Maximum likelihood estimation of protein kinetic parameters under weak assumptions from unfolding force spectroscopy experiments

Daniel Aioanei,* Bruno Samorì, and Marco Brucale†
Department of Biochemistry "G.Moruzzi," University of Bologna, Via Irnerio 48, 40126 Bologna, Italy
(Received 14 August 2009; published 23 December 2009)

Single molecule force spectroscopy (SMFS) is extensively used to characterize the mechanical unfolding behavior of individual protein domains under applied force by pulling chimeric polyproteins consisting of identical tandem repeats. Constant velocity unfolding SMFS data can be employed to reconstruct the protein unfolding energy landscape and kinetics. The methods applied so far require the specification of a single stretching force increase function, either theoretically derived or experimentally inferred, which must then be assumed to accurately describe the entirety of the experimental data. The very existence of a suitable optimal force model, even in the context of a single experimental data set, is still questioned. Herein, we propose a maximum likelihood (ML) framework for the estimation of protein kinetic parameters which can accommodate all the established theoretical force increase models. Our framework does not presuppose the existence of a single force characteristic function. Rather, it can be used with a heterogeneous set of functions, each describing the protein behavior in the stretching time range leading to one rupture event. We propose a simple way of constructing such a set of functions via piecewise linear approximation of the SMFS force vs time data and we prove the suitability of the approach both with synthetic data and experimentally. Additionally, when the spontaneous unfolding rate is the only unknown parameter, we find a correction factor that eliminates the bias of the ML estimator while also reducing its variance. Finally, we investigate which of several time-constrained experiment designs leads to better estimators.

DOI: 10.1103/PhysRevE.80.061916 PACS number(s): 87.15.−v, 82.37.Np, 02.50.−r

I. INTRODUCTION

The kinetics of protein unfolding under mechanical stress represents a very important topic in the field of biophysics as can be seen in the large number of reviews surveying the problem [1–13]. The most widely employed model for studying mechanical protein unfolding sees bond rupture as a decay of a metastable state with reaction kinetics given by

$$\frac{d\eta}{dt} = -k(f(t))\,\eta(t), \tag{1}$$

where $\eta(t)$ is the survival probability up to time t, $f(t)$ stands for the force at time t, and $k(f(t))$ is the dissociation rate [14]. The dependence of the dissociation rate on force was given in [15] the analytical formula

$$k(f) = k_0 e^{\alpha f}, \tag{2}$$

where $\alpha = x_\beta/(K_b T)$ with x_β standing for the position of the transition state along the mechanical reaction coordinate and k_0 being the spontaneous dissociation rate.

The two parameters k_0 and α are usually extracted by either of two approaches. The first one, sometimes called the "standard method" (SM) (see, e.g., [16,17]), involves Gaussian fits of the rupture force distributions for various loading rates and a linear regression between the most probable rupture forces and the logarithm of the loading rate, without taking into account the joint effects of multiple modules that unfold sequentially in the context of polyproteins. The second approach is based on Monte Carlo (MC) simulations (see, e.g., [1,18–23]), and in this case the mentioned joint effects are properly accounted for. However, both traditional methods have intrinsic shortcomings: they either throw away useful information by summarizing the data into statistics that are not *sufficient* or geometrically fitting as closely as possible quantities that are not of prime interest, such as linear dependencies or rupture force distributions, rather than focusing directly on finding the most probable kinetic parameters [24]. To overcome these shortcomings a maximum likelihood (ML) approach has been previously proposed [24,25].

We have further developed the ML approach in order to address the following problems:

(1) The probability to observe an unfolding event is a contextual feature of homomeric polyproteins.

(2) In real experiments a unique force-time, and likewise force-displacement, characteristic does not exist (see, e.g., [25]). When either the cantilever tip or the surface is not functionalized, which is often the case, this is in fact predicted by the theoretical models since they depend on microscopic parameters that vary based on the length of the subrange under mechanical stress and even from spot to spot depending on the local properties of the soft protein layer [26,27].

(3) Choosing one among the many existing theoretical force models for idealized polymeric chains, as reviewed, e.g., in [24,28–31], with various corrections of the interpolation formulas [32,33], is not trivial, and neither is deriving an empirical force model from the experimental data itself [25].

We tackle the first problem by taking into account the number of not-yet-unfolded modules when computing the survival probability in Eq. (3). We solve the second problem by allowing in Eq. (4) a different force-time function to de-

*aioaneid@gmail.com
†marco.brucale@unibo.it

scribe the stretching time range leading to each unfolding event as long as they are considered known (i.e., not introducing *nuisance parameters* into the likelihood function). Finally we address the third problem by constructing the force-time functions in an automated, fully objective way as increasing, continuous piecewise linear approximations to the AFM-recorded data points using Eq. (A4).

We show that α can be estimated just by maximizing the univariate function in Eq. (7), after which k_0 immediately comes out from Eq. (6). Since the statistical estimation procedure would not be complete without a way to compute the uncertainty of the estimated values [34], with Eq. (9) we show how to extract a *Bayesian credible region*, i.e., a fixed two-dimensional area that contains with a given probability the random point (k_0, α) (see, e.g., [35]). The approach is computationally feasible even for complex theoretical models such as wormlike chain (WLC) [36,37] that require numeric integration for the evaluation of the likelihood function. In fact WLC has never been used before in the context of ML estimation of kinetic parameters, but it can be easily applied with our framework by solving Eq. (A3).

It should be emphasized that it is common practice to fix α to a known value and estimate only k_0 in situations that are believed not to alter the position of the transition state: replacing water by deuterium oxide [38], certain protein mutations [22,39,40], and stretching proteins under the effect of chemical denaturants [41]. For this particular case we propose the *unbiased* and *more efficient* estimator given by Eq. (8).

II. THEORY

Next we are going to present the analytical form of the likelihood function; we will explain how it can be maximized and how to compute a credible region for the two parameters.

A. Likelihood function

When a monomeric protein is stretched starting with time t_s, from Eq. (1) and imposing that $\eta(t_s)=1$, we obtain

$$\eta(t) = \exp\left[- \int_{t_s}^{t} k(f(u))du \right], \quad t \geq t_s.$$

For a multimeric construct made up of identical tandem repeats behaving independently, let us consider an unfolding event after which, chronologically, there are $m-1 \geq 0$ more unfolding events in the single molecule force spectroscopy (SMFS) curve. The probability that all m modules survive becomes

$$\eta_m(t) = \exp\left[- m \int_{t_s}^{t} k(f(u))du \right], \quad t \geq t_s. \tag{3}$$

Assuming that $f(t)$ is continuous and increasing with $f(t_s)=y_s$, we can change the integration domain to force:

$$\eta_m(y) = \exp\left[- m \int_{y_s}^{y} k(z)(f^{-1})'(z)dz \right], \quad y \geq y_s.$$

The probability density $r_m(y)$ to observe a rupture event at force $y \geq y_s$ is

$$r_m(y) = -\frac{d}{dy}\eta(y) = mk(y)(f^{-1})'(y)\exp\left[- m \int_{y_s}^{y} k(z)(f^{-1})'(z)dz \right].$$

Notation 1. Let n be the total number of unfolding peaks in the whole data set, and for each unfolding event $1 \leq i \leq n$ we denote by t_{si} and y_{si} the time point and force at which we consider the stretching to start, by $t_i > t_{si}$ and $y_i > y_{si}$ the rupture time instant and force of rupture, and by m_i the number of modules that will unfold after i in the same curve, plus 1. The force-time function for peak i, from t_{si} to t_i, is described by f_i.

Since the unfolding events are independent of each other, the joint probability density function associated to the rupture forces $\vec{y}=(y_1 \ldots y_n)$ is

$$L(\vec{y};k) = \exp\left[- \sum_{i=1}^{n} m_i \int_{y_{si}}^{y_i} k(z)(f_i^{-1})'(z)dz \right] \prod_{i=1}^{n} m_i k(y_i) \times (f_i^{-1})'(y_i). \tag{4}$$

At this point we introduce α and k_0 explicitly into the joint probability density function by using Eq. (2):

$$L(\vec{y};k_0,\alpha) = \exp\left[- k_0 \sum_{i=1}^{n} m_i \int_{y_{si}}^{y_i} e^{\alpha z}(f_i^{-1})'(z)dz \right] \times k_0^n \prod_{i=1}^{n} m_i e^{\alpha y_i}(f_i^{-1})'(y_i). \tag{5}$$

Note that f_i can be any continuous increasing function. Appendix A contains more details about the computation of the derivative of the inverse force-time function when the WLC model is assumed [Eq. (A3)] and explicit formulas for the likelihood function when applied to the linear force-displacement characteristic $f(t)=\kappa v t$ [Eq. (A1)] or the piecewise linear force-time approximation [Eq. (A4)].

Briefly, the piecewise linear force-time approximation is a linear interpolation of a *longest increasing subsequence* (see, e.g., [42,43]) of the force values reported by the AFM during stretching, which is computed by removing the minimal number of data points such that the remaining ones show increasing force with time, and breaking ties by calling for increased time resolution toward the rupture event (see Appendix A). This approach eliminates most of the noise and it has the nice theoretical property that if applied to a set of forces that is already increasing, it becomes a simple linear interpolation.

B. Point estimation

Regarding $L(\vec{y};k_0,\alpha)$ as a function of k_0, the *conditional maximum likelihood estimate* of k_0 is the argument for which the function achieves the global maximum on $(0,\infty)$ and can be computed as

$$\hat{k}_0(\alpha) = \frac{n}{\sum_{i=1}^{n} m_i \int_{y_{si}}^{y_i} e^{\alpha z} (f_i^{-1})'(z) dz}. \quad (6)$$

Substituting in Eq. (5) we obtain the *profile likelihood* for α, which needs to be maximized numerically to obtain the estimator $\hat{\alpha}$:

$$L_p(\alpha) = \hat{k}_0(\alpha)^n \exp\left[\alpha \sum_{i=1}^{n} y_i\right] e^{-n} \prod_{i=1}^{n} m_i (f_i^{-1})'(y_i). \quad (7)$$

While ML estimators are known to have very good asymptotic properties when used with *i.i.d.* (independent and identically distributed) random variables (see, e.g., [44,45]), our rupture forces y_i are not identically distributed because each unfolding event i is assigned its own y_{si}, m_i, and f_i. As a result a more complex theory, such as perhaps that developed in [46], would be needed to study the asymptotic behavior of our ML estimators, but a rigorous treatment of the problem would exceed the scope of the present paper.

However, we do show in Appendix B that when α is fixed and known, and under the conditions of proposition 1, which can be shown to hold for the WLC interpolation formula in Eq. (A2) and for any force-time function that increases linearly after an arbitrary time point, the estimator $\hat{k}_0(\alpha)$ is biased. For this situation we propose the following unbiased estimator of k_0:

$$\tilde{k}_0(\alpha) = (n-1)\hat{k}_0(\alpha)/n, \quad n \geq 2. \quad (8)$$

Since $\sigma^2[\tilde{k}_0(\alpha)] = [(n-1)/n]^2 \sigma^2[\hat{k}_0(\alpha)]$, where σ^2 stands for variance, the unbiased estimator is also *more efficient*.

C. Bayesian credible region

We show here that the particular shape of our likelihood function makes it feasible to numerically compute a (rectangular) credible region for (k_0, α) containing the respective point estimates.

A key operation in the numerical computation of credible regions is the ability to efficiently integrate the likelihood function over (potentially infinite) rectangular regions. For this purpose we make the following observation:

$$\int_a^b x^n e^{-cx} dx = \frac{1}{c^{n+1}} \int_{ca}^{cb} y^n e^{-y} dy$$

$$= \frac{n!}{c^{n+1}} [P(n+1, bc) - P(n+1, ac)]$$

for any $0 \leq a < b \leq \infty$, $c > 0$, and integer $n \geq 0$, where P is the *incomplete gamma function* defined as

$$P(h,x) = \frac{1}{\Gamma(h)} \int_0^x t^{h-1} e^{-t} dt, \quad h > 0.$$

The integral of the likelihood function on a rectangular region then simplifies as

$$L_{(k_{0s},\alpha_s)}^{(k_{0e},\alpha_e)} = \int_{k_{0s}}^{k_{0e}} \int_{\alpha_s}^{\alpha_e} L(\vec{y}; k_0, \alpha) d\alpha dk_0$$

$$= \int_{\alpha_s}^{\alpha_e} \frac{P(n+1, k_{0e} c(\alpha, \vec{y})) - P(n+1, k_{0s} c(\alpha, \vec{y}))}{c(\alpha, \vec{y})^{n+1}}$$

$$\times \exp\left[\alpha \sum_{i=1}^{n} y_i\right] d\alpha n! \prod_{i=1}^{n} m_i (f_i^{-1})'(y_i), \quad (9)$$

with

$$c(\alpha, \vec{y}) = \sum_{i=1}^{n} m_i \int_{y_{si}}^{y_i} e^{\alpha z} (f_i^{-1})'(z) dz.$$

A $(1-p)$ credible region of (k_0, α) can then be found as a rectangular area $(k_{0s}, k_{0e}) \times (\alpha_{0s}, \alpha_{0e})$ that includes the point estimates $(\hat{k}_0, \hat{\alpha})$ such that $L_{(k_{0s},\alpha_s)}^{(k_{0e},\alpha_e)} / L_{(0,0)}^{(\infty,\infty)} = 1-p$. Alternatively k_0 and α can be restricted to a finite, more physically feasible region, and it is indeed common practice to do so with Monte Carlo methods which sample only a particular domain of interest (see, e.g., [22]).

III. VALIDATION

We present below three applications: a synthetic experiment for the situation when α is fixed and known, another synthetic experiment to check the suitability of the linear force-displacement model and the piecewise linear approximation with WLC-conforming data under a few time-constrained design strategies, and finally a real SMFS experiment with a well characterized protein.

A. Unbiased estimator $\tilde{k}_0(\alpha)$ is indeed a better estimator than the biased one

In order to confirm the theoretical prediction that $\tilde{k}_0(\alpha)$ is not only unbiased but also a better estimator than $\hat{k}_0(\alpha)$ in terms of showing smaller root-mean-square error (RMSE), we simulated pulling a multimeric construct made up of 20 identical modules whose length, spontaneous unfolding rate, and position of the transition state were chosen to match those previously reported for a real protein, namely, the B1 immunoglobulin-binding domain of protein G from streptococcus (GB1) [47].

Therefore, we used $k_0 = 0.039\ \mathrm{s}^{-1}$, $x_\beta = 0.17$ nm, $T = 301.15$ K, and for each n in 2,…,101 we generated 10 000 data sets of n unfolding events each following the linear force vs displacement characteristic with a cantilever spring constant of 0.07 N/m. The n events were generated giving roughly equal shares to each of the velocities 2^{-i} 2180 nm/s, $i = 0, \ldots, 5$. The starting force of pulling was randomly chosen within a range compatible with what is commonly observed experimentally, and the number of not-yet-unfolded modules was varied between 1 and 20.

Under these conditions the observed mean of $\tilde{k}_0(\alpha)$ was always very close to the theoretical expected value of 0.039, even for $n=2$, while the observed RMSE went down from about $0.0744\ \mathrm{s}^{-1}$ for $n=2$ to about $0.0039\ \mathrm{s}^{-1}$ for $n=101$

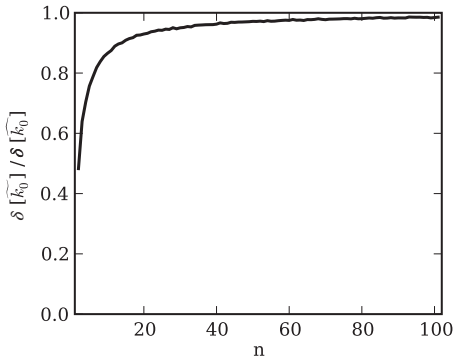

FIG. 1. Ratio of $\delta[\widetilde{k_0}(\alpha)]$ over $\delta[\widehat{k_0}(\alpha)]$ as a function of the number of unfolding events, where δ stands for root-mean-square error.

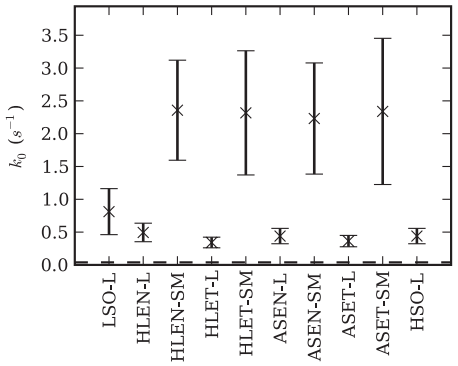

FIG. 2. Estimation of k_0 using the linear force-displacement model. Data were generated using the strategies LSO, HLEN, HLET, ASEN, ASET, and HSO, from left to right. Estimation was performed using maximum likelihood (L) and the standard method (SM). The dashed horizontal line indicates the actual k_0 value used for data generation. The crosses mark the mean and the error bars extend one standard deviation in both directions.

(data not shown). The biased estimator instead showed large bias for small values of n and higher RMSE all throughout (see Fig. 1) thus confirming the theoretical prediction that the unbiased estimator $\widetilde{k_0}(\alpha)$ is better than the biased estimator $\widehat{k_0}(\alpha)$.

B. Synthetic WLC data are well approximated by the piecewise linear function

We performed a more comprehensive simulation in order to compare the ability to recover the kinetic parameters with the various approaches discussed so far, when the data are generated by using the WLC model. We focused specifically on the importance of the intermediate data points from the start of the stretching process up to the rupture event, which are roughly approximated by the piecewise linear force-time function, but not taken into account by the widespread linear force-displacement model.

A second goal of this simulation was to investigate the efficiency of a few different experiment designs in terms of spreading a fixed amount of experimental time across different pulling velocities and checking which strategy leads to estimates with better performance.

The kinetic parameters and the cantilever spring constant were kept the same as in the previous synthetic data experiment, while the protein pick-up rate was set to 100%. Each trial simulated about 107.87 s of experimental time with a surface delay of 200 ms, an approach speed of 4360 nm/s, four unfolding modules in each curve, a piezo range of 500 nm, and 2048 sample points per curve, all of which are very reasonable values commonly used in real experiments. Six retraction speeds have been used, namely, 125, 249, 545, 1090, 2180, and 4360 nm/s, and six experiment design strategies were covered: lowest speed only (LSO) with 25 curves at the lowest velocity, highest and lowest equal number (HLEN) with 22 curves at the lowest speed and 23 at the highest one, highest and lowest equal time (HLET) with 12 curves at the lowest speed and 126 at the highest one, all speeds equal number (ASEN) with 11, 11, 11, 12, 12, and 12 curves, respectively, in increasing velocity order, all speeds equal time (ASET) with 4, 7, 15, 24, 34, and 42 curves, respectively, in increasing velocity order, and finally highest speed only (HSO) with 251 curves at the highest velocity. The number of trials was 1000 and all the data were generated using the WLC interpolation formula of Eq. (A2) with a persistence length of 0.35 nm.

The data were analyzed using our ML framework with the following force function types: linear force-displacement (L), piecewise linear force (PL), WLC, and finally Gaussian piecewise linear force (GPL). For the last-mentioned one we kept the rupture force unchanged, to allow for a reasonably fair comparison to the other approaches, but added noise with a standard deviation of 20 pN [22] to all the other data points in order to check how well the piecewise linear force approximation is able to tackle noise by selecting only the longest increasing subsequence of force values or indeed how much the remaining inaccuracies matter.

Additionally, we also analyzed the HLEN, HLET, ASEN, and ASET data sets using the SM as reviewed in [16] and the MC method analyzing the speed dependence of the unfolding force as reviewed in [18,20–23]. Briefly, the standard method consists in fitting a linear dependence between the most probable rupture force and the logarithm of the loading rate κv. The two kinetic parameters are then computed from the slope and intercept of the fitting line implicitly adopting the assumption that the force behavior during stretching can be satisfactorily approximated by the linear force-displacement characteristic. The Monte Carlo method instead consists in the simulation of thousands of synthetic curves on a two-dimensional grid of k_0 and x_β parameters and then selecting the combination of parameters that best match the experimental mean unfolding force dependence on velocity.

The SM approach produced the worst results where applicable, next followed by the L approach, the results of both being displayed in Figs. 2 and 3. The observed bias was around one to two orders of magnitude for k_0 and not too small for x_β either, therefore, raising a signal flag about the dangers of applying the wrong theoretical force model, in this case using the linear force-displacement characteristic when the underlying data have been generated using the WLC model. The assumed general applicability of the linear

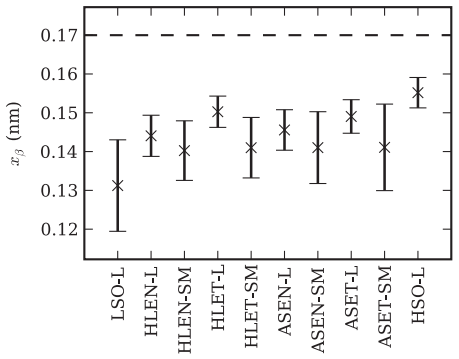

FIG. 3. Estimation of x_β using the linear force-displacement model. Data were generated using the strategies LSO, HLEN, HLET, ASEN, ASET, and HSO, from left to right. Estimation was performed using maximum likelihood (L) and the standard method (SM). The dashed horizontal line indicates the actual x_β value used for data generation. The crosses mark the mean and the error bars extend one standard deviation in both directions.

force-displacement characteristic has also been previously disproved with experimental data in [25].

Figures 4 and 5 contain the results for the other methods. Except for a little bias at the lowest speed, the ML-based WLC approach worked very well, as expected since the data were generated with the same model, and no noise was added.

Surprisingly the performance of the PL approach was almost indistinguishable from that of the WLC-assuming approach. This is quite significant since the PL approach does not take into account the fact that the data were generated with WLC suggesting that the results might be just as good with data conforming to any other theoretical model. It means that the time resolution in our synthetic data, which is typical of AFM instrumentation, is high enough so that the error performed by making a piecewise linear approximation to the WLC curve is negligible.

GPL, which is identical to PL except that it receives noisy input, also gave good results, although the estimates were noticeably more biased than the WLC or PL ones throughout all design strategies, while the variance was slightly larger.

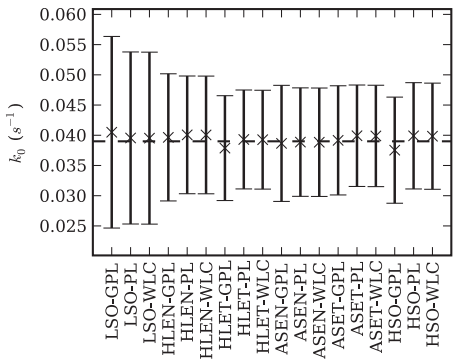

FIG. 4. Estimation of k_0 using GPL, PL, and WLC. Data were generated using the strategies LSO, HLEN, HLET, ASEN, ASET, and HSO, from left to right. The dashed horizontal line indicates the actual k_0 value used for data generation. The crosses mark the mean and the error bars extend one standard deviation in both directions.

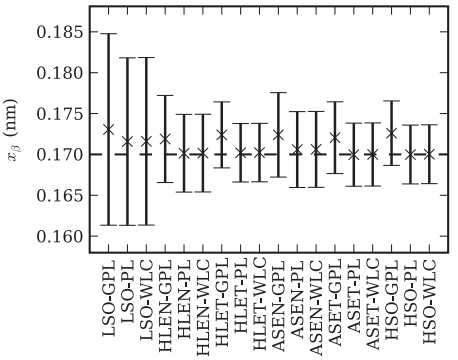

FIG. 5. Estimation of x_β using GPL, PL, and WLC. Data were generated using the strategies LSO, HLEN, HLET, ASEN, ASET, and HSO, from left to right. The dashed horizontal line indicates the actual x_β value used for data generation. The crosses mark the mean and the error bars extend one standard deviation in both directions.

The bias in the GPL k_0 estimate ranged from 2.3% to 3.5%, while for x_β from 1.1% to 2.0%, which in the presence of noise can be considered as very small.

Excluding the unsatisfactory L approach, for each of the other three ML-based approaches (GPL, PL, and WLC) the HLET experiment design strategy showed the smallest RMSE for both k_0 and α when compared to the other five design strategies (LSO, HLEN, ASEN, ASET, and HSO) covered in our simulation. That suggests that a very efficient experiment design consists in equally splitting the experimental time across two velocities, one very high and one very low.

Since the Monte Carlo method implies the same WLC model also used to generate the synthetic data of the simulated time-constrained experiments, the Monte Carlo method performed quite well (Figs. 6 and 7). The best experiment design strategy for MC turned out to be HLEN instead of HLET, followed closely by the latter one, for both k_0 and x_β. Comparing the results of the best design strategy of each approach, the Monte Carlo method achieved an RMSE about 50% higher for k_0 and 128% higher for x_β compared to our

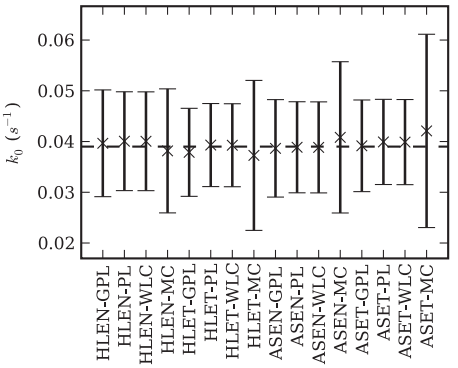

FIG. 6. Estimation of k_0 using GPL, PL, WLC, and MC. The GPL, PL, and WLC data are the same as in Fig. 4 and are reproduced here for easy visual comparison. Data were generated using the strategies HLEN, HLET, ASEN, and ASET, from left to right. The dashed horizontal line indicates the actual k_0 value used for data generation. The crosses mark the mean and the error bars extend one standard deviation in both directions.

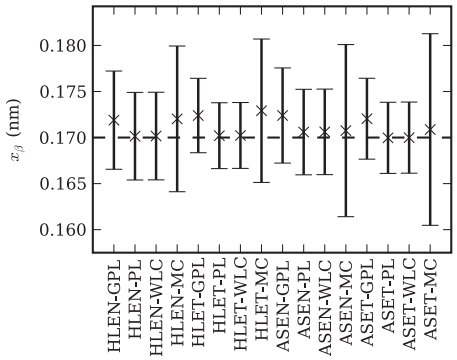

FIG. 7. Estimation of x_β using GPL, PL, WLC, and MC. The GPL, PL, and WLC data are the same as in Fig. 5 and are reproduced here for easy visual comparison. Data were generated using the strategies HLEN, HLET, ASEN, and ASET, from left to right. The dashed horizontal line indicates the actual x_β value used for data generation. The crosses mark the mean and the error bars extend one standard deviation in both directions.

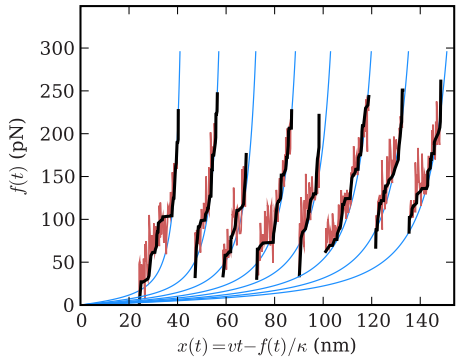

FIG. 8. (Color online) Experimental curve obtained by pulling the $(GB1)_8$ construct at $v=2180$ nm/s. The light blue (light gray) lines meeting in the origin represent WLC fits. The points interpolated by the thin red (medium gray) lines represent AFM force readings. The thick black lines are piecewise linear approximations of the longest increasing subsequences of force values. The detachment peak and nonspecific interactions at the start are not shown.

ML-based WLC and PL approaches, and about 40% higher for k_0 and 74% higher for x_β compared to GPL, that is, the piecewise linear method with noisy input. We would like to stress once more that while the MC approach in our simulation had access to the exact force model (WLC) with the exact parameters (contour length and persistence length) used for data generation, for GPL noise was present in the data and no information whatsoever about the underlying model was available. The difference in statistical performance is only expected to increase in real experimental settings where the force response is generated by the (linker-)protein-cantilever system rather than a unique, known theoretical model.

We can draw four conclusions from this synthetic data experiment. First, the general applicability of the linear force-displacement characteristic is disproved, whether applied through either the standard method or the maximum likelihood framework. Second, the piecewise linear approximation works well with WLC-conforming data. Third, for a fixed amount of experimental time, using only the highest feasible velocity to get as many unfolding events as possible is not the best experiment design strategy; instead it is better to allot half of the experimental time to a much lower velocity. Fourth, the ML approach, including the model-independent piecewise linear approximation, is better than the Monte Carlo method even when a unique force model exists and is known, as was the case in our simulation, and the advantage remains solid when noise is added only to the input of the piecewise linear approach.

C. GB1 kinetic parameters were correctly recovered from an SMFS experiment with polyprotein $(GB1)_8$

We further tested our ML approach with experimental data we obtained by pulling a multimeric construct consisting of eight GB1 modules [27,47,48]. For the experiment, a drop of the $(GB1)_8$-containing solution (20 μL, $\sim$0.1 g/L) was deposited on a flame cleaned glass coverslip for about 30 min. The velocity-clamp mechanical unfolding SMFS experiment was performed using Picoforce AFM with Nanoscope IIIa controller (Digital Instruments, Plainview, NY, USA) with a V-shaped silicon nitride cantilever (NP; Digital Instruments) whose spring constant was calibrated by the thermal noise method [49]. The buffer used was Tris/HCl (10 mM, pH 7.5).

We used the open source project Hooke [50] with locally made modifications to extract the relevant information from the AFM-recorded files, after which we applied an automated filtering step mostly based on the protocol specified in [21]. A total of 250 unfolding events passed the filtering stage, about a quarter of them at a retraction velocity of 125 nm/s, a quarter at 249 nm/s, and half at 2180 nm/s.

Making the WLC assumption we obtained the point estimates $k_0 \simeq 0.0475$ s^{-1}, $x_\beta \simeq 0.1661$ nm, and a 70%-credible region of $(0.0415, 0.0653)$ s$^{-1} \times (0.1583, 0.1687)$ nm, in very good agreement with the values reported in the literature of 0.039 s^{-1} and 0.17 nm that had been previously extracted from a larger data set consisting of 1826 unfolding events via WLC-assuming Monte Carlo simulations [47]. Using the piecewise linear force-time approach we obtained instead the point estimates $k_0 \simeq 0.0622$ s^{-1}, $x_\beta \simeq 0.1627$ nm, and a 70%-credible region of $(0.0466, 0.0777)$ s$^{-1} \times (0.1576, 0.1677)$ nm. Each of the two credible regions contains the point estimates obtained with both approaches. Figure 8 shows the piecewise linear approximation and the WLC fits for a $(GB1)_8$ curve with eight unfolding events from our experimental data set.

Using the WLC-based Monte Carlo method we obtained similar values of $k_0 \simeq 0.05$ s^{-1} and $x_\beta \simeq 0.16$ nm. Instead, using the standard method for the estimation task we obtained $k_0 \simeq 8.1440$ s^{-1}, which is two orders of magnitude larger than expected, and also a smaller distance to the transition state $x_\beta \simeq 0.1302$ nm.

We conclude that when WLC describes well the experimental data, as is the case with GB1 [47], the estimates obtained via the piecewise linear approximation and those extracted by making the WLC assumption are compatible within statistical uncertainty, while the k_0 estimate computed

via the standard method can be off by a couple of orders of magnitude.

IV. SUMMARY

We have set forth an ML framework for the analysis of SMFS experiments with homomeric polyproteins, where the protein kinetic parameters of the monomeric module are of interest. For the restricted case when only the spontaneous dissociation rate is unknown, we found an unbiased estimator that is also more efficient than the plain ML estimator.

To account for the heterogeneity of force behaviors seen in SMFS experiments we propose a piecewise linear approximation to the forces recorded by the AFM during stretching and showed via extensive simulation that the approach is able to correctly recover both kinetic parameters. That obviates the need to assume a predetermined force increase model which, as our tests show, can result in large estimation errors if the wrong one is chosen, thus, disproving the widespread practice of assuming the linear force-displacement characteristic.

Our framework does, however, allow one to specify a predetermined force model, and we validated this use case in the context of the WLC model with both synthetic and experimental data. The latter was obtained by pulling a polyprotein made up of identical tandem repeats of a protein domain that had been previously characterized via WLC-assuming Monte Carlo simulations. By imposing the same WLC model, from our data set we recovered almost the same kinetic parameters under the ML framework and also by applying the Monte Carlo method. Since WLC describes well the behavior of $(GB1)_8$ [47], this confirms the correctness of both our Monte Carlo implementation and the ML framework we propose. In order to compare the statistical properties of our approach against the Monte Carlo method we turned to synthetic data experiments which show that the WLC-assuming ML estimators are better than the WLC-assuming Monte Carlo estimators in terms of RMSE.

Without imposing a theoretical force model instead we obtained from the experimental data a slightly larger spontaneous unfolding rate, but still within the 70%-credible region of the WLC-based estimator. The compatibility, but with some difference, between the two sets of estimates can be at least partly attributed to the ability of the WLC model [strictly speaking, the approximation formula of Eq. (A2)] to describe well, but not perfectly, the behavior of the studied protein. This constitutes an approach for testing the applicability of a theoretical force model to any protein when geometric curve fits by themselves do not provide a definitive answer.

To do a comparison in terms of statistical performance with the Monte Carlo method we turned again to synthetic data experiments which proved that our ML approach, even when used without any information about the underlying force increase model, and with noise added to the data, performs better than the Monte Carlo method configured with the correct force increase model that was used for data generation, with the correct parameters so that no fitting is necessary. This is a clear proof of the superiority of the ML estimation: it requires less information as input while at the same time leading to better estimators even under clearly disadvantageous conditions.

Finally we approached the problem of long experimental times in two ways. First, by using ML estimation one is likely to need fewer rupture events for the estimation task compared to traditional approaches that do not benefit from the *likelihood principle* that guarantees that no information is lost. Second, our synthetic data experiments suggest that rather than using the highest feasible pulling velocity to get as many unfolding events as possible, it is more efficient to also use one much lower retraction velocity and to allot to the two velocities equal shares of the experimental time.

V. CONCLUSIONS

We conclude with a short review of the advantages our proposed method brings over existing ones:

(1) It is particularly well suited for the analysis of experiments where a master curve cannot be easily identified such as those involving ligand-receptor complexes [25] (see Appendix A 1).

(2) It is the only method for which an unbiased estimator of the dissociation rate has been provided when the distance to the transition state is known, thus, making it particularly attractive for the analysis of protein unfolding under the effect of certain chemical denaturants [41], protein mutations [22,39,40], and different solvents of equal molecule size [38] [see Eq. (8)].

(3) In all the tested settings it leads to better estimators of the kinetic parameters in terms of RMSE when compared to existing methods even under disadvantageous conditions (see Sec. III B).

(4) It comes with a clear recommendation about how to design experiments based on the well accepted statistical criterion of reducing the RMSE of the obtained estimators (see Sec. III B).

(5) As a simple numerical maximization of a univariate function [see Eq. (7)], point estimation is very fast in practice, in our experience orders of magnitude faster than the more established Monte Carlo method which instead requires extensive data generation on a two-dimensional grid of parameters.

Because of the generality of the last three mentioned advantages we recommend our approach as the method of choice in the analysis of all velocity-clamp experiments with polymers made up of one or more identical domains.

ACKNOWLEDGMENTS

We thank Professor Hongbin Li, University of British Columbia, Vancouver, Canada for kindly providing the $(GB1)_8$ plasmid, Dr. Isabella Tessari and Professor Luigi Bubacco, University of Padova, Italy for providing the $(GB1)_8$ construct, and Dr. Aldo Rampioni, University of Bologna, Italy for helpful comments on the paper.

APPENDIX A: APPLICATION TO SOME FORCE-TIME FUNCTIONS

We show next how the likelihood function can be computed with two force models and the piecewise linear approximation.

1. Linear force-displacement characteristic

Under the linear force vs displacement characteristic $f(t)=\kappa v t$ the likelihood function in Eq. (5) becomes

$$L(\vec{y};k_0,\alpha)=\exp\left[-\frac{k_0}{\alpha}\sum_{i=1}^{n}\frac{m_i}{\kappa_i v_i}(e^{\alpha y_i}-e^{\alpha y_{si}})\right]k_0^n\prod_{i=1}^{n}\frac{m_i}{\kappa_i v_i}e^{\alpha y_i}. \tag{A1}$$

2. Wormlike chain

The wormlike chain describes the force dependence on the distance over contour length ratio, and we adopt the well-known interpolation formula with less than 10% error proposed in [37]:

$$f(t)=\frac{K_bT}{4p}\{[1-x(t)/L_c]^{-2}+4x(t)/L_c-1\}, \tag{A2}$$

where $x(t)=vt-f(t)/\kappa$ is the distance at time t [51]. By substitution we get the cubic equation

$$-4az^3+[(12a/b+4)y+9a]z^2-[y^2(12a/b+8)/b+y(18a/b+8)+6a]z+y^3(4a/b+4)/b^2+y^2(9a/b+8)/b+y(6a/b+4)=0, \tag{A3}$$

where $a=K_bT/p$, $b=\kappa L_c$, and $z=(f^{-1})(y)v/L_c$.

It is possible to show that there is exactly one root in the interval of interest $(y/b,1+y/b)$, so $(f^{-1})(y)$ can be obtained without ambiguity. Many ways to solve the cubic polynomial equation exist including closed-form solutions [52]. Then $(f^{-1})'(y)$ can be computed by implicit differentiation, thus, making possible the numerical computation of the likelihood function using Eq. (5).

3. Piecewise linear force

The problem of finding the longest increasing subsequence of a sequence is classical in computer science and for this purpose we use an $\mathcal{O}(n\log n)$ algorithm as described, e.g., in [42]. Now let us consider one unfolding event with the longest increasing subsequence of forces $y_1,\ldots,y_p$ at increasing times $t_1,\ldots,t_p$, with y_p being the rupture force. The piecewise linear force-time function is then assembled as

$$f(t)=\begin{cases}y_j+(t-t_j)(y_{j+1}-y_j)/(t_{j+1}-t_j), & \text{if } t_j\le t<t_{j+1},\quad 1\le j<p-1\\ y_{p-1}+(t-t_{p-1})(y_p-y_{p-1})/(t_p-t_{p-1}), & \text{if } t\ge t_{p-1}.\end{cases}$$

Assuming that unfolding event i has longest increasing subsequence of force values $(t_{i1},y_{i1}),(t_{i2},y_{i2})\ldots(t_{ip_i},y_{ip_i})$ with the connection to the notation throughout the rest of the paper being that $(t_{i1},y_{i1})=(t_{si},y_{si})$ and $(t_{ip_i},y_{ip_i})=(t_i,y_i)$, the likelihood function can be written as

$$L(\vec{y};k_0,\alpha)=\exp\left[-\frac{k_0}{\alpha}\sum_{i=1}^{n}m_i\sum_{j=1}^{p_i-1}\frac{t_{ij+1}-t_{ij}}{y_{ij+1}-y_{ij}}(e^{\alpha y_{ij+1}}-e^{\alpha y_{ij}})\right]\times k_0^n\prod_{i=1}^{n}m_i\frac{t_{p_i}-t_{p_i-1}}{y_{p_i}-y_{p_i-1}}e^{\alpha y_{ip_i}}. \tag{A4}$$

APPENDIX B: EXPECTATION OF $\widehat{k_0}(\alpha)$ FOR α FIXED

Proposition 1. For $\alpha>0$ fixed and $n\ge 2$, if $\int_{y_{si}}^{\infty}e^{\alpha z}(f_i^{-1})'(z)dz=\infty$ for all $1\le i\le n$, then $E[\widehat{k_0}(\alpha)]=nk_0/(n-1)$.

To prove the above result we start from the definition

$$E[\widehat{k_0}(\alpha)]=\int_{y_1\ge y_{s1}\ldots y_n\ge y_{sn}}\widehat{k_0}(\alpha)L(\vec{y};k_0,\alpha)dy_1\ldots dy_n.$$

Using Eqs. (5) and (6), and making the changes of variables

$$x_i=k_0m_i\int_{y_{si}}^{y_i}e^{\alpha z}(f_i^{-1})'(z)dz,\quad 1\le i\le n$$

we obtain

$$E[\widehat{k_0}(\alpha)]=nk_0\int_{x_1\ge 0\ldots x_n\ge 0}\frac{\exp\left[-\sum_{i=1}^{n}x_i\right]}{\sum_{i=1}^{n}x_i}dx_1\ldots dx_n,$$

where the multiple integral can be computed as $1/(n-1)$, thus, obtaining the desired result.

[1] M. Rief, J. M. Fernandez, and H. E. Gaub, Phys. Rev. Lett. **81**, 4764 (1998).
[2] M. Carrion-Vazquez, A. F. Oberhauser, T. E. Fisher, P. E. Marszalek, H. Li, and J. M. Fernandez, Prog. Biophys. Mol. Biol. **74**, 63 (2000).
[3] T. E. Fisher, P. E. Marszalek, A. F. Oberhauser, M. Carrion-Vazquez, and J. M. Fernandez, J. Physiol. (London) **520**, 5 (1999).
[4] W. A. Linke and A. Grutzner, Pfluegers Arch. Eur. J. Physiol. **456**, 101 (2008).
[5] T. E. Fisher, A. F. Oberhauser, M. Carrion-Vazquez, P. E. Marszalek, and J. M. Fernandez, Trends Biochem. Sci. **24**, 379 (1999).
[6] K. Wang, J. G. Forbes, and A. J. Jin, Prog. Biophys. Mol. Biol. **77**, 1 (2001).
[7] A. F. Oberhauser and M. Carrion-Vazquez, J. Biol. Chem. **283**, 6617 (2008).
[8] T. E. Fisher, P. E. Marszalek, and J. M. Fernandez, Nat. Struct. Biol. **7**, 719 (2000).
[9] I. Tinoco, Jr., P. T. X. Li, and C. Bustamante, Q. Rev. Biophys. **39**, 325 (2006).
[10] I. Tinoco, Jr., Annu. Rev. Biophys. Biomol. Struct. **33**, 363 (2004).
[11] T. E. Fisher, M. Carrion-Vazquez, A. F. Oberhauser, H. Li, P. E. Marszalek, and J. M. Fernandez, Neuron **27**, 435 (2000).
[12] T. R. Strick, J. F. Allemand, D. Bensimon, and V. Croquette, Annu. Rev. Biophys. Biomol. Struct. **29**, 523 (2000).
[13] A. Janshoff, M. Neitzert, Y. Oberdorfer, and H. Fuchs, Angew. Chem. **39**, 3212 (2000).
[14] E. Evans and K. Ritchie, Biophys. J. **72**, 1541 (1997).
[15] G. I. Bell, Science **200**, 618 (1978).
[16] M. Raible, M. Evstigneev, F. W. Bartels, R. Eckel, M. Nguyen-Duong, R. Merkel, R. Ros, D. Anselmetti, and P. Reimann, Biophys. J. **90**, 3851 (2006).
[17] M. Evstigneev and P. Reimann, Phys. Rev. E **68**, 045103(R) (2003).
[18] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, Science **276**, 1109 (1997).
[19] M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **96**, 3694 (1999).
[20] R. B. Best and J. Clarke, Chem. Commun. (Cambridge, U.K.) 2002, 183.
[21] R. Rounsevell, J. R. Forman, and J. Clarke, Methods **34**, 100 (2004).
[22] R. B. Best, S. B. Fowler, J. L. Toca-Herrera, and J. Clarke, Proc. Natl. Acad. Sci. U.S.A. **99**, 12143 (2002).
[23] A. F. Oberhauser, P. E. Marszalek, H. P. Erickson, and J. M. Fernandez, Nature (London) **393**, 181 (1998).
[24] S. Getfert and P. Reimann, Phys. Rev. E **76**, 052901 (2007).
[25] A. Fuhrmann, D. Anselmetti, R. Ros, S. Getfert, and P. Reimann, Phys. Rev. E **77**, 031912 (2008).
[26] M. Carrion-Vazquez, P. E. Marszalek, A. F. Oberhauser, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **96**, 11288 (1999).
[27] Y. Cao and H. Li, Nature Mater. **6**, 109 (2007).
[28] T. Hugel, M. Rief, M. Seitz, H. E. Gaub, and R. R. Netz, Phys. Rev. Lett. **94**, 048301 (2005).
[29] E. M. Puchner, G. Franzen, M. Gautel, and H. E. Gaub, Biophys. J. **95**, 426 (2008).
[30] R. Merkel, Phys. Rep. **346**, 343 (2001).
[31] S. Cui, Y. Yu, and Z. Lin, Polymer **50**, 930 (2009).
[32] C. Bouchiat, M. D. Wang, J. Allemand, T. Strick, S. M. Block, and V. Croquette, Biophys. J. **76**, 409 (1999).
[33] R. W. Ogden, G. Saccomandi, and I. Sgura, Comput. Math. Appl. **53**, 276 (2007).
[34] R. B. Best, D. J. Brockwell, J. L. Toca-Herrera, A. W. Blake, D. A. Smith, S. E. Radford, and J. Clarke, Anal. Chim. Acta **479**, 87 (2003).
[35] P. Garthwaite, I. Jolliffe, and B. Jones, *Statistical Inference*, 2nd ed. (Oxford University Press, New York, 2002).
[36] J. F. Marko and E. D. Siggia, Macromolecules **28**, 8759 (1995).
[37] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, Science **265**, 1599 (1994).
[38] L. Dougan, G. Feng, H. Lu, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **105**, 3185 (2008).
[39] R. B. Best, S. B. Fowler, J. L. T. Herrera, A. Steward, E. Paci, and J. Clarke, J. Mol. Biol. **330**, 867 (2003).
[40] S. P. Ng, R. W. S. Rounsevell, A. Steward, C. D. Geierhaas, P. M. Williams, E. Paci, and J. Clarke, J. Mol. Biol. **350**, 776 (2005).
[41] Y. Cao and H. Li, J. Mol. Biol. **375**, 316 (2008).
[42] D. Gusfield, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, 1st ed. (Cambridge University Press, Cambridge, UK, 1997), Chap. 12.
[43] M. L. Fredman, Discrete Math. **11**, 29 (1975).
[44] A. DasGupta, *Asymptotic Theory of Statistics and Probability*, 1st ed. (Springer Science +Business Media, LLC, New York, 2008).
[45] G. A. Young and R. L. Smith, *Essentials of Statistical Inference*, 1st ed. (Cambridge University Press, Cambridge, UK, 2005).
[46] B. Hoadley, Ann. Math. Stat. **42**, 1977 (1971).
[47] Y. Cao, C. Lam, M. Wang, and H. Li, Angew. Chem. **45**, 642 (2006).
[48] H. Li, Org. Biomol. Chem. **5**, 3399 (2007).
[49] E.-L. Florin, M. Rief, H. Lehmann, M. Ludwig, C. Dornmair, V. T. Moy, and H. E. Gaub, Biosens. Bioelectron. **10**, 895 (1995).
[50] M. Sandal, F. Benedetti, M. Brucale, A. Gomez-Casado, and B. Samorì, Bioinformatics **25**, 1428 (2009).
[51] M. Carrion-Vazquez, A. F. Oberhauser, H. Diez, R. Hervas, J. Oroz, J. Fernandez, and D. Martinez-Martin, *Advanced Techniques in Biophysics* (Springer-Verlag, Berlin, 2006), pp. 163–245.
[52] R. Nickalls, Math. Gaz. **77**, 354 (1993).