Nonkinetic Modeling of the Mechanical Unfolding of Multimodular Proteins: Theory and Experiments

F. Benedetti,[†] C. Micheletti,[‡§] G. Bussi,[§] S. K. Sekatskii,[†] and G. Dietler[†]*

[†]Laboratory of Physics of Living Matter, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland; and [‡]Scuola Internazionale Superiore di Studi Avanzati and [§]Consiglio Nazionale delle Ricerche-Istituto per l'Officina dei Materiali Democritos and Italian Institute of Technology, SISSA Unit, Trieste, Italy

ABSTRACT We introduce and discuss a novel approach called back-calculation for analyzing force spectroscopy experiments on multimodular proteins. The relationship between the histograms of the unfolding forces for different peaks, corresponding to a different number of not-yet-unfolded protein modules, is exploited in such a manner that the sole distribution of the forces for one unfolding peak can be used to predict the unfolding forces for other peaks. The scheme is based on a bootstrap prediction method and does not rely on any specific kinetic model for multimodular unfolding. It is tested and validated in both theoretical/computational contexts (based on stochastic simulations) and atomic force microscopy experiments on $(GB1)_8$ multimodular protein constructs. The prediction accuracy is so high that the predicted average unfolding forces corresponding to each peak for the *GB*1 construct are within only 5 pN of the averaged directly-measured values. Experimental data are also used to illustrate how the limitations of standard kinetic models can be aptly circumvented by the proposed approach.

INTRODUCTION

During the last decade, single-molecule force spectroscopy experiments based on optical tweezers or atomic force spectroscopy have acquired increasing importance for characterizing properties of individual proteins, as well as protein complexes. Among the hundreds of such studies carried out so far, it is particularly worth mentioning force spectroscopy investigations of multimodular proteins. These constructs typically consist of a series of protein modules that are covalently linked at their ends (see, e.g., (1–12)). Upon pulling the constructs at both ends, a series of unfolding events are observed. The forces at which the unfolding events occur carry a wealth of information about the unfolding mechanics and kinetics of the construct modules (3,13,14). Customarily, this information is extracted by analyzing the force distribution obtained by gathering together the succession of unfolding forces over repeated stretching experiments and analyzing them with different methods, such as Monte Carlo simulation and regression to zero force (5,15–19). The scope and utility of these commonly employed analysis techniques can be considerably extended by examining separately the distribution of the forces associated with the first, second, etc., unfolding event in the constructs. This approach, which so far has been applied only limitedly (5,20), is particularly appropriate and useful when the construct consists of repeats of the same type of globular protein (such as *I*27 or *GB*1). In fact, because of the identical nature of the modules, it is expected that the forces associated with the various unfolding events depend on the number of unfolded modules still present on the molecule, but that the statistical distributions should nevertheless be tied by a definite relationship, as pointed out in previous studies (5,20,21,22). To the best of our knowledge, such dependence has not yet been adequately explored or exploited in experimental contexts. Furthermore, one may envisage using only the limited information contained in the experimental distribution of one single group of unfolding forces to predict with high accuracy the average unfolding forces of all other groups. This issue also has not been addressed before, and we therefore investigate it in this study.

The problem is here attacked at two levels. First we adopt a simplified analytical scheme, which implicitly relies on a standard kinetic model for the unfolding of the protein modules (Evans's theory). This method, which builds on a treatment introduced in previous studies (20,21,23), combines a transparent analytical formulation with the simplicity of implementation and use. Yet the simplifying assumptions that allow for the exact analytical treatment of the model come at a disadvantage, since the predicted probability distributions for the unfolding forces of the various peaks can be significantly different from the measured ones.

This limitation can be overcome by using the alternative and more general phenomenological approach introduced and discussed here for the first time that we know of. The scheme, based on the bootstrap statistics and termed back-calculation, is parameter-free and does not rely on any specific kinetic model. The method merely uses the probability distribution of forces associated with one of the unfolding events (the first, the second, etc.) and predicts the distribution of forces of all other events. The method is validated against data obtained from stochastic simulations (both Langevin and Monte Carlo) and from atomic force microscopy (AFM) experiments carried out on multimodular *GB*1 constructs. In all cases, the average forces associated with any

Submitted April 27, 2011, and accepted for publication July 1, 2011.

*Correspondence: giovanni.dietler@epfl.ch

Editor: Laura Finzi.

0006-3495/11/09/1504/9 $2.00

doi: 10.1016/j.bpj.2011.07.047

unfolding event are well predicted by back-calculation. Deviations from the experimental measured values are of only 5 pN, a quantity that is smaller than the uncertainty typically associated with experimental estimates for protein unfolding forces.

MATERIALS AND METHODS

Experiment

Our experiments were performed on multimeric constructs consisting of eight *GB*1 modules, hereafter denoted as $(GB1)_8$, and dissolved in Tris/HCl buffer (10 mM, pH 7.5) at a concentration of ~20 μg/ml (1,2,24). The force-extension curves of $(GB1)_8$ were measured by means of a commercially available AFM system (Picoforce AFM Nanoscope IIIa, Bruker, Madison, WI) using a V-shaped silicon nitride cantilever (NP, Bruker). The spring constant of the lever was measured from thermal fluctuation measurements (25) as part of the AFM calibration procedure and was found to be equal to 0.0575 N/m. The constructs were pulled along the x direction at the speed $x = 2180$ nm/s. Further details of the standard experimental protocol that was followed can be found in a recently published note (21).

A typical experimental force-extension curve is presented in Fig. 1. Because the AFM tip will not necessarily pick the construct at its free end, the number of modules trapped between the anchored end and the AFM tip can be <8. As a matter of fact, among the curves presenting a clear detachment peak, the most numerous group was the one displaying six unfolding peaks. We therefore limited considerations to this set of force-extension curves. The curves were analyzed using Hooke (26) an open-source software package designed to analyze the force spectroscopy curves. Hooke was further used to analyze the data from Langevin simulations of the stretching of multimodular protein constructs (see below).

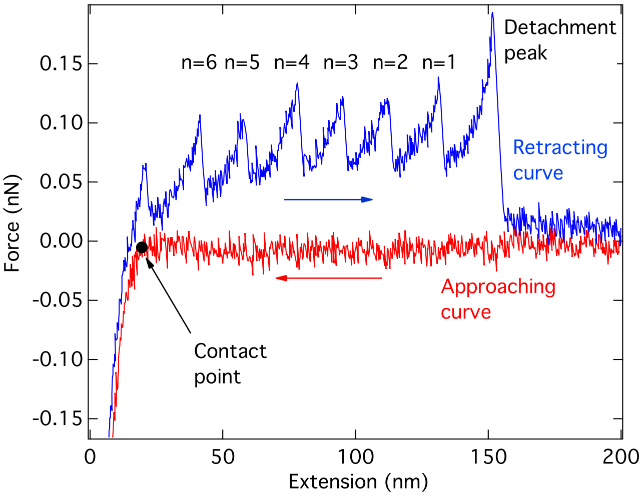

FIGURE 1 Typical force-extension curve recorded in an AFM stretching experiment on a $(GB1)_8$ construct. The lower curve shows the force as the AFM tip approaches the substrate until the contact is established. The upper curve represents the force while the tip is retracted from the substrate and shows a series of unfolding events of the construct picked up by the tip. Notice that this trace displays six unfolding events. The effect of an aspecific interaction at the beginning of the retracting trace is observed. The peak forces leading to the unfolding of the various modules are measured with respect to the background provided by the constant part of the retracting curve.

Numerical simulations

Two different computational approaches, namely Monte Carlo and Langevin simulations, were used to study the mechanical unfolding of multimodular protein constructs. In both cases, the pulled construct is assumed to be anchored at one end while the other is pulled at fixed speed. The end-to-end distance of each protein module projected along the pulling direction, x, is used as an effective order parameter to describe the module state. This corresponds to considering the system as being effectively one-dimensional, as in the sketch of Fig. 2 *a*. This is a good approximation, since for the typical unfolding forces at play in our experiments, the end-to-end distance of our construct, as obtained from a wormlike-chain (WLC) model,

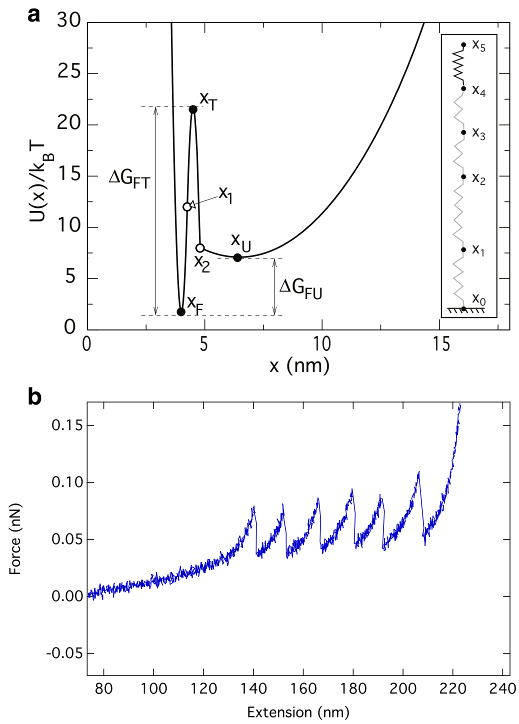

FIGURE 2 (*a*) Illustration of the anharmonic spring potential for one module of the construct in the Langevin simulation (see Eq. 1). A large value of A, equal to 100 pN $\times$ nm^2, was used in Eq. 1 to enforce the constraint that the modules cannot be stretched beyond the nominal *GB*1 contour length, $L_c = 18$ nm. The reference end-to-end separation of the folded state, x_F, is set equal to 4 nm and the the end-to-end separation between the folded state and the transition (T) state, $\Delta x = x_T - x_F$, is 0.5 nm. The reference end-to-end separation of the unfolded state, x_U, is equal to $(L_c - x_T)/2$. Consistent with what has been established in previous studies, the barrier separating the folded and transition states is set equal to $\Delta G_{FT} = 20 k_B T$, whereas the barrier between the folded and unfolded states has the value $\Delta G_{FU} = 5 k_B T$, with the temperature, T, equal to 300 K (i.e., $k_B T \approx 4.2$pN nm). For simplicity, the curvature, k_T, is set equal to k_F. The value of k_F in turn is set to $4\Delta G_{FT}/(x_T - x_F)^2 = 1344$ pN/nm to ensure the continuity of the potential and its derivative at the midpoint, $x_1 = (x_F + x_T)/2$, where the first two parabolas in Eq. 1 meet. The value of k_U was set to be much smaller than k_F, at $k_U = k_F/500 = 2.69$ pN/nm. The value of x_2 was finally obtained by the requirement of continuity of the potential. To avoid an excessive parameterization of the model, flexible linkers in the construct are described as unfolded protein modules. (*Inset*) Protein model used in Langevin simulations. (*b*) A force-extension curve obtained with the Langevin simulation applied to a model construct that initially comprised six folded modules intercalated by seven linkers (each linker has the same length as an unfolded module).

is expected to be almost equal to its contour length, so that fluctuations in the y and z directions can be neglected.

Depending on the value of the end-to-end separation, each module is considered as being folded (F) or unfolded (U); these two states are separated by a barrier of potential energy whose height is modulated by the applied tensile force. The effective potential energy, $U(x)$, is modeled explicitly in the Langevin scheme, where one integrates the stochastic equation of motion for each of the tethered modules in the construct that is being pulled. By contrast, no explicit representation of the construct is considered in the Monte Carlo approach. The latter, in fact, is employed to model the succession of discrete unfolding events occurring at force-dependent rates.

The two methods clearly embody rather different strategies for simulating the stretching experiments and, also in view of the different parameters used in the corresponding stochastic simulations, are useful to probe the generality and transferability of the back-calculation method (BC) proposed here.

A detailed description of the two methods is provided hereafter.

Langevin simulations

With reference to the sketch in the inset of Fig. 2 *a*, the anchored end of the construct is located at $x_0 = 0$, while the other end (x_4 in the sketch) is attached to the moving AFM tip. For simplicity, to parallel what is done in the Monte Carlo scheme below, the latter is modeled as a Hookean spring (x_4–x_5 in the sketch) with spring constant $k_{\mathrm{AFM}} = 0.01$ N/m. Each protein module behaves as an anharmonic spring; the associated free-energy profile, $U(x)$, is shown in Fig. 2 *a* and described by the expression

$$U(x) = \frac{A}{L_c - x} + \begin{cases} \frac{1}{2}k_F(x - x_F)^2 \text{ if } x<x_1 \\ \Delta G_{FT} - \frac{1}{2}k_T(x - x_T)^2 \text{ if } x_1 < x < x_2 \\ \Delta G_{FU} + \frac{1}{2}k_U(x - x_U)^2 \text{ otherwise} \end{cases}. \tag{1}$$

The model parameters are chosen to be consistent with the overall shape of the potential energy typically found in proteins (27) and are provided in the caption to Fig. 2. In particular, the contour length of each module is equal to the nominal contour length of *GB*1, $L_c = 18$ nm, the reference end-to-end separation of the folded state is $x_F = 4$ nm, and its distance to the transition state is $\Delta x = 0.5$ nm.

In multimodular protein constructs, each protein module is connected to the next via a short peptidic linker of length 1.5 nm. To keep at a minimum the number of parameters in the model, we described these linkers, which clearly do not undergo any transition upon stretching, by unfolded protein modules. To do so, we initially prepared the pristine construct as a succession of folded modules with initial end-to-end separation equal to x_F, intercalated with unfolded modules with initial end-to-end separation equal to $x_{U.}$ The potential energy barrier separating the F and U states is sufficiently high that an initially unfolded module will not spontaneously refold over the short timespan of the model stretching experiment.

The total potential energy of the homomeric module chain composed of n protein modules, ℓ linkers, and the AFM tip is given by

$$\begin{aligned} H(x_1, x_2, ..x_n, x_{n+\ell+1}) &= \sum_{i=1}^{n+\ell} U(x_i - x_{i-1}) \\ &\quad + \frac{1}{2}k_{\mathrm{AFM}}(x_{n+\ell+1} - x_{n+\ell})^2. \end{aligned} \tag{2}$$

The time evolution of the key construct positions, $x_{i=1,\ldots,n+\ell}$, follows the overdamped Langevin dynamics:

$$\gamma \dot{x}_i = -\frac{\partial H}{\partial x_i} + \eta(t), \tag{3}$$

where $\gamma = 4.4 \times 10^{-5}$ pN s/nm is the friction coefficient appropriate to yield (according to Kramer's theory) a spontaneous unfolding rate (at zero applied force) equal to $k_{\mathrm{off}} = 10^{-2}$ s^{-1}. $\eta(t)$ is a Gaussian white noise with zero mean and variance equal to $2k_{\mathrm{B}}T\gamma$ (k_{B} is the Boltzmann constant and $T = 300$ K is the temperature). Notice that the derivative of the potential U (entailed by the derivative of H) is not continuous in x_2.

The stochastic equations of motion were integrated numerically with a time step of 1 ns. After an initial equilibration, the position of the AFM tip, $x_{n+\ell+1}$ is moved at constant velocity, $x_{n+\ell+1}(t) = vt$, with $\nu = 500$ nm/s. This velocity value is commonly employed in stretching simulations and falls in the typical range of pulling velocities used in experiments (28). The typical time span required to unfold all the $n = 6$ modules in the constant-velocity simulation was 0.25 s.

The force/extension curve of the system is obtained by recording the restoring force experienced by the AFM tip, $f = k_{\mathrm{AFM}}(x_{n+\ell+1} - x_{n+\ell})$, as a function of the AFM tip position, $x_{n+\ell+1}$, as shown in Fig. 2 *b*. Several hundred such curves were collected and analyzed with Hooke after performing a time average over windows of duration 0.15 ms to mimic the finite time resolution of a typical experiment.

Monte Carlo simulations

As anticipated at the beginning of the section, the Monte Carlo approach (here implemented as in studies by Rief and colleagues (28,29) and Zinober and colleagues (30)) provides a phenomenological approach to the kinetics of mechanical unfolding. The advantage of its transparent formulation is balanced by the highly simplified nature of the model. In particular, by contrast with the Langevin modeling of biopolymer stretching employed here and in other approaches (31), no explicit representation of the module constructs is considered, and the linkers are not accounted for. In addition, the pulling action is assumed to act equally on all the n modules, causing the same steady increase of the end-to-end separation for each of them. Notice that because of the limited sound velocity in the chain, this condition is only approximately realized in Langevin schemes and experiments (where other effects, such as viscosity, can be at play). In any case, the lower the pulling rate the better the approximation is expected to be.

Within the above assumption, the end-to-end distance (equal to zero at the initial time, $t = 0$) of each one of the n modules at time t is equal to $(vt - F(t)/k_{\mathrm{AFM}})/n$. In this study, we considered $n = 6$ and $v = 500$ nm/s, and the effective spring constant of the AFM tip is set to $k_{\mathrm{AFM}} = 0.01$ N/m. Notice that k_{AFM} is smaller than the nominal spring constant of the tip used in our typical stretching experiments. This is because k_{AFM} stands for an effective spring that, in addition to the AFM tip, accounts for stiffness of the folded modules, which are not explicitly included in our Monte Carlo scheme. With this simplified description, the loading rate is not dependent on the number of folded modules, which brings the Monte Carlo closer to the BC assumptions, as discussed later. We underline that our goal here is to provide a benchmark for the BC and not to reproduce the experimental data, so a qualitative picture is satisfactory at this stage. The instantaneous force experienced by each module is computed from the theoretical force-extension curve, $f_{\mathrm{WLC}}(x)$ of an equilibrated WLC with contour length $L_c = 18$ nm (appropriate for *GB*1) and a persistence length of $l_p = 0.4$ nm. The progressive loading of the modules is followed at time increments of duration $\Delta t = 1.6 \times 10^{-5}$ s. At the (discrete) time, t, the probability that one of the modules yields and becomes unfolded is computed within the Evans approximation (13) disregarding the refolding probability:

$$p(t) = k_{\mathrm{off}} \exp\left(\frac{f_{\mathrm{WLC}}(vt) \times \Delta x}{k_{\mathrm{B}}T}\right)\Delta t, \tag{4}$$

where k_{B} is the Boltzmann constant and $T = 300$ K is the system temperature. The effective values $k_{\mathrm{off}} = 0.11$ s^{-1} and $\Delta x = 1.44$ Å are obtained from a fit of the experimental data using Evans's theory as in Benedetti et al. (20). The fitting procedure ensures that the unfolding forces fall in a range similar to the experimental ones, although a precise match is neither expected nor sought. The Monte Carlo scheme consists of drawing a random number, uniformly distributed in the [0,1] interval, for each of the *n* protein modules and comparing it with *p*(*t*). An unfolding event occurs when one of the *n* random numbers is smaller than *p*(*t*). The associated unfolding force is recorded and the calculation is next repeated with the *n* – 1 modules. The statistical distribution of the unfolding forces for each value of *n* was obtained from 1000 repeats of the Monte Carlo unfolding simulations.

Analytically solvable model

Simple analytical expressions for the probability distributions of the unfolding forces, and the associated mean values and variance, as a function of the number of domains, *n*, can be obtained by introducing a further simplification besides the ones introduced for the Monte Carlo scheme. Specifically, each protein module is treated as a harmonic spring (as in the Langevin approach) rather than a WLC, and the unfolding process follows the Bell-Evans theory. Within these assumptions, the probability distribution of unfolding forces has been previously worked out both for single-chain stretching (see, e.g., Hummer and Szabo (23)) and for multimodular constructs (20,21). For completeness, and for the purposes of better discussing the phenomenological BC method, an analogous derivation is provided here.

Let us consider a model construct consisting of $n_0 + \ell$ harmonic springs: the n_0 initially folded modules have spring constant equal to k_F, whereas the remaining ℓ have a smaller spring constant, $k_{U,}$ as appropriate for unfolded modules. The model construct is subject to the AFM pulling action (the AFM tip is again modeled as a harmonic spring with constant $k_{\rm AFM}$). Because the tip is pulled at constant velocity, *v*, the tensile force experienced at time *t* by each construct is equal to

$$f(t) = \frac{vt}{n/k_F + (n_0 - n + \ell)/k_U + 1/k_{\rm AFM}} \equiv k_{\rm eff} vt, \qquad (5)$$

where $k_{\rm eff}$ is the effective spring constant of the construct in series with the AFM tip and its inverse decreases with *n* as $k_{eff}^{-1} = (n_0 - n + \ell)k_U^{-1} + k_{\rm AFM}^{-1} + nk_F^{-1} \equiv \bar{k}^{-1}(1 - An)$. Here $\bar{k}^{-1}$ is the inverse spring constant of the completely unfolded construct, and *A* is a correction term that describes the dependence of the spring constant on the number of folded modules.

Following Evans's theory, the survival probability that any one module has remained folded up to time *t* is equal to (23)

$$S_1(t) \equiv \exp\left[-\int_0^{f(t)} \frac{k_{\rm off} e^{\frac{f\Delta x}{k_BT}}}{vk_{\rm eff}} df\right]. \qquad (6)$$

The probability that all the *n* modules have remained folded up to time *t*, or equivalently up to the loading force $f(t) = vt\, k_{\rm eff}$, is simply obtained by raising the above expression to the power *n*,

$$S_n(t) = S_1(t)^n. \qquad (7)$$

By differentiating S_n with respect to *f*, one obtains the probability distribution, $p_n(f)$, for the force at which the first unfolding event occurs in a chain of *n* modules.

The sought expression is

$$p_n(f) \propto \exp\left(\frac{f\Delta x}{k_BT} - \frac{n(1 - An)k_{\rm off}}{\frac{\Delta x\bar{k}v}{k_BT}} e^{\frac{f\Delta x}{k_BT}}\right), \qquad (8)$$

where the proportionality factor, containing the normalization of the probability distribution, was omitted.

Since the function above is typically nonnegligible only for positive *f*, we can compute its average and variance integrating over $[-\infty, +\infty]$, which leads to the analytical result

$$\langle f \rangle_n = -\frac{k_BT}{\Delta x}\left[\gamma + \log\frac{k_{\rm off}}{\frac{\Delta x\bar{k}v}{k_BT}} + \log\left(n - An^2\right)\right] \qquad (9)$$

$$\sigma_n^2 = \frac{\pi^2}{6\left(\frac{\Delta x}{k_BT}\right)^2} \qquad (10)$$

where $\gamma \sim 0.577$ is the Euler-Mascheroni constant.

We remark here that the variance is independent of the number of folded modules, *n*, in the construct. This result is related to the empirical observation that in typical stretching experiments of a single protein construct, the variance of the unfolding force is largely independent of the loading rate (23).

If the dependence of the spring constant on the number of folded modules can be neglected, the average unfolding force acquires a particularly simple expression:

$$\langle f \rangle_n = -\frac{\gamma}{a} - \frac{\log[bn]}{a}, \qquad (11)$$

where the parameters *a* and *b* are obtained from the average force and variance for a given *n*: $a = \pi/\sqrt{6\sigma^2}$and $b = \exp[-\gamma - a\langle f \rangle_n]/n$.

Finally, we notice that for all values of *n*, the expression of Eq. 8 corresponds to a Gumbel extremal distribution (32) with the fat tail extending toward low values of the force, *f*. Accordingly, the viability of the analytical model to capture the statistical properties of the unfolding forces measured for a given value of *n* can be ascertained by checking whether the forces follow the Gumbel distribution. To address this point, we employed the Anderson-Darling test and computed the significance level to which one can support the null hypothesis that the data originate from a Gumbel distribution. According to custom, the threshold of 5% statistical significance was used to accept or reject the null hypothesis.

Back-calculation

The previous analytical results rely on a definite kinetic model (Evans's theory) and on the harmonic modeling of the elastic response of the AFM tip and the protein modules. These effects could be included in a more general theoretical framework which, however, would not yield simple analytical calculations.

This difficulty can be circumvented using a simple and physically appealing phenomenological approach, which we term the back-calculation method, described hereafter. The method is parameter-free, as it relies on the knowledge of the empirical probability distribution of the unfolding forces at one particular value of *n*. This reference distribution can be used straightforwardly to predict the average value of the force and its variance at all other values of *n*. The scheme is best illustrated assuming that the reference distribution is the one for $n = 1$, $p_{n=1}(f)$. This distribution is directly obtained from the data gathered in the stretching experiments or from the stochastic simulations (Fig. 3). In the same spirit of the Monte Carlo and the analytically solvable model, we assume that the loading rate is sufficiently low that at any given time, all modules experience the same instantaneous tensile force applied at their ends, *f*, and that each of them can unfold independently from the others. We also assume that the stiffness of the construct, defined as the derivative of *f* with respect to its length, is not dependent on the number of folded constructs, *n*. This is equivalent to considering $A = 0$ in the analytical model, and it is realistic for investigated cases (for counterexamples, see King et al. (21)). Under

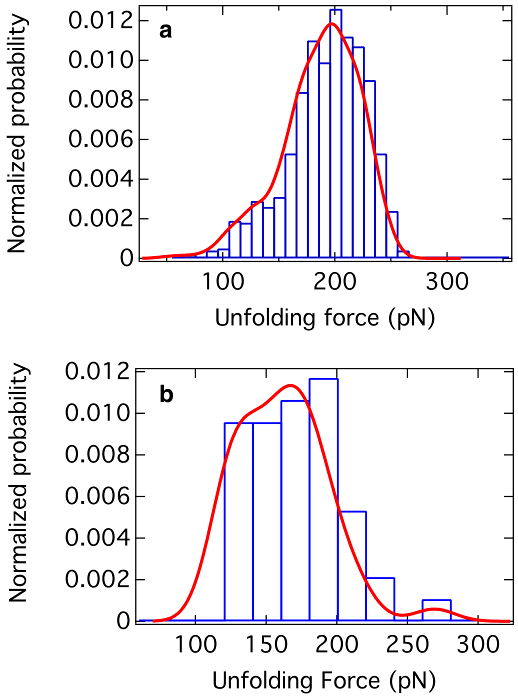

FIGURE 3 Normalized probability distribution of the unfolding forces for the $n = 1$ peaks (i.e., near the detachment point) obtained for (*a*) Monte Carlo simulation and (*b*) *GB*1pulling experiments. The continuous line in both cases represents the Gaussian KDE estimated from the raw data. The histograms are both normalized.

these assumptions, without resorting to any kinetic model or lengthy stochastic simulations, the average unfolding force associated with the *n*th peak, $\langle f \rangle_n$, is computed by drawing *n* random numbers distributed according to $p_{n=1}(f)$ and taking the smallest of them as the force at which one of the *n* modules first unfolds. The average value of the unfolding force, $\langle f \rangle_n$ (and its variance), is clearly obtained by repeating the batch force sampling process several times.

One may use the ordered list of *N* measurements to construct a cumulative probability distribution interpolated linearly between consecutive measured values. The cumulative distribution is next straightforwardly used (see Chapter 7.3 in Press et al. (33)) to sample, with the correct weight, the *n* force values. Describing the process in terms of the cumulative distribution has also the following important advantage. It is possible to exploit the simple relationship of Eq. 7 (which is based on the assumption of independence and hence valid regardless of the specific underlying kinetic process) to generate data for unfolding forces of the *n*th peak starting from the data obtained for a peak with a different order, say *m*th.

In fact, indicating by

$$Q_m(f) = \int_{-\infty}^{f} df' p_m(f')$$

the cumulative distribution for the unfolding forces of the *m*th peak, it can be determined that the corresponding cumulative distribution for the *n*th peak is

$$Q_n(f) = 1 - (1 - Q_m(f))^{n/m}. \tag{12}$$

It is important to stress that the above relationships are of high conceptual and practical interest for recovering the distribution of unfolding forces of one peak, say $n = 1$, starting from a peak of higher order, say $m = 2$. A detailed description of how this backward extrapolation can be practically implemented in a numerical scheme is provided in the Appendix, and the results are provided in the Supporting Material. The results discussed hereafter are produced with a more refined method where the probability $p_{n=1}(f)$ is obtained from fitting the histogram of the raw force measurements with a convolution of Gaussians using the kernel density estimation (KDE) (34) (Fig. 3). Data are sampled according to this distribution using either the cumulative distribution, or the rejection scheme (see Chapter 7.3 in Press et al. (33)).

RESULTS AND DISCUSSION

For all the three systems of interest (the *GB*1 experiment and the Monte Carlo and Langevin simulations), we analyzed the data of the force-versus-extension (or equivalently force-versus-time) curves. In all three cases, the data pertained to the stretching of constructs of $n_0 = 6$ modules, and therefore, the few curves that did not display a clear presence of six force peaks were discarded.

The peaks were indexed in an inverse order with respect to their order of appearance in the stretching experiment. Specifically, the peak of order $n = 6$ corresponds to the peak observed first (when six folded modules were present before the unfolding event), whereas peak order $n = 1$ corresponds to the unfolding event for which only one module was present before the unfolding event and occurring immediately before the construct detachment from the support. The peak force data for each value of *n* were next considered (see Benedetti et al. (20) for details on the automated peak division procedure) and used to compute the histograms reflecting the force distribution. The probability distribution is obtained with a convolution of Gaussians using the KDE method mentioned in the Methods section. The resulting normalized distribution of the forces, $p_n(f)$, at which a single module unfolds in the Monte Carlo scheme and *GB*1 experiments is shown in Fig. 3. The best-fit Gaussian convolutions were used to obtain a robust estimate of the average unfolding force and its standard deviation (SD) at each value of *n*. The results are provided in Tables 1–3 and Figs. 4–6.

The best-fit distribution for the last surviving peak, $n = 1$, was typically used as input for the back-calculation and analytically solvable methods to obtain predictions for the average unfolding forces at all values of *n*. For the case of highest practical interest, namely, the *GB*1 experiment, the distribution of unfolding forces of all other peaks, $n = 2, 3, 4, 5, 6$, was also used to predict the unfolding forces of other peaks (see Table S1 in the Supporting Material).

TABLE 1 Unfolding forces from Monte Carlo simulations

Comparison with $n = 1$ back-calculated values

n	6	5	4	3	2	1
Average, Monte Carlo data	142	146	150	157	166	187
Average, $n = 1$ BC	138	142	148	156	168	—
SD, Monte Carlo data	40	35	35	35	37	35
SD, $n = 1$ BC	29	31	32	31	33	—

All values are given in pN.

TABLE 2 Unfolding forces from Langevin simulations

Comparison with $n = 1$ back-calculated values

n	6	5	4	3	2	1
Average, Langevin simulation	70	73	74	78	84	92
Average, $n = 1$ BC	70	72	74	78	83	—
SD, Langevin simulation	16	14	13	13	14	15
SD, $n = 1$ BC	11	11	12	13	14	—

All values are given in pN.

Monte Carlo data

We start by discussing the application of the method to data generated using the Monte Carlo procedure. Of the three sets of data (from experiment and Langevin and Monte Carlo simulations), this set is the one that is expected to be most appropriately captured by back-calculation. The Monte Carlo scheme indeed builds on the identical kinetic status of all the modules, and during this process, only the total contour length changes, with very mild effect on the loading rate.

By using the $n = 1$ data, it is indeed seen in Table 1 that the mean values of the predicted and measured unfolding forces are in good agreement for all peaks $n = 2 \ldots 6$, with differences always <5 pN. The agreement is readily perceived in Fig. 4, where it is seen that the BC data up to $n = 4$ fall within the statistical uncertainty of the Monte Carlo data, and only the forces predicted at $n = 5$ and $n = 6$ present SDs of ~2.5 from the Monte Carlo data.

A more challenging quantity to compare is the second moment of the distribution, which is the variance or, equivalently, the SD. For the latter quantity, the agreement is still good. The deviation of the Monte Carlo and back-calculated values, $|\sigma_{BC} - \sigma_{data}|/(\sigma_{BC} + \sigma_{data})$, is typically within 10% and is worst for the last peak, $n = 6$, for which it is 16%.

The results of the analytical model present an accord with the Monte Carlo data that is comparable with their agreement with the BC. This is illustrated by the dashed line in Fig. 4, which reports the analytical predictions based on the Monte Carlo data for $n = 1$ (data for this case and other values of n are provided in the Supporting Material). The good accord is nontrivial in view of the fact that the simplified analytical treatment describes the folded protein domains as harmonic springs, whereas the Monte Carlo data were generated employing a WLC model for each domain. We carried out the Anderson-Darling statistical test described in the Methods section and established that the Monte Carlo data for $n = 1$ (and higher values, too) are compatible with an underlying Gumbel distribution. This reinforces the applicability of the simplified analytical scheme in the model Monte Carlo context.

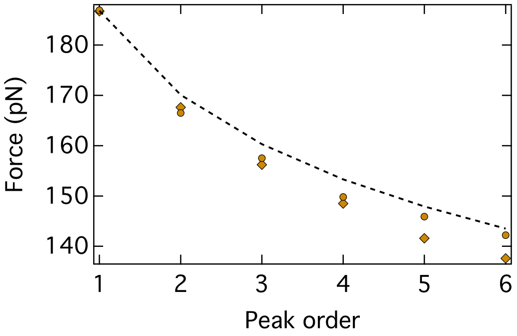

FIGURE 4 Average unfolding force versus peak order for the Monte Carlo (*circles*) and BC (*diamonds*) from the distribution of unfolding forces of the peak order $n = 1$ stemming from the Monte Carlo simulation and kinetic model (*dashed line*). The statistical error (mean ± SD) is the same size as the symbols, ~0.5 pN.

Langevin data

The same analysis was repeated for the data generated using the Langevin scheme, which contains several differences from the Monte Carlo scheme. Specifically, the Langevin scheme does not enforce either Evans's kinetics or the same precise behavior of all folded modules in the chain. In addition, it accounts for the presence of model linkers between the folded modules, and finally, values of Δx and k_{off} are appreciably different from those in the Monte Carlo case.

TABLE 3 Unfolding forces for GB1

Comparison with $n = 1$ and $n = 2$ back-calculated values

n	6	5	4	3	2	1
Average, experiment	124	128	129	137	146	162
Average, $n = 1$ BC	121	125	128	134	143	—
Average, $n = 2$ BC	123	127	131	137	—	160
SD – experiment	25	30	31	31	30	31
SD, $n = 1$ BC	17	18	19	21	25	—
SD, $n = 2$ BC	21	21	22	24	—	36

Accuracy in prediction of the SD is improved when data from $n = 2$ are used. All values are given in pN.

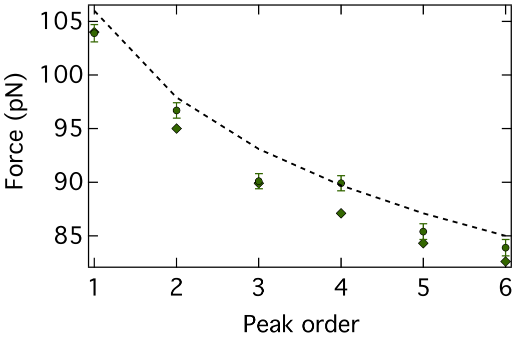

FIGURE 5 Average unfolding force versus peak order for the Langevin (*circles*) and BC (*diamonds*) from the distribution of unfolding forces of the peak order $n = 1$ stemming from the Langevin simulation and kinetic model (*dashed line*). The statistical errors (mean ± SD) are shown with error bars for the Langevin simulation, whereas for the BC data, the errors are the same size as the symbols (~0.5 pN).

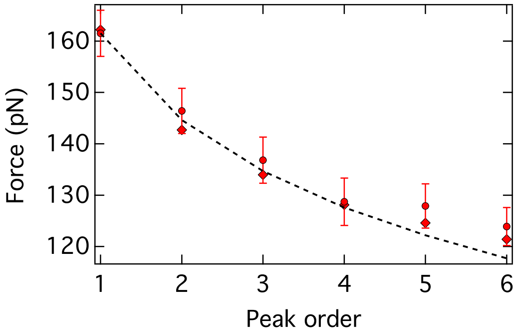

FIGURE 6 Average unfolding force versus peak order for the *GB*1 experimental data (*circles*) and BC (*diamonds*) from the distribution of unfolding forces of the peak order $n = 1$ stemming from the experiments and kinetic model (*dashed line*). The statistical errors (mean ± SD) are shown with error bars, whereas the BC data have errors of the same value as the size of the symbols (~0.5 pN).

As is visible from Table 2 and Fig. 5 also for the Langevin context, the performance of the back-calculation method is good and, with the exception of the point for the fourth peak (which compared to the trend of the other data points appears to be an outlier), the average predicted values of unfolding forces are all within about one SD of the Langevin data. As for the Monte Carlo data, the predicted SDs are also consistent with the measured ones, and the largest relative error, again found for the peak with the largest extrapolation, $n = 6$, is 14%.

As shown in Fig. 5, the performance of the analytical model based on the $n = 1$ data is not dissimilar from that of the back-calculation (the detailed results are again reported in the Supporting Material). Indeed, also in this context, the Anderson-Darling test indicates that distributions of the unfolding forces are compatible with a Gumbel distribution.

Experimental data on $(GB1)_8$

Finally, we turned to the experimental data, which clearly represent the greatest challenge. Because of the complex interplay of the several factors that impact on the stretching process, and because the pulling rate is not particularly low, it may not be expected a priori that the system unfolding response might be well captured by the back-calculation. In particular, it is not obvious a priori that the unfolding events of various peaks in the chain can be appropriately described as statistically independent events. In fact, correlations can arise in nearby protein moduli because of the limited sound velocity in the chain or because of contact interactions. Moreover, given the small number of experimental samples, 47 measurements for each force peak, it is not simple to obtain a reference histogram from the experiment or to pin a distribution, even when using the KDE interpolation scheme. Thus, any defect in the starting distribution is consequently amplified by the back-calculation method.

Despite these caveats, the predictive capability of the back-calculation method for the average unfolding forces was found to be very good also in this case. The level of agreement can be appreciated by examining Table 3 and Fig. 6. The increasing underestimation, as a function of n, of the sample SD (predicted from the $n = 1$ peak) is probably ascribable to the fewer-than-expected measurements at low forces. This is readily demonstrated by starting the back-calculation from the second peak, $n = 2$, which by covering lower values of unfolding forces can reproduce very well not only the mean unfolding forces at all other values of n, but also the corresponding SDs.

In light of this consideration, the very good consistency of the back-calculation data with the measured distribution is very remarkable, and testifies to the robust applicability of the method.

It is particularly instructive to discuss the performance of the analytical method as well. Neither the average unfolding forces nor their SDs are dissimilar from the experimental ones (see Fig. 6 and Supporting Material). However, unlike in the cases for the Monte Carlo and Langevin data, this agreement does not stand up to closer statistical scrutiny.

In fact, the Anderson-Darling statistical test indicates that the experimental data do not follow the Gumbel extremal statistics entailed by the analytical model at each value of n (see Eq. 8). In fact, the null hypothesis for the $n = 1$ peak is supported with a confidence level of $<1\%$. The same applies for the $n = 2$ peak as well (in spite of the fact that a more and more pronounced Gumbel-like character is expected as n increases).

The above observations demonstrate the utility of the back-calculation approach in the context of practical interest. Indeed, the phenomenology of systems such as multimodular constructs of *GB*1 can be too rich to be well accounted for by Evans's theory. In such contexts, a good control/prediction of the unfolding forces for varying numbers of surviving modules can be made only starting from the phenomenological distribution.

CONCLUSIONS

We present a systematic investigation of the statistical properties of the forces associated with the first, second, etc., unfolding events in a multimodular construct. We introduced a phenomenological scheme, termed back-calculation, which, using as sole input the distribution of the forces associated with a certain unfolding event (e.g., the first), predicts the force distribution of all other events. We stress that the method follows a bootstrap approach starting from the raw force-extension measurements. In particular, it does not rely on any model of mechanical response for protein unfolding kinetics.

At a general level, it is shown that the standard procedure of analyzing experimental stretching data by grouping together forces associated with all unfolding events, could be more profitably replaced by considering the events separately with equal order of appearance. To the best of our knowledge, the possibility of applying such a scheme to analyze experimental data has not been explored before. Second, a comparison of the experimental distributions of unfolding forces with that predicted by standard kinetic models reveals appreciable discrepancies, thus preventing their use as reliable descriptors of the mechanical unfolding process. This fact is consistent with previous independent investigations (27).

In addition, the approach has several implications for the design/analysis of stretching experiments of multimodular constructs. First, its simplicity makes the back-calculation particularly appealing as a simple and transparent scheme for the interpretation of experimental data. In this respect, an interesting applicative avenue is offered by heterogeneous multimodular constructs, for which the back-calculation can offer a term of reference apt for highlighting composition-dependent modulations of the mechanical response. Second, it offers a simple, parameter-free phenomenological approach for predicting the distributions of the various unfolding peaks using a negligible computational effort. In this respect, it presents major advantages compared to the more computationally intensive stochastic (Monte Carlo or Langevin) numerical approaches. Finally, it can be applied to the design of biomaterials starting from their molecular modular components (e.g., choosing an appropriate number of repeats), with unfolding forces falling in a desired range, or to precondition a pulling experiment (choice of pulling speed, stiffness of the AFM tip) so that the mechanical response is profiled with a desired resolution. A study of the latter aspects is underway.

The numerical implementations (C programming language) of the back-calculation techniques are available upon request from the authors.

APPENDIX

The procedure used to predict the force distribution for peak n given a set of experimental measurements for peak m is discussed here in detail. As a first step, an ordered table $F_i^{(m)}$, with $i = 1, \ldots, N$, is built that contains the N measured forces for peak m.

To resample new data from the same distribution, the procedure is as follows:

Extract a uniform random number $r \in [0, 1]$.
Find i such that $i < Nr < i + 1$.
The extracted force is computed as $(i - Nr + 1)F_i^{(m)} + (Nr - i)F_{i+1}^{(m)}$.

The last step is based on a liner interpolation of the cumulate of the distribution.

To extract data corresponding to the distribution of a different peak n, which can be larger or smaller than m, the procedure has to be modified as follows:

Extract a uniform random number $r \in [0, 1]$.
Compute $r' = 1 - (1 - r)^{m/n}$.
Find i such that $i < Nr' < i + 1$.
The extracted force is computed as $(i - Nr' + 1)F_i^{(m)} + (Nr' - i)F_{i+1}^{(m)}$.

SUPPORTING MATERIAL

One figure and one table are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(11)00938-6.

This work was supported by the Swiss National Science Foundation (grant No. 205320-131828) and by the Italian Ministry of Education (MIUR).

REFERENCES

1. Cao, Y., and H. B. Li. 2007. Polyprotein of GB1 is an ideal artificial elastomeric protein. *Nat. Mater.* 6:109–114.
2. Cao, Y., C. Lam, …, H. B. Li. 2006. Nonmechanical protein can have significant mechanical stability. *Angew. Chem. Int. Ed.* 45:642–645.
3. Carrion-Vazquez, M., A. F. Oberhauser, …, J. M. Fernandez. 1999. Mechanical and chemical unfolding of a single protein: a comparison. *Proc. Natl. Acad. Sci. USA.* 96:3694–3699.
4. Sandal, M., F. Valle, …, B. Samorì. 2008. Conformational equilibria in monomeric α-synuclein at the single-molecule level. *PLoS Biol.* 6:e6.
5. Brockwell, D. J., G. S. Beddard, …, S. E. Radford. 2005. Mechanically unfolding the small, topologically simple protein L. *Biophys. J.* 89: 506–519.
6. Marszalek, P. E., H. Lu, …, J. M. Fernandez. 1999. Mechanical unfolding intermediates in titin modules. *Nature.* 402:100–103.
7. Li, H., W. A. Linke, …, J. M. Fernandez. 2002. Reverse engineering of the giant muscle protein titin. *Nature.* 418:998–1002.
8. Carrion-Vazquez, M., H. B. Li, …, J. Fernandez. 2003. The mechanical stability of ubiquitin is linkage dependent. *Nat. Struct. Mol. Biol.* 10: 738–743.
9. Li, H., M. Carrion-Vazquez, …, J. Fernandez. 2000. Point mutations alter the mechanical stability of immunoglobulin modules. *Nat. Struct. Mol. Biol.* 7:1117–1120.
10. Lee, G., K. Abdi, …, P. E. Marszalek. 2006. Nanospring behaviour of ankyrin repeats. *Nature.* 440:246–249.
11. Oberhauser, A. F., P. K. Hansma, …, J. M. Fernandez. 2001. Stepwise unfolding of titin under force-clamp atomic force microscopy. *Proc. Natl. Acad. Sci. USA.* 98:468–472.
12. Sorce, B., S. Sabella, …, P. P. Pompa. 2009. Single-molecule mechanical unfolding of amyloidogenic β2-microglobulin: the force-spectroscopy approach. *ChemPhysChem.* 10:1471–1477.
13. Evans, E., and K. Ritchie. 1997. Dynamic strength of molecular adhesion bonds. *Biophys. J.* 72:1541–1555.
14. Garg, A. 1995. Escape-field distribution for escape from a metastable potential well subject to a steadily increasing bias field. *Phys. Rev. B.* 51:15592–15595.
15. Imparato, A., F. Sbrana, and M. Vassalli. 2008. Reconstructing the free energy landscape of a polyprotein by single-molecule experiments. *Europhys. Lett.* 82:58006.
16. Oberhauser, A. F., and M. Carrión-Vázquez. 2008. Mechanical biochemistry of proteins one molecule at a time. *J. Biol. Chem.* 283: 6617–6621.
17. Best, R., D. J. Brockwell, …, J. Clarke. 2003. Force mode AFM as a tool for protein folding studies. *Anal. Chim. Acta.* 479:87–105.
18. Brockwell, D. J., G. S. Beddard, …, S. E. Radford. 2002. The effect of core destabilization on the mechanical resistance of I27. *Biophys. J.* 83:458–472.
19. Aioanei, D., B. Samorì, and M. Brucale. 2009. Maximum likelihood estimation of protein kinetic parameters under weak assumptions from unfolding force spectroscopy experiments. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 80:61916.
20. Benedetti, F., S. K. Sekatskii, and G. Dietler. 2011. Single-molecule force spectroscopy of multimodular proteins: a new method to extract kinetic unfolding parameters. *J. Adv. Microsc. Res.* 6:1–6.

21. King, W. T., M. H. Su, and G. L. Yang. 2010. Monte Carlo simulation of mechanical unfolding of proteins based on a simple two-state model. *Int. J. Biol. Macromol.* 46:159–166.

22. Dietz, H., F. Berkemeier, …, M. Rief. 2006. Anisotropic deformation response of single protein molecules. *Proc. Natl. Acad. Sci. USA.* 103:12724–12728.

23. Hummer, G., and A. Szabo. 2008. Thermodynamics and kinetics of single-molecule force spectroscopy. *In* Theory and Evaluation of Single-Molecule Signals. E. Barkai, F. L. H. Brown, M. Orrit, and H. Yang, editors. World Scientific, Singapore. 139–175.

24. Li, H. 2007. Engineering proteins with tailored nanomechanical properties: a single molecule approach. *Org. Biomol. Chem.* 5:3399–3406.

25. Florin, E., M. Rief, …, H. Gaub. 1995. Sensing specific molecular interactions with the atomic force microscope. *Biosens. Bioelectron.* 10:895–901.

26. Sandal, M., F. Benedetti, …, B. Samorì. 2009. Hooke: an open software platform for force spectroscopy. *Bioinformatics.* 25:1428–1430.

27. Schlierf, M., and M. Rief. 2006. Single-molecule unfolding force distributions reveal a funnel-shaped energy landscape. *Biophys. J.* 90: L33–L35.

28. Rief, M., M. Gautel, …, H. E. Gaub. 1997. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science.* 276: 1109–1112.

29. Rief, M., J. M. Fernandez, and H. E. Gaub. 1998. Elastically coupled two-level systems as a model for biopolymer extensibility. *Phys. Rev. Lett.* 81:4764–4767.

30. Zinober, R. C., D. J. Brockwell, …, D. A. Smith. 2002. Mechanically unfolding proteins: the effect of unfolding history and the supramolecular scaffold. *Protein Sci.* 11:2759–2765.

31. Berkovich, R., S. Garcia-Manyes, …, J. M. Fernandez. 2010. Collapse dynamics of single proteins extended by force. *Biophys. J.* 98:2692–2701.

32. Gumbel, E. 2004. Statistics of Extremes. Dover, Mineola, NY.

33. Press, W., S. Teukolsky, …, B. Flannery. 2007. Numerical Recipes, 3rd ed.: The Art of Scientific Computing. Cambridge University Press, Cambridge, United Kingdom.

34. Silverman, B. W. 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC, Boca Raton, FL.