Received: 15 November 2008, Revised: 1 April 2009, Accepted: 2 April 2009, Published online in Wiley InterScience: 28 May 2009

(www.interscience.wiley.com) DOI:10.1002/jmr.959

Analyzing single-molecule manipulation experiments

Christopher P. Calderon[a]*, Nolan C. Harris[b], Ching-Hwa Kiang[b] and Dennis D. Cox[c]

Single-molecule manipulation studies can provide quantitative information about the physical properties of complex biological molecules without ensemble artifacts obscuring the measurements. We demonstrate computational techniques which aim at more fully utilizing the wealth of information contained in noisy experimental time series. The "noise" comes from multiple sources e.g., inherent thermal motion, instrument measurement error, etc. The primary focus of this paper is a methodology that uses time domain based methods to extract the effective molecular friction from single-molecule pulling data. We studied molecules composed of eight tandem repeat titin I27 domains, but the modeling approaches have applicability to other single-molecule mechanical studies. The merits and challenges associated with applying such a computational approach to existing single-molecule manipulation data are also discussed. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: single-molecule manipulation; stochastic differential equation; effective friction; local maximum likelihood

INTRODUCTION

An understanding of how individual molecules behave mechanically and chemically can be of importance to both nanotechnology and materials' applications (Becker *et al*., 2003; Ackbarow *et al*., 2007; Muller and Dufrene, 2008). These types of studies are also important in understanding the fundamental biophysics of biomolecules like proteins and nucleic acids (Clausen-Schaumann *et al*., 1999; Hegner *et al*., 1999; Marszalek *et al*., 1999; Collin *et al*., 2005; Evans and Calderwood, 2007; Harris *et al*., 2007; Liu *et al*., 2007; Greenleaf *et al*., 2008). Single-molecule experiments have allowed researches to explore small scale systems with high spatial and temporal resolution and have provided useful mechanical information about individual molecules. Single-molecule experimental techniques are making rapid advances and time series resulting from these experiments contain a rich amount of physical information.

In this paper we demonstrate how time domain based computational modeling techniques utilizing local maximum likelihood methods (Calderon, 2007a; Calderon and Chelli, 2008; Calderon *et al*., 2009a,b) can be used to extract kinetic information from single-molecule time series where a molecule is stretched by an external force. Specifically, we use an AFM tip to mechanically unfold/refold a molecule composed of eight serially linked I27 domains from human cardiac titin using constant velocity pulling experiments. Although we focus on studying a protein with AFM, the computational methods we present are applicable to other systems e.g., nucleic acids (Calderon *et al*., 2009b) or optical tweezer manipulation experiments (Collin *et al*., 2005; Seol *et al*., 2007).

The primary interest in this paper is in extracting the effective molecular friction induced by a protein molecule as a function of molecular extension. From a biological standpoint, the I27 domain of titin acts as a molecular "shock absorber" (Li *et al*., 2000). The hydrogen bond network in the beta sandwich secondary structure is believed to help dissipate energy caused by rapid extension of biological materials such as cardiac muscle and to also allow extension of tissue without inducing rupture (Li *et al*., 2000). The amount of energy damped depends on the effective molecular friction of the molecule. The elastic and frictional properties of molecules can be tuned and this shows great promise in nanotechnology and materials' applications where one desires to design synthetic adhesives or spider-silk like materials (Li *et al*., 2000; Becker *et al*., 2003). Single-molecule techniques provide one means to characterize these types of novel materials, but directly estimating the effective molecular friction from single-molecule data can be challenging due to the multiscale noise sources influencing these types of measurements (Kawakami *et al*., 2006; Khatri *et al*., 2008; Calderon *et al*., 2009a).

Previous works (Higgins *et al*., 2006; Kawakami *et al*., 2006; Khatri *et al*., 2008) have demonstrated that the contribution of the effective friction due to the molecule attached to the AFM cantilever can be decoupled from that associated with the solvent interacting with the cantilever in experiments where the cantilever is oscillated at a constant frequency. Frequency domain techniques which utilize a simple harmonic oscillator

* *Correspondence to: C. P. Calderon, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA. E-mail: calderon@rice.edu*

[a] *C. P. Calderon*
Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005, USA

[b] *N. C. Harris, C.-H. Kiang*
Department of Physics and Astronomy, Rice University, Houston, TX 77005, USA

[c] *D. D. Cox*
Department of Statistics, Rice University, Houston, TX 77005, USA

model were used to quantify the effective friction due to the molecule. Here we demonstrate how time domain based methods (Calderon *et al*., 2009a,b) can be used to estimate the effective molecular friction. We borrow many of the ideas laid out in (Kawakami *et al*., 2005; Kawakami *et al*., 2006), but our techniques do differ in several important aspects which we highlight throughout. The primary motivations for providing an alternative computational method for extracting effective molecular friction from experimental times series are as follows: (1) data coming from constant loading rate experiments (Evans and Calderwood, 2007) do not utilize an oscillating cantilever. The time domain based methods we demonstrate can utilize existing time series to estimate an effective friction; (2) in other circumstances, the use of an oscillating external force may complicate the analysis of rupture time distributions in various single-molecule experiments (Evans and Calderwood, 2007; Dudko *et al*., 2008). Possessing computational tools which can estimate rupture time distributions, reaction rate constants, and state-dependent effective friction from the same experimental data are desirable because one can then make better use of the information content contained in experimental data (Evans and Calderwood, 2007; Dudko *et al*., 2008). In addition to estimating the effective friction from experimental data, we also demonstrate how the models we use can quantify other sources of variation in single-molecule experiments. For example, we can quantitatively measure how the tip attachment location influences the measured effective friction.

The remainder of this paper is organized as follows: the "Materials and Methods" section reviews the models, discusses computational details, and presents the experimental setup. This is followed by the "Results and Discussion" and "Conclusions" sections.

MATERIALS AND METHODS

Local maximum likelihood estimation (state space/time domain)

The time domain method we utilize was first presented in (Calderon *et al*., 2009a,b). We summarize the main features here and in the "Results and Discussion" section we present a new discussion expanding on our motivation for using this type of modeling framework. Nonlinear stochastic differential equations (SDEs) are fit to data coming from AFM experiments where an external force is added into the system (Balsera *et al*., 1997; Li and Makarov, 2003). The global SDE (Calderon, 2007a; Calderon and Chelli, 2008) representing the dynamics of a single time series is assumed to be a nonlinear diffusion of the form

$$dz_t = \mu(t, z_t)dt + \sqrt{2}\sigma(z_t)dB_t \tag{1}$$

$$y_{t_i} = z_{t_i} + \varepsilon_{t_i} \tag{2}$$

where z_t represents the experimental observable value at time t (throughout we use the end-to-end extension of the titin molecules as this observable), B_t represents the standard Brownian motion, $\mu(\cdot,\cdot)$ the time dependent drift function, and $\sigma^2(\cdot)$ represents the diffusion coefficient.[1] The drift term is time dependent because we are adding an external force. The "instrument noise" (ε_{t_i}) does not allow us to directly observe, z_t, instead we observe $y_{t_i} = z_{t_i} + \varepsilon_{t_i}$ where the subscript on the time index is used to stress that our observations are discrete.

In single-molecule systems, the complexity of the atomistic system often cannot be ignored and this causes problems in developing physically based, accurate parametric SDE models (Hummer and Kevrekidis, 2003; Park and Schulten, 2004; Hummer, 2005; Calderon, 2007b) from *a priori* considerations. For example, the "force-hump" associated with a titin unfolding intermediate (Lu *et al*., 1998; Higgins *et al*., 2006) cannot be predicted by simple coarse-grained polymer models. We assume that the global dynamics are completely unknown *a priori*, so appealing to a standard parametric estimation scheme is problematic. To overcome the difficulty of unknown global drift and diffusion functions. We use local models (Calderon, 2007a,b; Fan *et al*., 2007; Calderon and Chelli, 2008) to fit the coefficients of polynomial SDEs whose functional form is motivated by the overdamped Langevin equation (Calderon, 2007a). We should stress that the basic local modeling idea presented here does not require an overdamped Langevin structure. If there is a compelling reason to try other local models' structures, they can be entertained. The relevant expressions for the local models used here are

$$\sigma^{\text{loc}}(z) = (C + D(z - z^o)) \tag{3}$$

$$F^{\text{Ext}}(t, z) = k_{\text{pull}}(\lambda(t) - z) \tag{4}$$

$$F^{\text{Int}}(z) = (A + B(z - z^o)) \tag{5}$$

$$\mu^{\text{loc}}(t, z) = \frac{(\sigma^{\text{loc}}(z))^2}{k_BT}\left(F^{\text{Int}}(z) - F^{\text{Ext}}(t, z)\right) \tag{6}$$

where k_BT represents Boltzmann's constant times the system temperature, F^{Ext} the external force applied to the system, $\lambda(t)$ is the pulling protocol (common to all experiments/simulations), k_{pull} is the spring constant associated with the harmonic constraint used to apply the external force, F^{Int} the force due to internal molecular interactions, and $\theta \equiv (A, B, C, D)$ is the local parameter vector estimated by approximate maximum likelihood estimation (Jimenez and Ozaki, 2006). z^o is a free parameter used only for estimation purposes. We set z^o to the average (temporal) value observed in a given local time series window. The windows were formed by dividing a single global time series into M "time windows."[2] The modeling ideas behind the two-scale realized volatility estimator (TSRV) (Zhang *et al*., 2005) were used to approximate the variance of the instrument noise process in each local window.

The information from a TSRV inspired analysis is used in conjunction with likelihood based methods to estimate quantities describing the z dynamics.[3] The fitting criterion used was motivated by the local linear maximum likelihood type method outlined by Jimenez and Ozaki (2006). The stage location

[1] The thermal noise in the "internal system" has contributions coming from internal molecular fluctuations as well as solvent bombardment on the molecule and the cantilever tip. Methods for decoupling these noise sources are discussed later in this article.

[2] The width of the time series window was determined by the length of the force trace time series, but each local model consisted of roughly 300 observations and each unfolding event was associated with 20–40 sets of local SDE model parameters.

[3] An analysis of the autocorrelation of the temporal differences (e.g., $\{y_{t_{i+1}} - y_{t_i}\}_i^N$) of the time series suggested that the assumption of white measurement noise was reasonable and the goodness-of-fit tests employed also suggested the white noise proxy was statistically acceptable for our observed time series (results available upon request).

$\lambda(t)$ was denoised by using the Daubechies (five vanishing moments) wavelet family to smooth the signal.[4] All measurement noise was assumed to be contained in y after this smoothing procedure. The validity of the various assumptions (diffusive noise, local linearity, etc.) were tested in an *a posteriori* fashion using the probability integral transform based Q-test developed by Hong and Li (2005).

To obtain a global SDE model, a penalized spline was used to stitch the piecewise polynomial models together (Ruppert *et al.*, 2003). The full numerical details of this spline procedure have been outlined by Calderon *et al.* (2009c). Briefly, a sequence of estimated local θs measured along one experimental time series are used to construct the global functions (via smoothing splines (Calderon *et al.*, 2009c)) needed to provide μ and σ. Information about the parameter uncertainty is used to determine a regularized global model from the local θs. The procedure is then repeated for each observed experimental time series. An additional discussion on the local windows is presented in the "Results and Discussion" section.

Effective friction computations

The overdamped Langevin equation in Equation 3–6 implicitly assumes that the diffusion coefficient is related to the effective friction via the Einstein relation: $\sigma^2 = k_BT/\zeta$ where the effective friction is denoted by ζ. Recall that we do test the validity of the local model assumptions using goodness-of-fit tests (Hong and Li, 2005). After the global $\hat{\sigma}$ function was fit using our smoothing spline procedure, the resulting spline was subsequently used to compute the effective friction $(\hat{\zeta})$ from the Einstein relation; hats above a variable are used to remind us that the function comes from time series estimates. Recall that our time domain based model accounted for the instrument measurement noise, but the effective friction that we compute has contributions coming from both the effective friction induced by the molecule (the quantity of interest) and the friction experienced by solvent bombarding the cantilever ("classical" thermal noise). In order to decouple these sources, we use the approach of Kawakami *et al.* (2006), namely we assume that the friction acting on the system can be thought of as two parallel springs. Under this assumption, we can subtract the effective friction experienced by the cantilever/"protein containing some folded I27 domains" system, denoted by $\hat{\zeta}$, from that associated with the cantilever/"protein with all I27 domains unfolded" system $(\hat{\zeta}^{\text{Unfold}})$ and use this difference to estimate the effective friction due to the protein containing folded I27 domains i.e., $\hat{\zeta}^{\text{Folded}}(z) = \hat{\zeta}(z) - \hat{\zeta}^{\text{Unfold}}$. We used a common estimate of the scalar $\hat{\zeta}^{\text{Unfold}}$ for all computations.

AFM experiments

10 μl of protein solution, 50–100 μg/ml containing eight serially linked repeats of the I27 domain of human cardiac titin (Athena ES) was incubated on a gold substrate at room temperature for 20 min. A Multimode AFM with picoforce option (Veeco Instruments) was used for experimental measurements. Individual protein chains were attached to a silicon nitride cantilever tip with spring constant, $k_{\text{pull}} = 50$ pN/nm. The spring constant was determined using methods appealing to the equipartition theorem (Hutter and Bechhoefer, 1993; Butt and Jaschke, 1995).

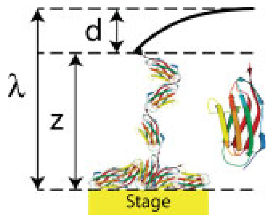

Figure 1. Schematic of experimental apparatus stretching the I27 domain of titin.

The experimental setup is illustrated in Figure 1. The attached molecule was stretched to unfold several domains, allowed to relax back close to the substrate surface, and held at a constant position to allow the molecule to refold before repeating the cycle. The stretch and relax portions of the cycle were performed at a constant velocity of 50 nm/s, followed by a rest time of 30 s. Discrete time series were recorded at the frequency of 20 kHz.

Sample raw AFM data is shown in Figure 2. In our experiments we attempted to retain the same molecule for multiple cycles. This was done in an effort to minimize variation due to the tip attachment point and other random factors not typically within our control in AFM pulling experiments. In the results presented, we analyze three batches of data. Each batch contains results from holding the same molecule for seven fold/refold cycles (the different traces shown in Figure 2 are associated with a batch of cycles obtained with one molecule). In our analysis we only use the second and third peak for modeling because the first

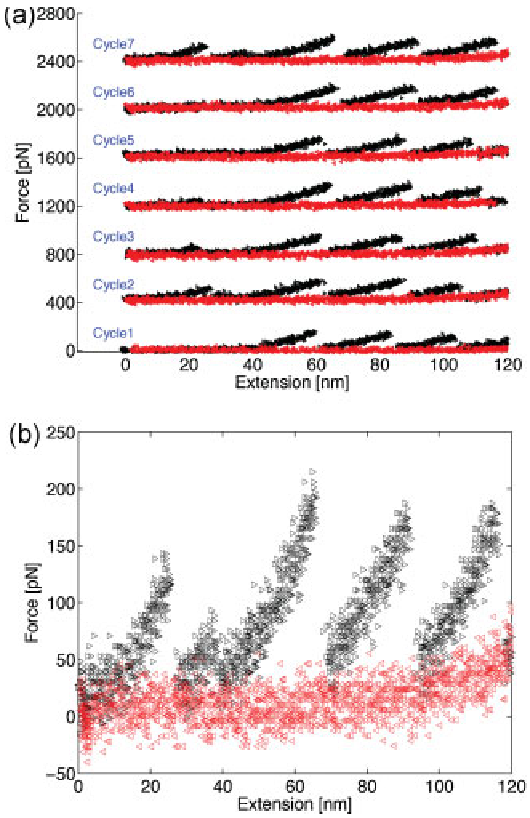

Figure 2. (a) Sample force extension trace on one titin molecule. A titin molecule was attached to an AFM tip and retained for seven cycles; each cycle consisting of an extension phase (data points denoted by ▷) and a relaxation phase (◁). We vertically shifted the force extension data by 400 pN on each cycle for display purposes only. (b) We zoom in on cycle 7.

[4] We used a "level six" wavelet approximation for smoothing the stage, wavelet coefficient thresholding could also be entertained.

peak can contain artifacts from nonspecific binding (Kawakami *et al*., 2006). We did not attempt to analyze all eight peaks because of the experimental difficulty associated with retaining a single titin molecule for multiple force extension cycles. That is, if we successfully held onto a molecule and observed three unfolding events, we reversed the AFM stage direction in order to minimize the chances of the molecule detaching from the AFM tip as opposed to attempting to unfold all eight domains.

RESULTS AND DISCUSSION

Effective force and friction computations

Figure 3a plots the estimated effective internal force in and the effective friction coming from the titin molecule is shown in Figure 3a–b. In these plots, three different titin molecules were subjected to seven unfolding/refolding cycles. We refer to the seven experiments where one molecule is retained as a "batch;" in each batch we only report the unfolding portion of the cycle. The term "Relative Extension" is used on the *x*-axis because we shifted each extension to start the coordinate system after the first force peak (see MATERIALS AND METHODS) resulting in a common initial extension for each cycle observed.

Even within each batch there is significant variation and this variation likely has physical relevance in addition to the unavoidable sampling noise associated with a finite discretely sampled data set. Note that the force hump (Lu *et al*., 1998; Marszalek *et al*., 1999) that is believed to correspond to an intermediate is not always detected. It is known that the likelihood of observing the signature of this intermediate decreases as the number of domains unfold (Higgins *et al*., 2006) and the strength of this signal depends on unobserved conformation details. We did not attempt to align the force traces from different experiments (we only shifted the extension as discussed above), but as one can see there are clearly different slopes (and curve shapes) in the different force curves. The different slopes can physically be interpreted as different "effective stiffness coefficients." Also the presence of a force hump in only some of the force curves demonstrates that there are indeed significant differences in the force responses that alignment would not help resolve. The variation within each batch is large enough in the force curves so as to make the variation induced by the random attachment point negligible. This is not surprising given that several groups have been able to reproduce similar results with this model system (Li *et al*., 2000). The variation induced by the randomness of the attachment point does seem to be slightly detectable when inspecting the effective internal friction, but further experiments are needed to determine if these differences in the batches are due to the attachment point and the underlying conformational details or to instrument artifacts.

Recall that our primary interest is in the effective friction due to the molecule. The time domain estimate of the friction is very noisy as can readily be seen by looking at the global smoothing splines plotted in Figure 3 (in the "Time domain computational issues" section we discuss possible techniques for reducing the variance), but before getting into this technical discussion we prefer to focus on the physical interpretations of this result and compare our result to frequency domain results obtained elsewhere (Kawakami *et al*., 2006).

Effective friction estimates compared to previous work

To facilitate our discussion we plot results "averaging" over the different unfolding events within each batch in Figure 4 (see caption for details). The percentage of friction estimated due to the folded polymer construct where the aforementioned percentage is computed according to $\hat{\zeta}^{\text{Folded}}(z)/\hat{\zeta}^{\text{Unfold}} \times 100 + \phi$ where ϕ is a "shift constant" used to shift the percentage at zero relative extension to zero. This shift constant was used because we used one single $\hat{\zeta}^{\text{Unfold}}$ for all data points and it is likely that this value is not appropriate for all experiments due to unavoidable experimental drift; the shift constant is only used to register the curves and facilitate making comparisons (the unshifted data using a common baseline is contained in Figure 3).

First we discuss similarities with other experiments and theoretical predictions. In (Khatri *et al*., 2008) a polymer model, quantitatively explaining why a contribution to the effective friction should increase as a function of extension was presented and the authors supported their theoretical predictions with force clamp experiments. A wide degree of variation was observed between different experiment batches when an effective friction was measured (Khatri *et al*., 2008); note that

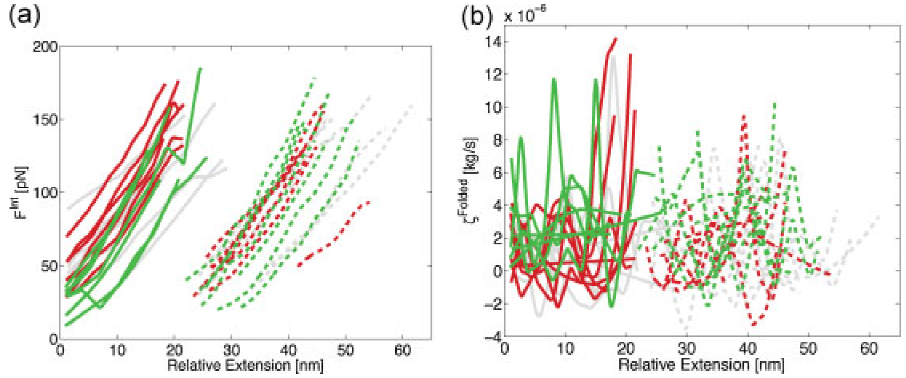

Figure 3. (a) The effective internal force and (b) effective internal friction coefficient were estimated using global SDE models calibrated from AFM time series. In every case reported, a single titin molecule remained attached to the AFM tip for seven force extension cycles. The different colors indicate replicated experiments where a new titin molecule was attached to the tip. The solid lines denote the second peak observed (all extensions are shifted relative to this peak) and the dotted lines correspond to the third force peak observed. The first peak was not analyzed due to nonspecific binding which can affect the results.

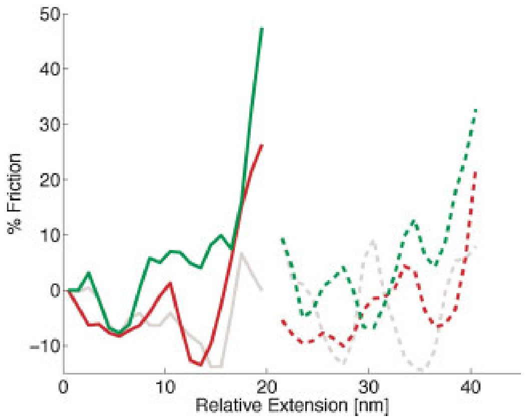

Figure 4. The average percentage of the internal friction. Since the length of the unfolded molecule was different in each case due to various factors (different tip attachment point, random breakage times, etc.), this average was computed by first estimating the global friction $\hat{\zeta}$ at 20 discrete points evenly spaced between the maximum and minimum extensions reported in each curve shown in Figure 3. The average over this grid was computed and the result is reported on a common rescaled grid. Since the common baseline $\hat{\zeta}^{\text{Unfold}}$ may not be good for all experiments, we shifted the first unfolding event to zero for the three cases studied. The different curves correspond to subjecting a titin molecule to seven different cycles (retaining the same attachment point throughout) and averaging the results in the fashion described above. The colour and line types used are identical to those of the corresponding experiments reported in Figure 3.

the tip attachment point changes in each batch. We also observe that the effective friction appears to generally increase as a function of extension and that the tip attachment point appears to introduce significant variation. In another study probing the effective friction of a titin molecule containing five serially linked I27 domains (Kawakami *et al*., 2006), the authors found that effective friction tended to decrease as the number of domains unfolded by force increased. Our results in Figures 3 and 4 also support this finding. A series of I27 domains acts as an effective "shock absorber" and the amount of energy this network can dissipate depends in a nontrivial fashion on the number of folded domains. Recall that the individual I27 domains each consists of a beta sandwich structure possessing a large hydrogen bond network which is believed to absorb/dissipate a significant amount of energy in various situations (Lu *et al*., 1998; Becker *et al*., 2003; Ackbarow *et al*., 2007).

Now we discuss some of the differences between our experiments/analysis and previous studies. We studied titin molecules containing eight repeat I27 domains instead of five as was done in the other viscoelastic studies we are comparing to. Also we do not oscillate the cantilever tip in our experiments and appeal to time domain/local maximum likelihood techniques using nonlinear SDEs accounting for state-dependent effective noise whereas the other studies use frequency domain methods that appeal to a linear harmonic oscillator model assuming constant effective noise. With this said, we proceed to present some of the more notable differences in results.

Our computations of the effective friction are noisier than the methods we compare to. This is one of the motivations for presenting the average curve in Figure 4. After we aligned the curves and computed this average we notice that in each case the effective friction slightly "dips" around the 75% relative extension before it rapidly increases. The large amount of noise and the various approximations employed here are suspect, but the reproducibility of this feature and the known existence of an intermediate suggest that it is possible that the formation of the intermediate temporarily reduces the effective friction. Note that we did subject our models to hypothesis tests (Hong and Li, 2005) appropriate for nonstationary time series and the models performed adequately. For example, out of 1400 local models (data coming from one batch of experiments) we rejected 6.07% of the local models on par with the rejection rate that one would expect under the "null" (assuming the estimated maximum likelihood parameter gives the "true" model). An increased amount of experimental data samples and tests possessing better power properties may help identify/diagnosing potentially poor assumptions in our models using goodness-of-fit tests. The aforementioned additional experimental data can be obtained by various means e.g., using a higher discrete time sampling rate or using slower pulling speeds. The availability of "omnibus" time series tests (Hong and Li, 2005) is one appealing feature of the time domain methods we present.

In experiments where the cantilever was oscillated (Kawakami *et al*., 2006), the "dip" was not observed but these authors estimated effective friction curves that appear to contain less uncertainty. It would be interesting to see if the modeling techniques we use here detect a slight dip using the same data sets and/or if the signal to noise ratio in our estimated models is improved in oscillating cantilever experiments. Recall, that in some situations it might not be desirable to introduce an oscillating cantilever if the primary interest is not in the effective molecular friction, but instead in something like the net effective dissociation or unfolding rates (Dudko *et al*., 2008). Possessing a technique which can utilize existing data sets to infer multiple physical quantities can help one in better characterizing natural or man-made systems, and our method offers this possibility.

Also, the absolute value of the effective friction we obtained for the second force peak is estimated to be roughly an order of magnitude larger than two studies carried about by other researchers referenced above. This could be in part due to the larger number of repeat I27 domains. Also, both the frequency domain and time domain methods we use are subject to systematic bias. For example, we estimate $\hat{\sigma}(z)$ and use this to compute the state-dependent friction. Even if a diffusive process truly generates the data, the measurement noise, finite number of samples, and machine drift will make a nontrivial relationship between our estimate and the "truth" i.e., $\hat{\sigma}(z) = \sigma(z) + b(z) + \varepsilon$ where $b(z)$ is a bias function and ε is a zero mean process meant to account for unavoidable variation due to finite sample sizes. If $b(z) > 0$ everywhere we would tend to be underestimating the effective friction. This could happen if the instrument noise associated with the photon detector measuring the cantilever deflection "leaked" into the $\hat{\sigma}(z)$, which is only supposed to contain thermal noise sources due to solvent bombardment and molecular friction. If $b(z) < 0$ everywhere we would tend to overestimate the effective friction. This could happen in our models if we over-smoothed the stage location, $\lambda(t)$, too much making the system dynamical noise seem smaller than it should be due to the underlying thermal noise. In practice $b(z)$ will likely be a complicated function taking both signs and will be hard to determine from a finite set of samples when laboratory data is analyzed, but higher precision instrumentation, increased sampling rates, and goodness-of-fit tests can help us in approximating this bias in the future. These types of issues are discussed in the next section.

Time domain computational issues

Figure 5 provides a schematic illustrating the basic idea behind our method. Each time series is divided into "*M*" total blocks (the variable *M* is used to index the blocks). In the titin cases presented, we selected *M* so that each block contained about 400 observations uniformly spaced in time. Within each block, there was an average state value associated with the time series, denoted by z^0. The local internal force and $\sigma(z)$ function were estimated using maximum likelihood type methods in this window (Jimenez and Ozaki, 2006). Effectively we obtained an estimate of the function (e.g., $\sigma(z^0) \simeq C$) and its derivative (e.g., $\partial\sigma(z)/\partial z\mid_{z^0} \simeq D$) at z^0. We then used SDE simulations to approximate the covariance associated with the estimated local parameters (assuming a normal error distribution of the estimated parameters). This was carried out for all *M* blocks, and then a smoothing spline procedure (Calderon *et al*., 2009c) using all of this information to construct $\hat{\sigma}(z)$ which was our estimate of the unknown "truth" $\sigma(z)$. We observed that this estimate was noisy; this noise is indirectly reflected in Figure 3b. A more direct representation of the noise observed in $\hat{\sigma}(z)$ is presented by Calderon *et al*. (2009a).

Increasing the rate at which we sample from the AFM device would likely improve this situation. Reducing the pulling speed would also allow us to obtain more data points in a given extension range which again may help our estimates. In addition, we could entertain the possibility of using overlapping blocks i.e., blocks $m = 1$ and 2 in Figure 5 could be constructed to share entries. The local estimation procedure would not change, but accurately accounting for the covariance in the uncertainty estimates used in the smoothing spline procedure would become more computationally intensive. This is one possibility for reducing the variance in our estimates. We could also entertain other methods for parameterizing our local SDEs (i.e., model the effective friction function directly as opposed to the diffusion term). If the problem under consideration requires high accuracy the extra computational load or alternative procedures may be worth pursuing.

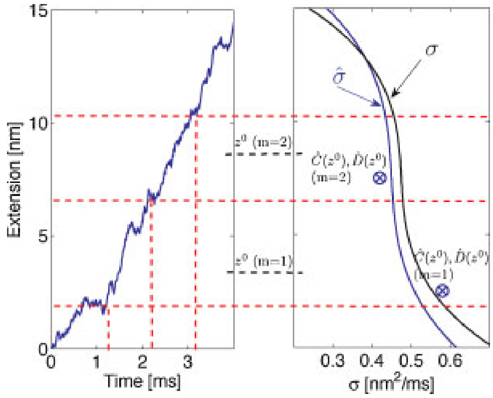

Figure 5. Illustration of the procedure used to estimate global SDEs from local models. The windows "m = 1" and "m = 2" each contained an equal number of *N* discrete time observations. In both windows the function and its derivative (with respect to the "extension" *z*) are estimated from the *N* time series samples at the point z^0 (in each window this corresponded to the average extension observed). We assume that for each time series there is a "truth" (the function $\sigma(z)$) corresponding to the SDE of interest. The notation $\hat{\sigma}(z)$ is used to remind us that we have used a sequence of these windows to approximate the "truth."

Note that we also introduced smoothing into our approximation procedure at several points. We smoothed the stage location $\lambda(t)$ and assumed the resulting smoothed estimate contained no measurement uncertainty. Different types of smoothing and/or including instrument noise explicitly in this quantity can moderately influence the computations we presented. For example, the absolute value of the effective friction changed by about 20% using a different smoothing method for $\lambda(t)$ (though the percentage of friction due to the folded protein did not appreciably change because this smoothing influences both $\hat{\zeta}$ and $\hat{\zeta}^{\mathrm{Unfold}}$). We reported the "best" results using the Q-test statistic (Hong and Li, 2005) to guide our choice. This test is appealing to our situation because it does not require one to assume a stationary time series and it is easy to compute using the maximum likelihood approximation employed in estimating (*A*, *B*, *C*, *D*) in each local window. We also used spline smoothing in approximating the global estimates of $\sigma(z)$ and $F^{\mathrm{Int}}(z)$. Recall that the method presented utilizes "point function estimates" (e.g., *A* or *C*) and "derivatives or linear sensitivities" (*B* or *D*) in constructing the global functions (Calderon *et al*., 2009c). One could also entertain applying standard smoothing procedures to the linear sensitivities directly to assist in understanding physical features. For example, if the "polymer stiffness" (Kawakami *et al*., 2006) is a complicated function of extension (Calderon *et al*., 2009c), then one might want to obtain a smoothing spline approximating $\partial F^{\mathrm{Int}}(z)/\partial z$ directly. In addition, one may want to consider local models that allow more flexibility than a simple harmonic oscillator or an overdamped Langevin equation. The local models can readily be changed in our setup, and the basic modeling ideas can be applied to other local model structures.

CONCLUSIONS

We demonstrated how an SDE modeling method (Calderon *et al*., 2009a,b) accounting for instrument and inherent thermal noises can be applied to estimate the effective friction due to a single molecule in an AFM experiment. The findings were qualitatively in agreement with previous results using different experimental setups and analysis. One appealing feature of the method we presented is that it can be used to estimate effective friction without oscillating the cantilever. This has relevance to constant loading rate experiments (Evans and Calderwood, 2007; Dudko *et al*., 2008) which are becoming more popular in single-molecule studies; data generated in constant loading rate experiments can be re-analyzed using our methods and provide additional physical information from existing data. Another attractive feature of our tests is that we can subject our modeling assumptions to goodness-of-fit tests. We also discussed how estimating the derivative (linear sensitivity) of the effective diffusion and/or force could help in identifying state-dependence of these quantities. State-dependent noise is common in complex systems where we only monitor a crude system observable like an end-to-end extension of the molecule (Calderon *et al*., 2009a,b).

These types of computational tools can also help in identifying "outliers" in single-molecule time series. Atypical measurements of the effective friction as a function of extension can sometimes be attributed to unobserved collective conformational coordinates modulating the effective dynamics (Calderon *et al*., 2009b). Work of this type may help in identifying signatures associated with refolding events (Muller and Dufrene, 2008). Techniques for quantitatively characterizing the response of single-molecule systems can also assist our understanding of natural and man-made biomaterials (Li *et al*., 2000).

Acknowledgements

Christopher P. Calderon thanks NSF DMS 0240058, NSF ACI-0325081, and the computer support provided by NSF CNS-0421109 and a partnership between Rice AMD and Cray. The work of Nolan C. Harris and Christopher P. Calderon was supported in part by a training fellowship from the Nanobiology Training Program of the W. M. Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (NIH T90DK70121-04 and 5 R90 DK71504-04). Ching-Hwa Kiang thanks NSF DMR-0505814 and Welch Foundation C-1632 for support. Dennis D. Cox thanks NSF DMS-0505584.

REFERENCES

Ackbarow T, Chen X, Keten S, Buehler M. 2007. Hierarchies, multiple energy barriers, and robustness govern the fracture mechanics of alpha-helical and beta-sheet protein domains. *Proc. Natl. Acad. Sci. USA* **104**: 16410–16415.

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

Becker N, Oroudjev E, Mutz S, Cleveland J, Hansma P, Hayashi C, Makarov D, Hansma H. 2003. Molecular nanosprings in spider capture-silk threads. *Nat. Mater.* **2**: 278–283.

Butt H-J, Jaschke M. 1995. Calculation of thermal noise in atomic force microscopy. *Nanotechnology* **6**: 1–7.

Calderon C. 2007a. Fitting effective diffusion models to data associated with a “glassy potential”: estimation, classical inference procedures and some heuristics. *Mutliscale Model. Simul.* **6**: 656–687.

Calderon C. 2007b. On the use of local diffusion for path ensemble averaging in potential of mean force computations. *J. Chem. Phys.* **126**: 084106.

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

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

Calderon C, Martinez J, Carroll R, Sorensen D. 2009c. P-splines using derivative information. *Ann. Appl. Stat. pages submitted, preprint: http://www.caam.rice.edu/tech_reports/2009/TR09–13.pdf*

Calderon C, Chelli R. 2008. Approximating nonequilibrium processes using a collection of surrogate diffusion models. *J. Chem. Phys.* **128**: 145103.

Clausen-Schaumann H, Rief M, Gaub HE. 1999. Sequence-dependent mechanics of single DNA molecules. *Nat. Struct. Biol.* **6**: 346–349.

Collin D, Ritort F, Jarzynski C, Smith S, Tinoco I Jr, Bustamante C. 2005. Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature* **437**: 231–234.

Dudko O, Hummer G, Szabo A. 2008. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proc. Natl. Acad. Sci. USA* **105**: 15755–15760.

Evans E, Calderwood D. 2007. Forces and bond dynamics in cell adhesion. *Science* **316**: 1148–1153.

Fan J, Fan Y, Jiang J. 2007. Dynamic integration of time- and state-domain methods for volatility estimation. *J. Am. Stat. Assoc.* **102**: 618–631.

Greenleaf W, Frieda K, Foster D, Woodside M, Block S. 2008. *Direct observation of hierarchical folding in single riboswitch aptamers. Science* **319***:* 630–633.

Harris N, Song Y, Kiang C. 2007. Experimental free energy surface reconstruction from single-molecule force spectroscopy using Jarzynski’s equality. *Phys. Rev. Lett.* **99**: 068101.

Hegner M, Smith S, Bustamante C. 1999. Polymerization and mechanical properties of single reca-dna filaments. *Proc. Natl. Acad. Sci. USA* **96**: 10109–10114.

Higgins M, Sader J, Jarvis S. Frequency modulation atomic force microscopy reveals individual intermediates associated with each unfolded I27 titin domain. *Biophys. J.* 2006. **90**: 640–647.

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

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

Hummer G, Kevrekidis I. 2003. Coarse molecular dynamics of a peptide fragment: free energy, kinetics, and long-time dynamics computations. *J. Chem. Phys.* **118**: 10762–10773.

Hutter JL, Bechhoefer J. 1993. Calibration of atomic-force microscope tips. *Rev. Sci. Instrum.* **64**: 1868–1873.

Jimenez J, Ozaki T. 2006. An approximate innovation method for the estimation of diffusion processes from discrete data. *J. Time Ser. Anal.* **27**: 77–97.

Kawakami M, Byrne K, Khatri B, Mcleish T, Radford S, Smith D. 2005. Viscoelastic measurements of single molecules on a millisecond time scale by magnetically driven oscillation of an atomic force microscope cantilever. *Langmuir* **21**: 4765–4772.

Kawakami M, Byrne K, Brockwell D, Radford S, Smith D. 2006. Viscoelastic study of the mechanical unfolding of a protein by AFM. *Biophys. J.* **91**: L16–L18.

Khatri BS, Byrne K, Kawakami M, Brockwell D, Smith D, Radford S, McLeish T. 2008. Internal friction of single polypeptide chains at high stretch. *Faraday Discuss.* **139**: 35.

Li H, Oberhauser A, Fowler S, Clarke J, Fernandez J. 2000. Atomic force microscopy reveals the mechanical design of a modular protein. *Proc. Natl. Acad. Sci. USA* **97**: 6527–6531.

Li P, Makarov D. 2003. Theoretical studies of the mechanical unfolding of the muscle protein titin: bridging the time-scale gap between simulation and experiment. *J. Chem. Phys.* **119**: 9260–9267.

Liu S, Bokinsky G, Walter N, Zhuang X. 2007. Dissecting the multi-step reaction pathway of an RNA enzyme by single-molecule kinetic fingerprinting. *Proc. Natl. Acad. Sci. USA* **104**: 12634–12639.

Lu H, Isralewitz B, Krammer A, Vogel V, Schulten K. 1998. Unfolding of titin immunoglobulin domains by steered molecular dynamic simulations. *Biophys. J.* **75**: 662–671.

Marszalek P, Lu H, Li H, Carrion-Vazquez M, Oberhauser A, Schulten K, Fernandez J. 1999. Mechanical unfolding intermediates in titin modules. *Nature* **402**: 100–103.

Muller D, Dufrene Y. 2008. Atomic force microscopy as a multifunctional molecular toolbox in nanobiotechnology. *Nat. Nanotechnol.* **3**: 261–269.

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

Ruppert D, Wand M, Carroll R. 2003. Semiparametric Regression. Cambridge University Press: New York.

Seol Y, Li J, Nelson P, Perkins T, Betterton M. 2007. Elasticity of short DNA molecules: theory and experiment for contour lengths of 0.6–7 microm. *Biophys. J.* **93**: 4360–4373.

Zhang L, Mykland P, Ait-Sahalia Y. 2005. A tale of two time scales: determining integrated volatility with noisy high-frequency data. *J. Am. Stat. Assoc.* **100**: 1394–1411.