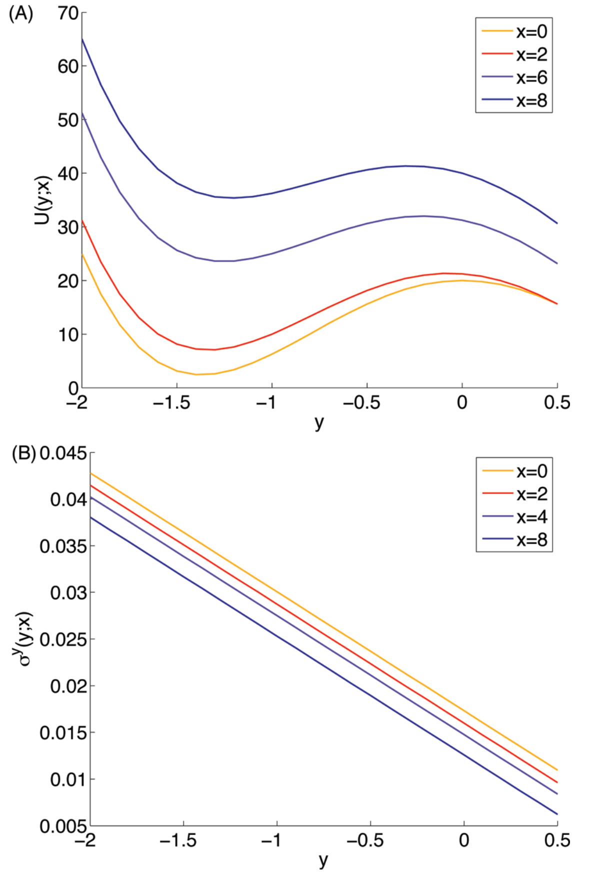

Figure 1. (A) Slices of the potential energy function, $U(x, y)$ of Case 1 (see eq 3) for four different fixed values of x. (B) plots Similar slices for the noise magnitude function, $\sigma^y(x, y)$.

B. Data-Generating Processes. Various dynamics for the unresolved x process will be considered (the equations are provided in the Appendix), but the main feature that all evolution rules considered have in common is that x starts near 0 and increases at a roughly constant rate over time. Both situations where inertia is important and one in which x evolves as a diffusive process (with drift) are considered. Cases where inertia in x is important will be labeled as A and those where x evolves diffusively will use the label B, for example, Case 1A or Case 1B. In all cases, the resolved y coordinate evolves according to a stochastic process having the form

$$\begin{aligned} \gamma^y(x,y)\frac{\mathrm{d}y}{\mathrm{d}t} &= -\frac{\partial U(x,y)}{\partial y} + N(t) \\ \langle N(t')N(t)\rangle &= 2k_\mathrm{B}T\gamma^y(x,y)\delta(t-t') \end{aligned} \tag{1}$$

The convention above is commonly used in statistical physics,[27−29] where $-[\mathrm{d}U(x,y)/\mathrm{d}y]$ is the instantaneous "systematic force" acting on y, $N(t)$ denotes a Gaussian mean zero "random force" white noise process whose covariance is determined by the second line in the equation above (angle brackets denote an ensemble average), $\gamma^y(x,y)$ represents a state-dependent "friction" coefficient, and $\delta(\cdot)$ represents the Dirac delta function. In the mathematics and statistics communities, stochastic differential equations (SDEs),[30] interpreted in the Itô sense, are commonly used to describe random processes. A process analogous to that in eq 1 could be written as

$$\mathrm{d}y_t = -\sigma^y(x_t,y_t)^2/k_\mathrm{B}T\nabla_y U(x_t,y_t)\mathrm{d}t + \sqrt{2}\sigma^y(x_t,y_t)\mathrm{d}B_t \tag{2}$$

where B_t represents a standard Brownian motion (t subscripts denote the time index of a SDE). The two coefficient functions (one in front of dt and the other in front of dB_t) along with the instantaneous state values completely characterize the dynamics in this model. The square of the latter, $\sigma^y(x_t, y_t)^2$, will be referred to as the local diffusion coefficient function. Furthermore, the SDE above implicitly assumes the Einstein relation. However, it is to be noted that other functional forms (not appealing to this Einstein relationship) can be readily entertained as models using the statistical estimation and inference tools discussed later.

The aim is to estimate and infer how y influences the functions σ^y and $\partial U/\partial y$. Note that it is assumed that x is not observed. Later, the procedure for locally approximating these functions given discretely observed time series of y will be briefly reviewed.[8,17,31] Also an "overdamped Langevin"[14,32] structure is enforced in eqs 1 and 2. That is, in the data-generating process, inertia of the fast resolved coordinate y is unimportant by construction. Methods for testing the validity of an overdamped type assumption in cases where a y type coordinate potentially has inertial contributions were considered elsewhere.[8] The interest in this article is in quantifying how an evolving x process modulates the y dynamics using recently developed frequentist inference methods. The questions of interest are (1) How accurately can one detect changes in the evolution rules governing y's dynamics?, (2) How much does x need to change in value before the evolution rules change enough so that rejection of the null hypothesis occurs?, and (3) Does inertia in the unresolved coordinate help in detecting model misspecification due to the built-in (velocity-induced) memory of this non-Markovian noise source?

To help address these issues, the equations used to generate Figure 1 are used also to generate sample paths. The equations are

Case 1:

$$\begin{aligned} U(x,y) &= \alpha/2(y - x/2)^2 + \kappa(y^2 - \phi)^2 \\ \sigma^y(x,y)^2/k_\mathrm{B}T &\equiv 1/\gamma^y(x,y) = 1/\gamma \end{aligned}$$

Case 2:

$$\begin{aligned} U(y) &= \alpha/2y^2 + \kappa(y^2 - \phi)^2 \\ \sigma^y(x,y)^2/k_\mathrm{B}T &\equiv 1/\gamma^y(x,y) = 1/\gamma(1/2(\exp\zeta x + y/\psi))^2 \end{aligned} \tag{3}$$

Although the nonlinear equations specifying the SDEs in the various cases are stated above explicitly, we are interested in attempting to locally approximate these nonlinear equations (with simpler polynomial functions) due to the fact that we assume that the researcher will not have the luxury of this information.

C. Statistical Inference. The pathwise estimation and inference procedures used are described in refs 8 and 31. The estimation procedure is summarized briefly here. Given N discrete time-ordered observations of the resolved coordinate, $\{y_i\}_{i=0}^{N}$, divide that time series into W temporal windows $\{y_i\}_{i=N_0}^{N_1-1}$, $\{y_i\}_{i=N_1}^{N_2-1}$, ... $\{y_i\}_{i=N_{w-1}}^{N_w-1}$ with $N_0 = 0$, $N_i < N_{i+1}$, and $N_W - 1 = N$ (here, we use $W = 10$ and select the partitioning such that $\Delta N \equiv N_{i+1} - N_i = 500$). In each of these local windows, the surrogates

$$\begin{aligned} -\frac{\partial U(x,y)}{\partial y} &\equiv -\nabla_y U(x,y) \approx A + B(y - y^\mathrm{BASE}) \\ \sqrt{k_\mathrm{B}T/\gamma^y(x,y)} &\equiv \sigma^y(x,y) \approx C + D(y - y^\mathrm{BASE}) \end{aligned} \tag{4}$$

are used as proxies to the full nonlinear equations. The term y^BASE is a free parameter where we desire to approximate the value of the effective force (denoted by A) and its linear sensitivity with respect to y (denoted by B). To facilitate the presentation, we report results obtained using $y^\mathrm{BASE} = \psi \equiv$ the location of the well minima of $U(0, y)$. (The specification of the free parameter had virtually no numerical influence on the results reported.) Approximation of σ^y is similar, but note that the noise magnitude is allowed to change as a function of y.[8,33] With this surrogate local parametric structure for the two SDE functions, one can then plug the surrogate functions into eq 2 and approximate the transition density associated with the surrogate SDE using explicit closed-form expansions.[18] (This expansion is only guaranteed to converge under fairly general conditions which are violated here due to the possible degeneracy of the diffusion. The expansion used has been demonstrated to be accurate in applications demanding high accuracy, even in cases where formal convergence proofs are elusive.[18,19,31]) Note that plugging in these linear models results in a nonlinear surrogate SDE. The data and the transition density proxy are then used to construct an approximate maximum likelihood estimate (MLE). The parameter vectors $\theta \equiv (A, B, C, D)$ maximizing the corresponding likelihood are found for each temporal window. Denote the θ maximizing the likelihood in window N_i by $\hat{\theta}_{N_i}$.

The procedure is then repeated for each observed time series, resulting in an ensemble of estimated parameters, denoted by $\{\{\hat{\theta}_{N_i}^{(\mathscr{B}_j)}\}_{N_i=0}^{N_w}\}_{\mathscr{B}_j=1}^{\mathscr{B}}$, where the superscript $(\mathscr{B}_j)$ is used to identify the MLEs coming from time series batch (realization) $\mathscr{B}_j$. The estimated $\hat{\theta}_{N_i}^{(\mathscr{B}_j)}$ are used to formulate a null hypothesis, and the goodness-of-fit is assessed using Hong and Li's omnibus test[19] testing procedure. The method is applicable to a nonstationary time series; it can also treat models more involved than Markovian SDEs. If one is willing to assume that the time series is stationary, other tests have been shown to have better power;[20,21] however, the validity of the stationarity assumption is often questionable. In our controlled simulation study, it is known that there is an evolving unresolved degree of freedom causing the stationarity assumption to be invalid.

The distribution of Hong and Li's test statistic converges in distribution to a mean zero normal with variance 1, $\mathcal{N}(0,1)$, as the sample size tends to infinity.[19] The testing procedure attempts to check for any type of model misspecification, for example, non-Markovian noise sources and/or nonlinear dynamical effects. One practical problem associated with this test is that in finite samples, the test statistic distribution can differ substantially from $\mathcal{N}(0,1)$.[20,21] A computationally intensive parametric bootstrap procedure has been reported to alleviate this problem (and help incorporate parameter uncertainty into the test) in a stationary setting where the initial distribution of the resolved coordinate can be accurately modeled using observed data,[21] but this approach is not directly applicable in our case.

The procedure reported here uses a simple simulation-based approach to approximate the distribution of the test statistic. Two null hypotheses are tested (one each for Case 1 and Case 2 data). The surrogate SDE structure is used in all cases along with the corresponding transition density approximation. The selection of the parameter θ_0, which completes the specification of the null, is discussed in the Results and Discussion section. In order to approximate the test statistic distribution, $M = 1 \times 10^4$ batches of the times series, each containing $\Delta N = 500$ entries, were used to approximate the finite sample test statistic distribution under a prescribed null. That is, the surrogate SDE and a specific θ_0 value (known to the researcher) were used to simulate M sample paths using the SDE determined by eqs 2 and 4 and the specified θ_0. Discrete samples of the simulated SDE paths and the known (approximate) transition density were used to numerically compute M test statistics; the test statistic distribution under a known (fixed) data-generating process was available once this was done. The validity of using the null (SDE and specified parameter) can then be assessed using this null structure and the observed discrete data. Specifically, one computes the test statistics associated with realizations from a process where the data-generating process is not known to the researcher a priori, for example, the two-dimensional diffusions of Case 1 and 2. One can then make decisions on the validity of the null using test statistics computed from the data of interest and the test statistic distribution calibrated using the M idealized data sets. It should be noted that if the null is computed from observed data (as will be done later), then this procedure does not include parameter uncertainty information. Furthermore, this procedure only produces an approximation of the critical value associated with a desired significance level, α. A simple sensitivity analysis demonstrated that the approximate significance level could vary by as much as $\pm 5 - 10\%$ if perturbations commensurate in magnitude with the largest expected fluctuation in a parameter estimate obtained using only one set of 500 observation time series were made (i.e., a fluctuation due to sampling uncertainty associated with a single path) and the test statistic distribution was recomputed using M idealized paths. However, many perturbations consistent with the estimated parameter uncertainty had negligible influence on the computed null distribution (for example, the Case 1 and Case 2 null had different parameters, but the two distributions were hard to distinguish). Extensions of some computationally intensive procedures can help one in getting a better approximation of the critical value associated with a desired α,[21] but this is left to future research.

D. Comparison to Other Chemistry/Physics Stochastic Modeling Approaches. The use of time series to calibrate coarse-grained models has been introduced in the past.[27-29,33] One salient feature that distinguishes the approach used here and in related works[9,14,15,34,35] from previous efforts is that ensemble averaging is not appealed to in order to get either an effective force or a diffusion (or damping) coefficient. The methods used do not need to assume that the "orthogonal" degrees of freedom are sampled "ergodically" in a single times series (the entire collection of observed time series data does not need to ergodically sample phase space either for that matter). These types of approaches are often employed in physical-sciences-motivated/based. For example, in refs 28 and 29, the authors used estimates of the free energy and the definition of the friction coefficient depending on various quantities obtained using ensemble averaging; the hope was to calibrate low-dimensional mesoscopic models (capable or reaching larger time scales) using MD data on short time scales. The goal of this work is similar in nature but differs in several other respects, which are discussed below. It should be noted that earlier coarse-graining approaches, for example, refs 27−29, did not have the luxury of some of the recent advances in computational and theoretical mathematical statistics used here. Now that such tools are available, it is possible to ask new questions and seek higher resolution from models and at the same time pose new questions relevant to probing single-molecule dynamics to researchers in mathematical statistics which potentially open new avenues of research for both communities.

Some other distinguishing features of the statistical inference approaches employed are as follows: (i) they are applicable to single-trajectory realizations and hence do not require one to ensemble average (or aggregate time series from multiple realizations); (ii) they do not partition state space in a finite number of discrete states;[33,36] and (iii) they do not require that the underlying time series come from a stationary process. A single sample path can be used to estimate the coefficient functions associated with the SDE by utilizing expansions of the transition density[18] (i.e., the solution to the Fokker−Planck equation associated with the SDE). The maximum likelihood type fitting procedure does not assume that time averaging is equivalent to ensemble averaging, and this feature can be important in biomolecular systems where there are slow degrees of freedom that are not adequately sampled in a single trajectory. For example, if different regions of phase space exhibit substantially different "ruggedness"[13] in the energy landscape (or alter the kinetics in some other nontrivial fashion depending on the resolved coordinate), it is possible that this feature can complicate using a single characteristic relaxation time to construct predictive discrete-state Markov models.[36] The ideas outlined here can potentially be used in constructing discrete state partitions[36] without requiring an excessive amount of temporal coarse-graining by extracting a local effective damping or friction coefficient from a single sample path. If different paths exhibit different damping rates, but are otherwise similar in regards to the criterion defining a discrete state, then one might want to reconsider the definition of a given discrete state.

If one is optimistic, one may hope that the ergodic sampling occurs, that is, "conformational heterogeneity", and other unresolved degrees of freedom average out over the time scale of interest. Suppose a resolvable coordinate is a "good reaction coordinate" in the sense discussed in ref 37, namely, the quantity is the slowest coordinate of the system and all orthogonal degrees of freedom rapidly sample phase space over the time scale that the simulations are made over, and that the details of the orthogonal coordinates are not physically important. In such a case, there is hope for deriving simple effective evolution equations describing a more complex dynamical system. One can attempt to construct a single equation from observed data or from first-principles considerations. However, if a physically relevant lurking degree of freedom is present, eventually, this single equation using only the reaction coordinate will be rejected at a fine enough temporal resolution. This rejection can occur even if the system does not "hop" out of the free-energy well. Cases where the rejection occurs because of subtle changes in the dynamics over time due to the coupling with the unresolved coordinate are demonstrated in the Results and Discussion. Note that this type of analysis can be used to help in quantitatively determining if ergodic sampling is practically occurring over the time scale which observations are made.

In addition, it is stressed that a single goodness-of-fit test statistic can be constructed using a single time series. That is, even if only a small number (or one for that matter) of noisy trajectories are observed, one can still assess the fitted model without having to resort to ensemble averaging or an ergodic sampling assumption. One can exploit the time ordered samples associated with a single discretely observed path to test the statistical validity of an assumed model. The test used, that of Hong and Li,[19] simultaneously checks for both the shape of an assumed distribution and the correlation structure in the (generalized) residuals; the latter can be important in detecting kinetic surrogate model misspecifications. Standard tests, based on the well-known χ^2 statistic,[38,39] have difficulty in simultaneously checking for both the distribution shape and the temporal correlation structure of residual type quantities. As discussed later, Bayesian methods can only compare two or more assumed models;[36] if an unanticipated time correlation occurs in the residuals, this feature would be difficult to assess using a Bayesian framework.[36,39] Recall that the test used here does not depend on any type of stationarity or ergodic sampling assumptions in the underlying time series. This is particularly relevant to cases where the transition density is estimated nonparametrically;[33] in the nonstationary or nonergodic regime, nonparametric estimates are problematic. If one does indeed have ergodic sampling and a stationary time series, recent mathematically rigorous testing procedures which test the Markov assumption nonparametrically can be attempted.[21,40]

Furthermore, it is stressed that the methods shown here are data-driven in nature, that is, a model is fit to the observed output. The SDE coefficient functions estimated may not have a readily transparent physical interpretation. Appealing to the Einstein relation along with the particular "local overdamped" diffusion model structure to fit the local time series data helps one in loosely understanding the output,[9,35] but one does not need to assume such relationships. Other SDEs can be entertained for approximating the dynamics. Possessing a reliable means for assessing coarse-grained models (and quantifying the uncertainty in parameter estimates) is an important component to coarse graining and can help in assessing if a (local) Einstein-type relation is justified for the observed data. The goodness-of-fit appealed to[19] attempts to qualitatively test such assumptions given the data. However, as discussed in greater detail in ref 22, the SDE models fit to the data only provide a surrogate for the dynamics implied by one observed path. The fitted surrogate SDEs can then be used to approximate different random variables (point values or paths), and these surrogate random variables can be used with the established machinery of statistical mechanics to provide unambiguous estimates of traditional thermodynamic or kinetic quantities (e.g., the potential of mean force along a reaction coordinate[35]). If ergodic sampling does not occur on the time scale of the observations, the collection of the SDE model can also be used to infer how much conformational heterogeneity influences the observed data and makes fuller use of the time-ordered data.[8,9,22,35]

III. Results and Discussion

Figure 2 plots sample paths associated with Cases 1A and B, and Figure 3 plots paths associated with Cases 2A and B (the cases referred to are described in section II.B). In all cases reported in this article, Brownian sample paths are drawn and used along with eq 2 to evolve the y process. The same Brownian paths are used in all cases; therefore, the differences between the paths of the resolved process in the various cases are due to the different x evolution rules. This helps in minimizing sources of variation and facilitates comparing the estimation and hypothesis test results presented later. In Figures 2 and 3, gray lines are used to plot paths where the dynamics in x are "turned off", that is, the value of x is frozen to be 0, and the same Brownian paths are used again along with the SDE in eq 2. These cases are labeled as y_{ref}. The differences $y - y_{\mathrm{ref}}$ (where paths are matched by the underlying Brownian path) are plotted in panels C and F of Figures 2 and 3. In all samples analyzed in this paper, the resolved coordinate does not hop over the large energetic barrier near $y = 0$ in Figure 1A.

In Case 1, evolution of x causes a fairly dramatic change in the shape of $U(x, y)$, but the changes are not excessively pronounced when one observes sample paths. The most salient change in $U(x, y)$ that one can visually inspect in the time series is that the well bottom shifts slightly (the average value appears to slowly change). The trends in $\sigma^y(x, y)$ in Case 2 are associated with changes in the effective friction, and these changes are even subtler than those in Case 1, as is evident by panels C and F of Figure 3. Note that inertial versus diffusive dynamics in x dramatically influence the features in the x sample paths, but the influence that the (assumed unresolvable) x coordinate has on y does not result in dramatic differences in the global shape of the y sample paths. This is due mainly to the time scale separation imposed and the fact that the x process has the same general trend in all cases. However, it is possible that hypothesis tests can detect more localized signatures of the slow non-Markovian noise source induced by x. We return to this discussion later.

A. Quantifying Different Sources of Variation/Uncertainty. The previous figures displayed a handful of representative trajectories. In all results that follow, $\mathscr{B} = 100$ Brownian paths are simulated for a total of 1 ns, and discrete observations are made every 0.2 ps. The $N_W = 10$ local windows used for both estimation and goodness-of-fit testing each contain 500 temporal observations. With this specification, the $\{\{\hat{\theta}_{N_i}^{(\mathscr{B}_j)}\}_{N_i=0}^{N_w}\}_{\mathscr{B}_j=1}^{\mathscr{B}}$ can be computed.

The lines with symbols in Figure 4 correspond to averaging the estimated A and C parameters over the $\mathscr{B}$ samples in each local window; these parameters aim at locally approximating $\nabla_y U$ and σ^y at y^{BASE}. The solid lines correspond to adding the empirically determined standard deviation to the mean of the estimated A and C found in the $\mathscr{B}$ samples. The cases labeled as "Ref" in these figures correspond to using the average θ value estimated in time window 0 for each of the 100 samples in both Cases 1B and 2B. These Case 1B and 2B parameters associated with window 0 are used to formulate the null vector θ_0 discussed at the end of section 2.C. The null vectors are then used along with the same 100 Brownian paths to simulate the SDE corresponding to eq 2. This gives the ideal surrogate case, that is, the full nonlinear structure of eq 3 is not used to evolve y, but instead a surrogate model is used. From this data, the average and standard deviation of the A and C parameters over the different time windows are determined. The "Ref" data are used primarily to approximate how much uncertainty is associated with a finite number of time series observations and differs in meaning from the "ref" label used in the previous section (the "Ref" data was also used to verify that the bias introduced by finite sample sizes and the approximate transition density were

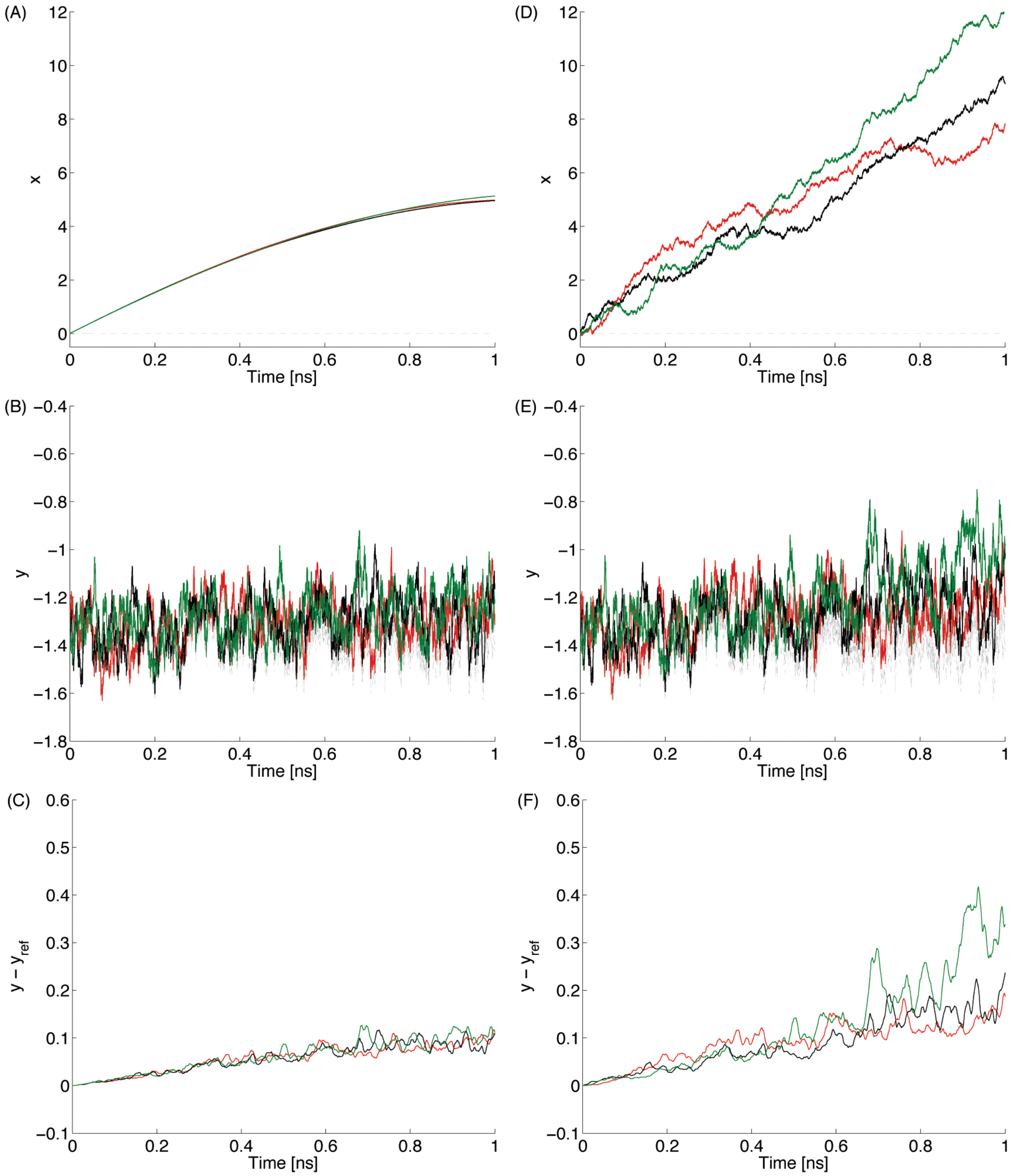

Figure 2. The colored trajectories represent four sample paths for Case 1A (left panels A−C) and Case 1B (right panels D−E) described in eq 3. The same Brownian realizations were also used to drive a reference system where x was forced to be frozen at 0. These paths are colored gray, referred to as “ref ”, and the difference between y and y_{ref} shows how the evolving unresolved coordinate x modulates the dynamics in a trajectory-wise sense.

small); the same null vectors are used later to generate M independent sample paths (using new Brownian paths) for hypothesis testing purposes.

In all cases, the unresolved coordinate x modulates the dynamics and hence changes the effective force (Case 1) or friction coefficient (Case 2) experienced by the y particle. The evolution of x takes two different forms. In the inertial A cases, the associated sample paths of x are much smoother, whereas in the B cases, the distribution of the x process is broader for larger times due to the diffusive dynamical rules used to evolve the x particle. This has a fairly small, but noticeable, effect in the parameter uncertainty quantified using the empirical second moments of the θ parameter estimated in each local window in both Cases 1 and 2. In all cases, the first two moments of the estimated parameters reach a value substantially different than those associated with the “Ref” case around time window 5. However, note that the surrogates permit a linear trend in both of these functions; therefore, this does not imply that the surrogate models are invalid. It simply shows that the constant associated with the point y^{BASE} has changed appreciably by time window 5 in relation to the resolution that one can obtain with this time series sample size and assumed model.

Comparison of the width of the estimated parameters in the Ref simulations to those in the other cases allows one to roughly determine if “conformational heterogeneity” is causing a statistically detectable change in the dynamics. For example, if the width in the estimated parameters is substantially larger than that in the Ref case, this might be the case. Increased uncertainty might also be a result of a crude surrogate model. The latter can be tested for, and this is demonstrated in the next subsection where the validity of the local surrogates over longer time scales is investigated.

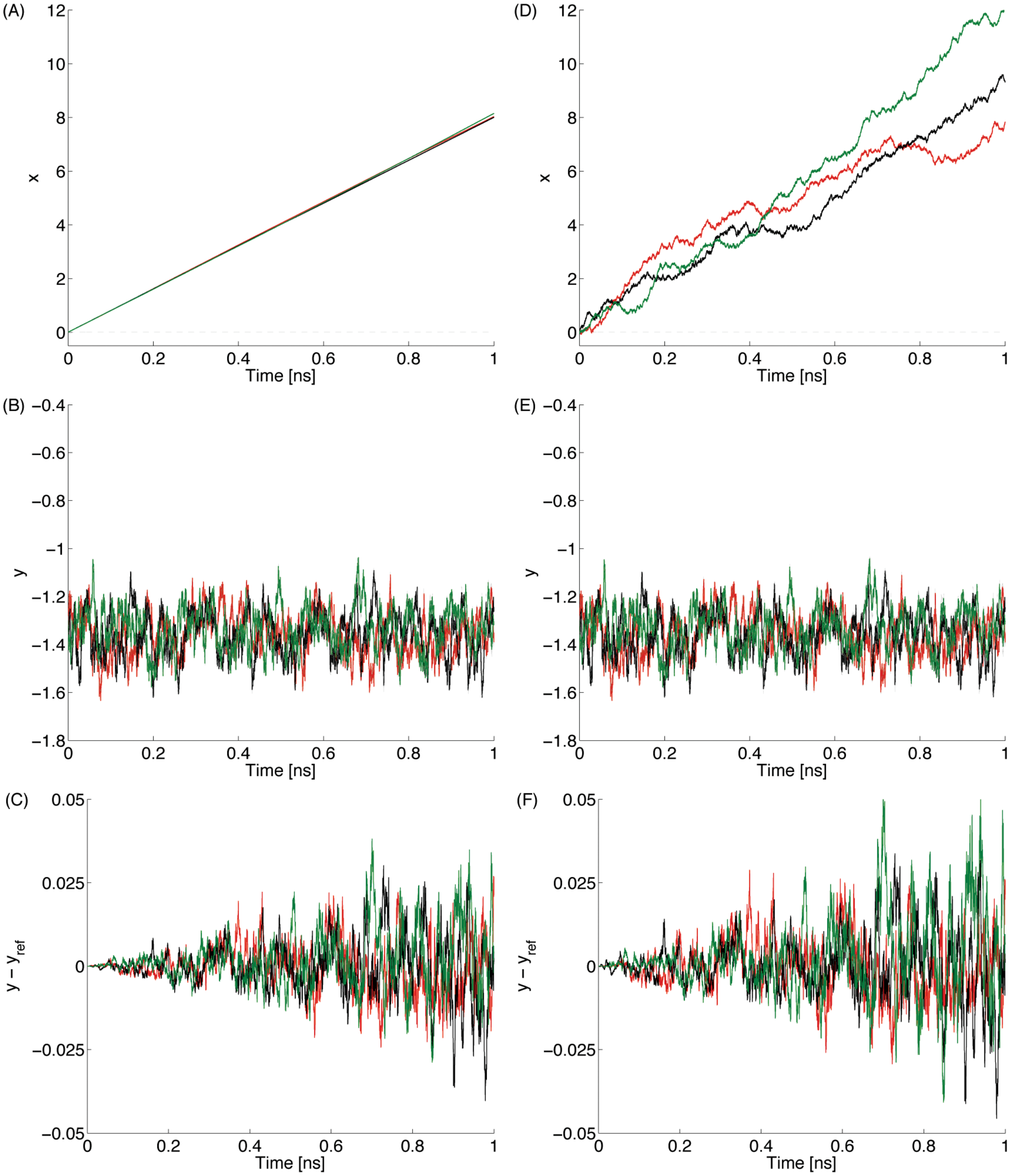

Figure 3. Similar to Figure 2, except for Case 2. Note that the differences in panels C and F here are fairly small in relation to both the typical fluctuations associated with the process and the corresponding panels in Figure 2.

In more realistic settings, the unresolved coordinate will introduce additional complications beyond what was considered here. Namely, the initial condition will not be the same for each realization, and the associated evolution for different initial values will not likely have a uniform trend as was enforced here. Also x's "memory" and the dependence of the dynamic rules on the initial condition may be much more complex.[7] Again, hypothesis testing methods show great promise in detecting these situations, and the methods presented here attempt to detect any substantial departure from an assumed surrogate model (e.g., non-Markovian versus Markovian dynamics). There does not need to be a persistent trend in multiple trajectories. The pathwise tests used do not require a distribution of trajectories; the information in each path realization is used along with the surrogate model structure to formulate a test statistic, and the validity of the assumed model can be assessed on a path by path basis to determine if the various transitions observed are consistent with the assumed or fitted model.

B. Quantifying the Influence of Unresolved "Conformational Heterogeneity". The utility of using a collection of surrogate models in various sampling settings has been discussed elsewhere.[8,9,14,15,17,35] The basic ideas are summarized in this subsection. A single (noisy) time series realization $\{y_i\}_{i=0}^{N}$ typically samples a small portion of phase space. By fitting the local surrogate models, one transforms the time series into a collection of parameters $\hat{\theta}_{N_i}$ which have a loose physical interpretation in terms of effective force and friction (see eq 4). Other works have discussed some subtleties that can arise when unresolved degrees of freedom are coupled to the resolved coordinate and modulate the dynamics either in a more traditional thermodynamic fashion[13] or influence kinetic properties by a dynamical mechanism; for example, see ref 24. Methods for detecting statistically significant changes in the dynamics of the resolved coordinate can help one in quantitatively testing various scientific models or hypotheses.

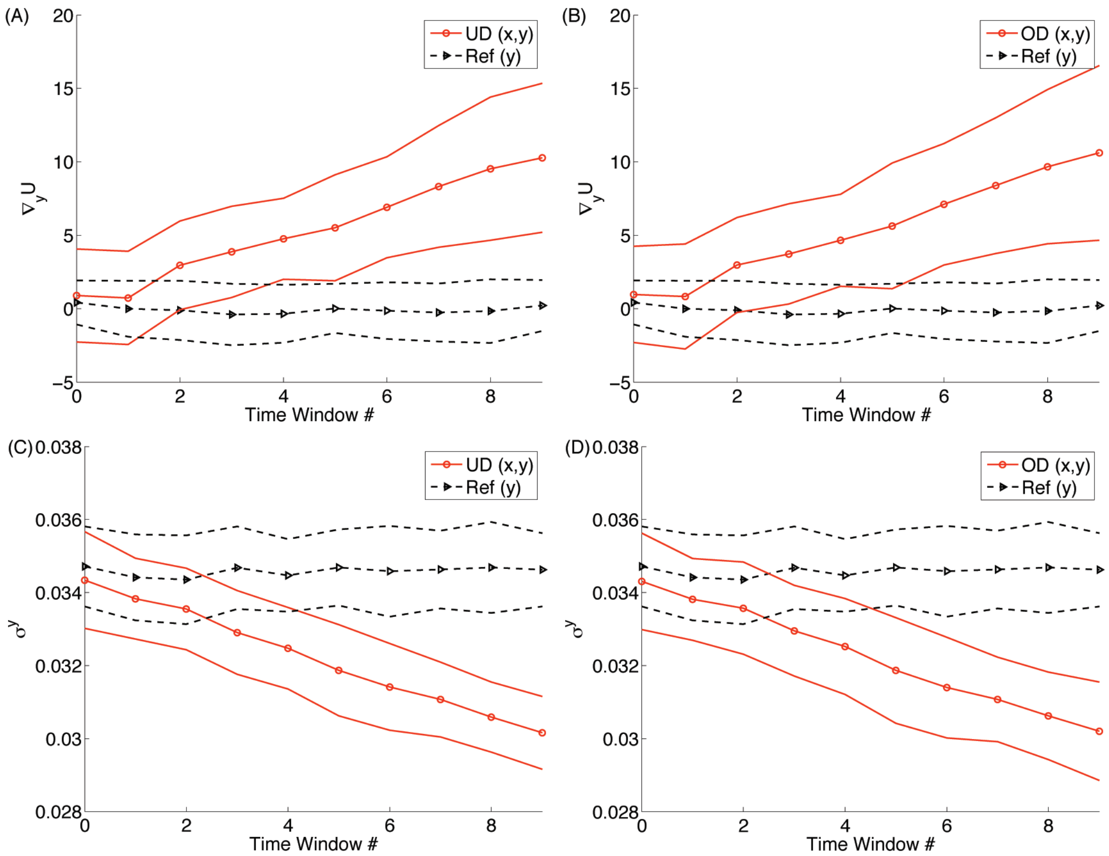

Figure 4. The average force ($\nabla_y U(x, \psi)$) and noise function ($\sigma^y(x, \psi)$) estimated using the expression in eq 3 in 10 different local time windows. The aforementioned quantities are approximated using A and C, respectively, when y^{BASE} is set to ψ. These estimates are plotted using symbols. The lines connecting the symbols are only to guide the eye, and the lines without symbols correspond to the empirical average value observed ± the sample standard deviation (taken over 100 paths). The data-generating processes studied are the full two-dimensional nonlinear SDEs in eq 2, and the "Ref" SDE in the legend is a one-dimensional diffusion simulated for the same reporting time as the observed data using parameters estimated from time window 1 data; this provides one with an idealized uncertainty estimate of the parameters.

It is stressed that once a new time series is in hand, a new sequence of surrogate parameters can be estimated. This estimated sequence summarizes some of the dynamical information associated with the region of phase space explored in each trajectory (the unresolved coordinate is likely substantially different in each time series/sample path). By analyzing the diversity in the $\{\{\hat{\theta}_{N_i}^{(\mathcal{B}_i)}\}_{N_i=0}^{N_W}\}_{\mathcal{B}_i=1}^{\mathcal{B}}$ population, one can learn about the dynamics associated with different portions of phase space without excessively coarse graining the data. Such an approach can also be helpful in cases where a continuum of states is believed relevant to single-molecule experiments.[26]

C. Goodness-of-Fit Tests. Here, the two θ_0 associated with the "Ref" cases described section III.A are used along with the assumed SDE structure in eq 2 to formulate the null hypothesis in Cases 1 and 2. The data in each different time window is used along with the single null parameter vector to construct a new time series, that is, the sequence $\{y_i\}_{i=0}^{N}$ is transformed to $\{z_i\}_{i=1}^{N}$ using the probability integral transform (PIT).[19] If the null is correct for all observations, then the PIT series consists of entries all coming from a single uniform distribution $U[0, 1]$, and each observation is statistically independent of the other entries. These properties hold regardless of the correlation structure in the original $\{y_i\}_{i=0}^{N}$ time series and are valid for both stationary and nonstationary time series if the assumed null model generates the data (these properties do not require large sample limits). Goodness-of-fit tests can then be applied (either globally[14] or locally[8,9]) to the PIT time series.[19] With this transformation, one can quantitatively determine if the surrogate model is statistically acceptable by computing a test statistic. The test statistic, denoted by Q, is constructed in an attempt to detect any departure from the PIT sequence that holds under the null. As we will discuss later, this is both a strength and a potential weakness of this PIT-based "omnibus" procedure[19−21] (so-called because it simultaneously checks the fit to the $U[01]$ and the independent assumption.[19]).

Before discussing this issue further, the empirical cumulative distribution of the test statistic computed in four time windows is plotted for the various cases studied in Figure 5. The smooth solid lines plot the null distributions determined by the Monte Carlo procedure discussed in the Background and Methods. The staircase plots are the empirical cumulative distribution function of the test statistic computed using the PIT sequence corresponding to the local windows (which are varied) and the null vector θ_0 associated with window 0 (one single fixed vector for Case 1 and another for Case 2). Recall that θ_0 contains information about a surrogate SDE calibrated when x is close to 0 and $y = \psi$. The physical interest here is in determining how substantially the changes in x affect y's dynamics under the null (near the well bottom corresponding to $x = 0$). The shaded region corresponds to the critical value Q associated with the approximate 10% significance level, that is, any test statistic to the right of the shaded region is rejected using the approximate Type I error rate, α = 0.10. The percentage of models rejected can be determined by inspecting the value of the empirical cumulative distribution function at the critical value and subtracting this quantity from unity. As time proceeds, the percentage of rejected models clearly increases dramatically. Note that the changes in the dynamics of the sample paths in

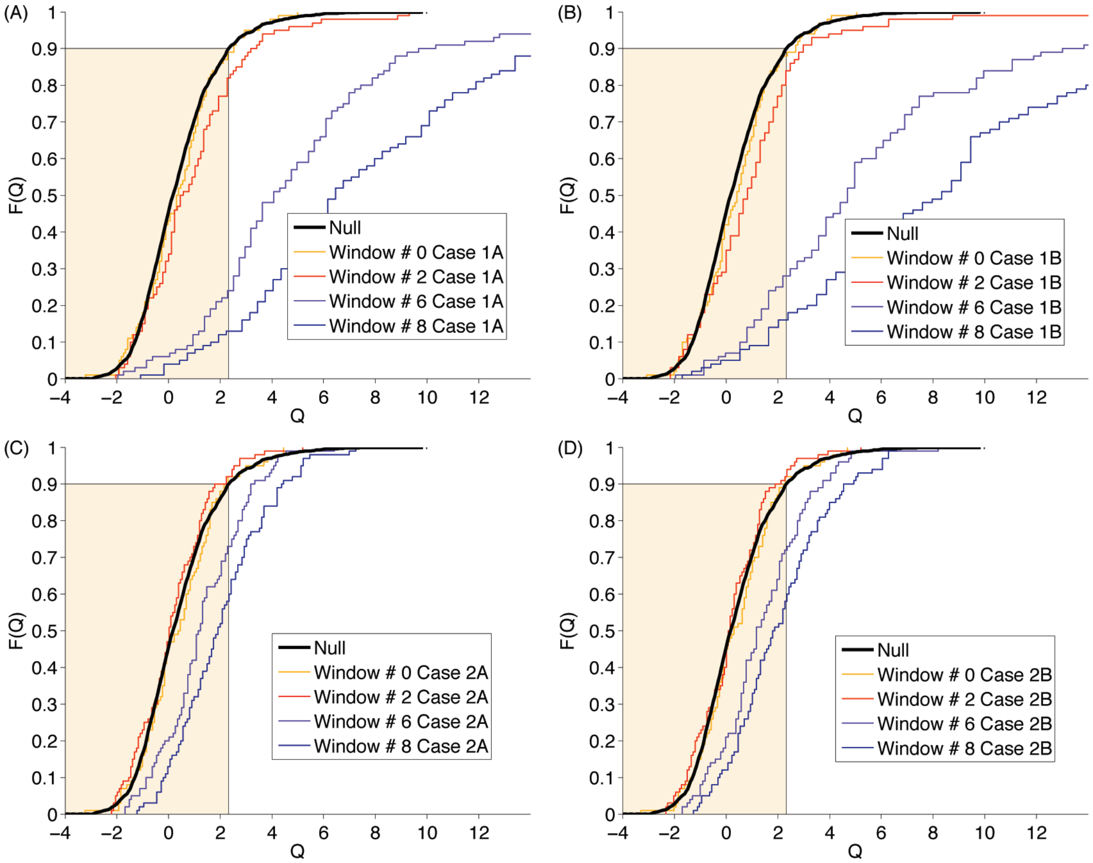

Figure 5. The empirical cumulative distribution function (staircase plots) of the test statistic Q for the various cases at four time windows along with the empirically determined null distribution (thick smooth curve). The shaded region corresponds to an approximation of the 10% significance level, that is, any test statistic greater to the right of the shaded region is rejected using an approximate Type I error rate, $\alpha = 0.10$.

Figures 2 and 3 are very subtle in time window 6, but there is strong rejection at the 10% level. The strength of this procedure is that the validity of a single dynamical model can be determined without requiring ergodic sampling. If an unexpected θ value is estimated in a surrogate, then this type of procedure can be useful to determine whether parameter variation is due to inherent uncertainty due to finite sample sizes or if something more systematic is lurking in the background. This procedure can be done using only one observed time series.

Note that the details of the evolution rules (with or without inertia) did not significantly affect the hypothesis test results reported. The test employed was not able to distinguish between different non-Markovian (dynamical) effects associated with a slowly evolving unresolved coordinate coupled to the resolved coordinate in the sample sizes considered here. The smoothness of the unresolved x paths did not influence the test (in each local window, x did not change appreciably from its initial value in the time window). In the inertial cases, the unresolved x coordinate made several (small) nearly constant deterministic perturbations to y's evolution rules. In the diffusive case, the discrete increments of a diffusive x type process were noisier and less correlated than comparable inertial cases (in our Case B models, the increments of x were independent and identically distributed (i.i.d.) by construction). However, the extra correlation in the inertial cases did not help in identifying the non-Markovian noise source because the change it induced in y's evolution rules was too small to be detected by the test applied. Recall that the test used is "omnibus",[19] and the price one often pays for this feature is power.[20,21] Future extensions of recent hypothesis tests to accommodate nonstationary data would likely assist single-molecule data analysis and may be able to help one gain power in detecting specific features, such as slow lurking coordinates modulating the dynamics in a "non-Markovian" fashion (e.g., the velocity of the unresolved coordinate is important), but this seems to be a formidable technical challenge. However, note that the test used[19] can detect other subtle features. Over time, x changes appreciably, and once this (continuous) change in either the effective force or friction is statistically significant, then the PIT based test is able to determine this subtle regime shift that would likely be missed by standard methods. This suggests that a more sudden regime switch (in either noise or force) induced by an x type coordinate would also be readily detected by this test.

It should be noted that Bayesian methods encounter technical challenges in making assessments of the overall goodness-of-fit. The correlated time series situation is more challenging (in either a stationary or nonstationary regime) in a Bayesian framework.[39] Bayesian methods can readily compare different dynamical models[36] but do not naturally detect factors outside of the model space considered. Misspecifications induced by a slowly evolving x type coordinate coupled to the y dynamics would likely be missed by a Bayesian analysis. If one has the luxury of observing every degree of freedom, as one does in molecular simulation, then creating an effective partitioning of state space to enforce a Markovian dynamical model is attractive.[36] However, if there are unresolvable degrees of freedom (the case common to experiments), making subtle, but physically important, changes to the dynamics due to a coupling to the resolvable coordinate(s), then this poses a serious problem to a Bayesian method. This is one area where frequentist methods are particularly attractive. The suitability of an assumed dynamical model can be compared directly to the observed data (even if the data set is only one or two times series), whereas a Bayesian would either have to compare different models in a limited class or somehow assess the prediction of the model given the time correlated data. Both aforementioned tasks are seriously complicated by heterogeneity in the unresolved coordinate and/or a limited amount of time series sample paths. Detecting the influence of the resolved coordinate's velocity on its own dynamics (an easier problem than the one considered here) is also problematic for a Bayesian method, whereas various frequentist methods can readily detect this type of model misspecification[8] without having to directly address the technical challenges associated with a hypoellpitic diffusion[41] or discretizing state space.[33,36] If one is willing to take the surrogate models on faith (or develop a reliable goodness-of-fit testing procedure), empirical or hierarchical Bayesian methods do offer great promise in describing single-molecule data where an unresolved degree of freedom is present, as discussed in ref 14. Another item worth noting is that an accurate estimate of the critical value associated with a desired α may not be practically relevant to the problem at hand (i.e., every proxy will be rejected given enough evidence from the true process). Putting a quantitative handle on "practical significance" is difficult in a frequentist framework, whereas Bayesian inference methods can often more easily treat this type of problem by utilizing prior information (again, this assumes the researcher already has knowledge of an accurate surrogate model). However, frequentist PIT-based methods show great promise in providing quantitative tools/metrics for assessing the fit and providing diagnostics[19] of parametric models in various situations where the time correlation in the data is fairly complex and the time series is nonstationary. Even if the process is technically stationary, some time series realizations can be "practically" nonstationary, meaning the process does not "ergodically" sample phase space in the observed data (the PIT-based methods are relevant here too).

IV. Conclusions and Outlook

This article was concerned with using a surrogate model structure and frequentist hypothesis tests to detect if an unresolved coordinate makes statistically significant changes to the dynamics of a resolved particle. The omnibus goodness-of-fit test studied[19] was used to test if features of an unresolved coordinate coupled subtly to the resolved coordinate could be detected in moderate sample sizes in various controlled simulations. The unresolved coordinate was allowed to evolve in the inertial and diffusive regimes. Furthermore, the unresolved particle was not associated with a stationary distribution; the unresolved particle increased at a roughly constant rate over the time scale of the simulations.

The tests were able to detect subtle (continuous) changes induced by the coupling in cases where the unresolved coordinate changed appreciably from its initial value. Previous studies highlighting thermodynamic[8] and kinetic[22] consequences of neglecting subtle features in all-atom MD simulations where "conformational fluctuations" have been suggested[12,24] to be important motivated aspects of this research. The details of how the unresolved coordinate x reached the new state did not significantly affect the performance of the goodness-of-fit test. It can be loosely stated that the time scale separation did not allow one to "dynamically detect" the presence of a non-Markovian noise source for the sample sizes and surrogate models considered (i.e., the "memory" of the slow x coordinate could not be detected in short time windows where x changed by a small amount regardless of how x or y evolved). In ref 8, it was demonstrated that the same test could dynamically detect the presence of certain non-Markovian noise sources when the unresolved degree of freedom was associated with "faster" time scales than the resolved particle. For example, the importance of including the unobserved velocity of the resolved particle in the dynamical model can be handled by the methods shown here and elsewhere.[8] Velocity correlation of the resolved coordinate is a type of memory that can often be accounted for using various methods that depend only on past temporal values of the resolved coordinate.[26] However, in cases where the unresolved coordinate is slow, the memory kernel is much more complex; for example, it can depend on the current state and can also depend in a complicated way on "orthogonal" coordinates (ref 8 discusses this issue as related to surrogate modeling and provides many references to established literature on this general topic). In cases where complications such as poor sampling of physically relevant orthogonal coordinates[8] occur on the time scale that one can experimentally resolve or in cases where one encounters a continuum of physically relevant states,[26] the surrogate modeling ideas[9,14,17] combined with functional data analysis[16] show promise for constructing computational tools that can help in summarizing (and making predictions from[9,22,35]) the rich information contained in the noisy single-molecule trajectories.

The test considered is fairly general; it only requires the evaluation of conditional densities and can readily handle nonstationary data. Different tests can possibly perform better and detect dynamic signatures of a slow non-Markovian noise source. If it turns out that the suggestion that inertial/dynamical contributions are important to enzymology and/or protein dynamics (see, for example, refs 10−12, 24), the development of new more powerful tests can be important in analyzing single-molecule data. However, if it turns out that inertia is not incredibly relevant to the physical process of interest and all momentum is highly dissipated over nanosecond time scales in systems with high energy barriers,[13] then the results shown here are even more promising (methods indirectly detecting signatures of inertial memory would not be physically important in such applications). The modeling and tests presented can be used to help in mapping out the energy landscape and effective position-dependent friction. With such information (and if inertia is unimportant), one can entertain using established tools related to transition-state theory to approximate various rates associated with biocatalysis.[42] If the conformational change occurs on time scales inaccessible to direct simulations, then proposed kinetic mechanisms should be tested with some quantitative criterion like a mean first passage time; see, for example, ref 13. With the wealth of simulation and experimental data available, it is important to develop quantifiable criteria to assess any proposed mechanism and select the simplest model describing the data consistent with the observations. Note also that there is no "one size fits all" test, and the criterion used to select suitable models depends heavily on the application at hand, but frequentist pathwise tools presented and discussed here can be used to get a better quantitative handle on coarse-grained model construction.

It should be noted that surrogate SDEs can be used in applications beyond approximating the effective force and noise of complicated processes in different regions of phase space. The collection of physically interpretable parameters associated with the surrogates can be used to characterize nanoscale systems,[17] identify events in noisy experimental time series induced by unresolved conformational coordinates,[15] and generate surrogate sample paths associated with nonequilibrium simulations[35] and experiments.[9] These models do not excessively coarse grain over "conformation" space. The applicability of a diffusive process does require some modest temporal coarse graining, but as shown here and elsewhere,[8,9] quantitative goodness-of-fit tests exist if one uses frequentist methods to temporally coarse grain a continuous (time and space) model. Approaches that minimize coarse graining remain in the spirit of single-molecule experiments and can make better use of the information-rich data coming from new experiments.[1−6] Studies using the collection of surrogate models in a transition path sampling context[23,43] would also be interesting avenues of future work.

Acknowledgment. The author extends his gratitude to an anonymous referee for helpful comments and references. This research was supported in part by the Director, Office of Advanced Scientific Computing Research of the U.S. Department of Energy under Contract DE-AC02-05CH11231.

Appendix

The evolution equations and parameters specifying the *x* dynamics are reported below. Recall that the label A is used to denote the cases where inertia of the unresolved coordinate is important, and for those labeled B, *x* evolves diffusively.

A Cases:

$$\begin{aligned}
&\mathrm{d}v_t^x = 1/M((-\gamma_{\mathrm{A}}^x v_t^x - \nabla_x U(x_t, y_t))\mathrm{d}t + \sqrt{2k_{\mathrm{B}}T\gamma_{\mathrm{A}}^x}\mathrm{d}\tilde{B}_t)\\
&\mathrm{d}x_t = v_t^x\mathrm{d}t\\
&v_0 = \mathscr{V}^{\mathrm{IC}} \qquad x_0 = 0 \qquad\qquad y_0 = \psi
\end{aligned} \tag{5}$$

B Cases:

$$\begin{aligned}
&\mathrm{d}x_t = 1/\gamma_{\mathrm{B}}^x\mathscr{F}\mathrm{d}t + \sqrt{2k_{\mathrm{B}}T/\gamma_{\mathrm{B}}^x}\mathrm{d}\tilde{B}_t\\
&x_0 = 0 \qquad y_0 = \psi
\end{aligned} \tag{6}$$

In the above, $\tilde{B}_t$ represents another standard Brownian motion process (statistically independent of B_t). All quantities are reported as dimensionless. M is the mass of the x particle (the value of 5×10^5 was used throughout), $\mathscr{V}^{\mathrm{IC}}$ is the initial velocity (=8), ψ is the point corresponding to the well minimum of $U(x, 0)$, and it takes a value of ~1.36, γ^x is the constant friction coefficient used (the subscript on this denotes different values are used in different cases, $\gamma_{\mathrm{A}}^x = 10$, $\gamma_{\mathrm{B}}^x = 500$), $\mathscr{F}$ represents a constant force (it was selected so as to force $1/\gamma_{\mathrm{B}}^x\mathscr{F} = \mathscr{V}^{\mathrm{IC}}$, that is, the velocity is comparable to the velocity of the inertial B cases).

These parameters along with those specifying eq 2, namely, $\alpha = 10$, $\kappa = 1.25$, $\phi = 2$, and $\zeta = 0.04$, completely characterize the dynamics. Note that in Cases 1A, 1B, 2A, and 2B, the equations are set up to have similar dynamics at time zero. As time evolves, x evolves at a roughly constant rate but modulates the dynamics in a subtle fashion (there is no sudden regime switch).

References and Notes

(1) Fuller, D. N.; Raymer, D. M.; Kottadiel, V. I.; Rao, V. B.; Smith, D. E. Single phage T4 DNA packaging motors exhibit large force generation, high velocity, and dynamic variability. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16868–16873.

(2) Walther, K. A.; Gräter, F.; Dougan, L.; Badilla, C. L.; Berne, B. J.; Fernandez, J. M. Signatures of hydrophobic collapse in extended proteins captured with force spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7916–21.

(3) Ke, C.; Humeniuk, M.; Gracz, H. S.; Marszalek, P. E. Direct measurements of base stacking interactions in DNA by single-molecule atomic-force spectroscopy. *Phys. Rev. Lett.* **2007**, *99*, 018302.

(4) Greenleaf, W. J.; Frieda, K. L.; Foster, D. A. N.; Woodside, M. T.; Block, S. M. Direct observation of hierarchical folding in single riboswitch aptamers. *Science* **2008**, *319*, 630–633.

(5) Henzler-Wildman, K. A.; et al. Intrinsic motions along an enzymatic reaction trajectory. *Nature* **2007**, *450*, 838–844.

(6) Hodges, C.; Bintu, L.; Lubkowska, L.; Kashlev, M.; Bustamante, C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science* **2009**, *325*, 626–628.

(7) Lu, H. P.; Xun, L.; Xie, X. S. Single-molecule enzymatic dynamics. *Science* **1998**, *282*, 1877–1881.

(8) Calderon, C. P.; Arora, K. Extracting kinetic and stationary distribution information from short md trajectories via a collection of surrogate diffusion models. *J. Chem. Theory Comput.* **2009**, *5*, 47.

(9) Calderon, C. P.; Harris, N.; Kiang, C.-H.; Cox, D. D. Quantifying multiscale noise sources in single-molecule time series via pathwise statistical inference procedures. *J. Phys. Chem. B* **2009**, *113*, 138.

(10) Li, Y.; Qu, X.; Ma, A.; Smith, G. J.; Scherer, N. F.; Dinner, A. R. Models of single-molecule experiments with periodic perturbations reveal hidden dynamics in RNA folding. *J. Phys. Chem. B* **2009**, *113*, 7579–7590.

(11) Miyashita, O.; Onuchic, J. N.; Wolynes, P. G. Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12570–12575.

(12) Arora, K.; Brooks, C. L., III. Large-scale allosteric conformational transitions of adenylate kinase appear to involve a population-shift mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 18496–18501.

(13) Pisliakov, A. V.; Cao, J. ; Kamerlin, S. C. Warshel, A. Enzyme millisecond conformational dynamics do not catalyze the chemical step. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 17359−17364.

(14) Calderon, C. P. On the use of local diffusion for path ensemble averaging in potential of mean force computations. *J. Chem. Phys.* **2007**, *126*, 084106.

(15) Calderon, C. P.; Chen, W.-H.; Harris, N.; Lin, K. J.; Kiang, C.-H. Analyzing DNA melting transitions using single-molecule force spectrscopy and diffusion models. *J. Physics: Condens. Matter* **2009**, *21*, 034114.

(16) Ramsay, J.; Silverman, B. W. *Functional Data Analysis*; Springer-Verlag: New York, 2005.

(17) Calderon, C. P.; Harris, N.; Kiang, C.-H.; Cox, D. D. Analyzing single-molecule manipulation experiments. *J. Mol. Recognit.* **2009**, *22*, 356.

(18) At-Sahalia, Y.; Fan, J.; Fan, J. Maximum-likelihood estimation of discretely-sampled diffusions: A closed-form approximation approach. *Econometrica* **2002**, *70*, 223–262.

(19) Hong, Y.; Li, H. Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *Rev. Financ. Stud.* **2005**, *18*, 37–84.

(20) Ait-Sahalia, Y.; Fan, J.; Peng, H. Nonparametric transition-based tests for jump-diffusions. *J. Am. Stat. Assoc.* **2009**, *104*, 1102–1116.

(21) Chen, S. X.; Gao, J.; Tang, C. Y. A test for model specification of diffusion processes. *Ann. Stat.* **2008**, *36*, 167–198.

(22) Calderon, C. P. A data-driven approach to decomposing complex enzyme kinetics. *Phys. Rev. E* **2009**, *80*, 061118.

(23) Metzner, P.; Schutte, C.; Vanden-Eijnden, E. Illustration of transition path theory on a collection of simple examples. *J. Chem. Phys.* **2006**, *125*, 084110.

(24) Vendruscolo, M.; Dobson, C. M. Dynamic visions of enzymatic reactions. *Science* **2006**, *313*, 1586.

(25) Tripathi, S.; Portman, J. J. Inherent flexibility determines the transition mechanisms of the EF-hands of calmodulin. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 2104–2109.

(26) Taylor, J. N.; Makarov, D. E.; Landes, C. F. Denoising Single-Molecule FRET Trajectories with Wavelets and Bayesian Inference. *Biophys. J.* **2010**, *98*, 164−173.

(27) Gardiner, C. W. *Handbook of Stochastic Models*; Springer-Verlag: Berlin, Germany, 1985.

(28) Balsera, M.; Stepaniants, S.; Izrailev, S.; Oono, Y.; Schulten, K. Reconstructing potential energy functions from simulated force-induced unbinding processes. *Biophys. J.* **1997**, *73*, 1281.

(29) Burykin, A.; Kato, M.; Warshel, A. Exploring the origin of the ion selectivity of the KcsA potassium channel. *Proteins* **2003**, *52*, 412–426.

(30) Kloeden, P.; Platen, E. *Numerical Solution of Stochastic Differential Equations*; Springer-Verlag: New York, 1992.

(31) Calderon, C. P. Fitting effective diffusion models to data associated with a "glassy potential": Estimation, classical inference procedures and some heuristics. *Mutliscale Model. Simul.* **2007**, *6*, 656–687.

(32) Park, S.; Schulten, K. Calculating potentials of mean force from steered molecular dynamics simulations. *J. Chem. Phys.* **2004**, *120*, 5946–5961.

(33) Hummer, G. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.* **2005**, *7*, 34.

(34) Calderon, C. P.; Chelli, R. Approximating nonequilibrium processes using a collection of surrogate diffusion models. *J. Chem. Phys.* **2008**, *128*, 145103.

(35) Calderon, C. P.; Janosi, L.; Kosztin, I. Using stochastic models calibrated from nanosecond nonequilibrium simulations to approximate mesoscale information. *J. Chem. Phys.* **2009**, *130*, 144908.

(36) Bacallado, S.; Chodera, J. D.; Pande, V. Bayesian comparison of markov models of molecular dynamics with detailed balance constraint. *J. Chem. Phys.* **2009**, *131*, 045106.

(37) Dudko, O. K.; Hummer, G.; Szabo, A. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 15755–15760.

(38) Chen, H.; Rhoades, E.; Butler, J. S.; Loh, S. N.; Webb, W. W. Dynamics of equilibrium structural fluctuations of apomyoglobin measured by fluorescence correlation spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 10459–10464.

(39) Johnson, Valen E; Bayesian, A. χ^2 test for goodness-of-fit. *Ann. Stat.* **2004**, *32*, 2361.

(40) Aït-Sahalia, A. Nonparametric tests of the markov hypothesis in continuous-time models. *Ann. Stat.*, in press.

(41) Pokern, Y.; Stuart, A. M.; Wiberg, P. Parameter estimation for partially observed hypoelliptic diffusions. *J. R. Stat. Soc., Ser. B* **2009**, *71*, 49–73.

(42) Olsson, M. H.; Mavri, J.; Warshel, A. Transition state theory can be used in studies of enzyme catalysis: lessons from simulations of tunnelling and dynamical effects in lipoxygenase and other systems. *Philos. Trans. R. Soc. London, Ser. B* **2006**, *361*, 1417–1432.

(43) Bolhuis, P. G.; Dellago, C.; Geissler, P. L. Transition path sampling: throwing ropes over mountains in the dark. *J. Phys.: Condens. Matter* **2000**, *12*, A147–A152.

JP911124Z