Force-Clamp Analysis Techniques Give Highest Rank to Stretched Exponential Unfolding Kinetics in Ubiquitin

Herbert Lannon,[†] Eric Vanden-Eijnden,[‡] and J. Brujic[†*]
[†]Department of Physics and Center for Soft Matter Research and [‡]Courant Institute of Mathematical Sciences, New York University, New York, New York

ABSTRACT Force-clamp spectroscopy reveals the unfolding and disulfide bond rupture times of single protein molecules as a function of the stretching force, point mutations, and solvent conditions. The statistics of these times reveal whether the protein domains are independent of one another, the mechanical hierarchy in the polyprotein chain, and the functional form of the probability distribution from which they originate. It is therefore important to use robust statistical tests to decipher the correct theoretical model underlying the process. Here, we develop multiple techniques to compare the well-established experimental data set on ubiquitin with existing theoretical models as a case study. We show that robustness against filtering, agreement with a maximum likelihood function that takes into account experimental artifacts, the Kuiper statistic test, and alignment with synthetic data all identify the Weibull or stretched exponential distribution as the best fitting model. Our results are inconsistent with recently proposed models of Gaussian disorder in the energy landscape or noise in the applied force as explanations for the observed nonexponential kinetics. Because the physical model in the fit affects the characteristic unfolding time, these results have important implications on our understanding of the biological function of proteins.

INTRODUCTION

Force-clamp spectroscopy using the atomic force microscope has proven to be a useful tool for following the unfolding trajectories of single polyprotein molecules (1–6). Previous studies have investigated the effect of the applied force (4,7–10), length of the polyprotein chain (11,12), and order statistics (3) on the unfolding kinetics of mechanically stable proteins. The simplest free energy landscape model for mechanical unfolding is a two-state reaction process over a single transition state barrier, which is tilted by the work done on the molecule (13). In such a reaction driven by simple diffusion, the probability distribution of the measured dwell times at a given force is exponential with a rate of decay that is determined by the barrier height. Moreover, the unfolding rate is exponentially dependent on the applied force. The majority of previous studies have interpreted their data using this two-state model to determine the height of the energy barrier and the distance to the transition state.

Apart from the two-state fitting of the unfolding kinetics of ubiquitin (7), more recent work has shown that a larger statistical pool of dwell times at a given force reveals important deviations from exponential kinetics and requires more sophisticated modeling. Surprisingly, these deviations have led to three alternative models with different physical interpretations for the unfolding of ubiquitin pulled under the same experimental conditions. The first physical interpretation considers unfolding via multiple pathways in a rough energy landscape, where the timescale of interconversion between the folded states is assumed to be slow compared to that of unfolding. This static disorder scenario predicts that the nonexponential dwell times at a force of 110 pN are consistent with exponentially distributed free energy barriers (8). By contrast, a more recent work assumes that the static disorder (14,15) has a Gaussian distribution of barriers and derives the corresponding function to fit the experimental dwell times over a range of constant forces (10). Alternatively, assuming that the Gaussian distribution comes from the noise in the applied force (11) leads to the same form of the nonexponential fitting function for the dwell times if the noise correlation time is longer than that of unfolding. In addition to these physical interpretations, in (12) a log normal distribution is proposed to be the best heuristic fit to the dwell times of both monomeric and polyubiquitin data.

A possible explanation for the apparent success of these four models in fitting the same data is that rigorous methods of analyzing and assessing force-clamp trajectories are lacking. For example, some studies average and normalize the measured end-to-end length trajectories as an estimate of the cumulative unfolding probability (7,9), whereas others export the individual dwell times and bin them into probability density distributions before fitting (3,10). Moreover, because the polyprotein chains vary in length and detach from the cantilever at random times, not all events are necessarily observed in the experiment. To account for the undetected events different filtering methods are applied to the data, each with their own associated uncertainties. In this work we quantitatively assess the errors in existing analysis protocols, develop new, to our knowledge, analysis methods that systematically take into account biases introduced by experimental artifacts, and evaluate the success of each model using not only graphical tests, but also rigorous

Submitted August 18, 2012, and accepted for publication October 19, 2012.

*Correspondence: jb2929@nyu.edu

Editor: Charles Wolgemuth.

0006-3495/12/11/2215/8 $2.00

http://dx.doi.org/10.1016/j.bpj.2012.10.022

statistical tests based on maximum likelihood estimation and Bayesian sampling. We show that tests of robustness against filtering the data provide an excellent indication of the validity of the underlying model and we illustrate the results in both real and synthetic data sets. To use the full experimental data set and avoid filtering, we additionally derive a likelihood function that calculates the probability of observing a sequence of dwell times followed by the measured detachment time of the molecule. This method allows us to rank the proposed models in terms of their consistency with observing the data set using standard statistical tests, such as those described in (16,17). Finally, we show the agreement between filtering techniques and the use of the likelihood function and propose a self-consistent recipe for data assessment in future experiments.

The importance of distinguishing between fitting functions is to deduce the correct physical picture for protein unfolding, which sets the mechanical response timescales in biology. Indeed, it is striking that the mean unfolding times for the four proposed distributions for ubiquitin span over two orders of magnitude at a given constant force, thus emphasizing the importance of determining the correct model.

MATERIALS AND METHODS

Force-clamp spectroscopy measurements are taken using the same atomic force microscope and experimental method described in (1,2,7–12). The ubiquitin polyprotein construct consists of 12 identical monomers and is synthesized according to the same procedure as that described in (11). In response to a constant stretching force, each of the protein domains in a polyprotein chain unfolds stochastically, leading to a stepwise elongation of the end-to-end length over time, as shown in the example in Fig. 1 *A*. Time zero is marked at the beginning of the first plateau in the end-to-end length after the constant stretching force of 110 pN is applied. The resulting staircase of unfolding events yields a set of dwell times t_1, t_2, …, which mark the rupture of the native state of each domain to the fully extended unfolded state. Only staircases with a minimum of three repeating steps are included in the analysis as the signature of the single polyprotein molecule. Plotting over 2000 unfolding times in the order in which they are collected leads to the scatter graph in Fig. 1 *B*. The logarithmic scale emphasizes the span over three orders of magnitude of the unfolding times, whereas the homogeneity of the data from experiments performed with distinct cantilevers and on different days gives validity to the force calibration and the stability of the protein, respectively.

RESULTS AND DISCUSSION

Unbiasing the unfolding data from experimental artifacts

To determine the probability $F(t)$ of observing an unfolding event before time t, a common way to analyze time series data is to plot the cumulative distribution function (CDF) of the dwell times. Experimentally this CDF is often constructed by averaging and normalizing the raw staircases, but this method gives an approximation of the CDF that is not monotonically increasing due to the presence of thermal noise and occasional drift in the experiment. Instead, the correct way to construct the CDF is to directly export the dwell times, sort and rank them from smallest to largest, and then plot the normalized rank against the dwell time as the empirical CDF. This procedure avoids loss of information by binning, given that this empirical CDF has a value at each measured dwell time.

However, in the case of force-clamp trajectories the empirical CDF of all the observed dwell times does not coincide with the unfolding probability $F(t)$ because of experimental artifacts. The experimental window (fixed by the time resolution at short times, $t_{\min}$, and the total duration of the experiment, $t_{\max}$) may not encompass the whole range of the unfolding probability $F(t)$. The empirical CDF is given by

$$\widehat{P}(t) = \frac{\#\{\text{dwell times}<t\}}{N}, \tag{1}$$

where N is the total number of dwell times in the data set and $\#\{\text{dwell times}<t\}$ denotes the number of such times that are less than t, and must therefore be fit with a $P(t)$, conditional on the time range of the experiment (18). Although $F(t)$ is zero at time zero and reaches one at infinity, the conditional $P(t)$ is fixed to zero at $t_{\min}$ in our experiments, reaches one at $t_{\max}$, and is defined as

$$P(t) = \begin{cases} 0 & \text{if } t<t_{\min} \\ \dfrac{F(t)-F(t_{\min})}{F(t_{\max})-F(t_{\min})} & \text{if } t_{\min}\le t\le t_{\max} \\ 1 & \text{if } t>t_{\max} \end{cases}. \tag{2}$$

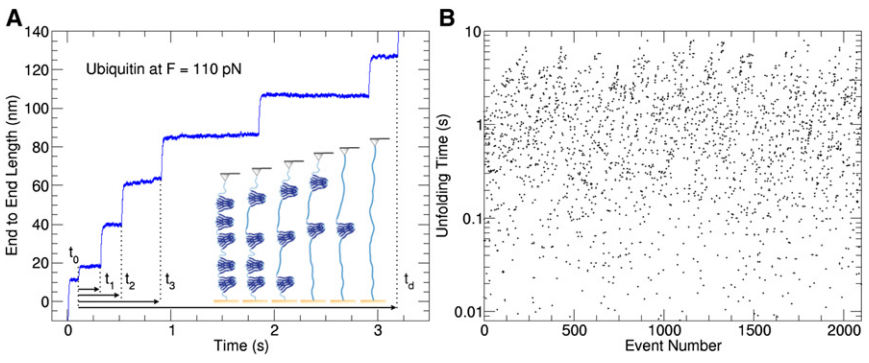

FIGURE 1 (*A*) A typical force-clamp unfolding trajectory of a single ubiquitin polyprotein pulled with a constant stretching force of 110 pN. The beginning of the plateau that precedes the staircase of unfolding events marks time zero t_0 as the moment when the molecule is held taught under the applied force. The dwell times are then measured as the time interval between t_0 and each of the unfolding steps. Finally, the molecule detaches at t_d. The stepwise unfolding is illustrated in the schematic diagram. (*B*) Unfolding dwell times from the staircases are plotted on a semilog scale in the order that they are collected and show a broad and homogeneous distribution of times.

Note that this conditional fitting of the data fixes the values of $F(t)$ at t_{min} and t_{max} without the need of introducing extra parameters. The functional form of $F(t)$ chosen for the fitting procedure self-consistently determines the range captured by the data, as shown in Fig. 2 *A*. If the experiment lasts long enough that the value of $F(t)$ approaches one at t_{max} then the conditioning has little effect on the parameters. However, even cases where $F(t)$ reaches 0.85 at t_{max} can alter the rate of an exponential function by 27% and change the shape of the distribution unless this conditioning is taken into account (see Fig. S1 in the Supporting Material).

Another artifact of force-clamp trajectories is that the molecules detach from the cantilever at random times t_d, which implies that some events are not observed in the experiment. If the total number of domains N in the polyprotein chain were known a priori (3), one could unbias the distribution of dwell times using order statistics, assuming that the unfolding events are independent of one another. This assumption for linear polyproteins has been proven correct for ubiquitin (11), NuG2 protein (3) as well as in numerical simulations (19). However, in our experiments the cantilever picks up polyproteins at random points on the surface such that any N (up to the full length N_{max}) can be exposed to a stretching force in a given experiment. This renders the unbiasing procedure difficult to resolve because different distributions $p(N)$ bias the empirical $\widehat{P}(t)$, which is illustrated on synthetic examples in Fig. S2. It is therefore necessary to filter the data, such that all events come from trajectories that correspond to the same time window in the experiment, to construct the $P(t)$ that corresponds to the underlying $F(t)$. The correct way to do so is to choose an experimental time window (e.g., from t_{min} to a cutting time t_c) and only consider those dwell times that i), occur within that range and ii), come from trajectories that lasted over the entire range, such that $t_d \geq t_c$, as shown in Fig. S3. Note that filtering the data by the detachment time alone by keeping all dwell times less than t_d leads to empirical CDFs that give inaccurate values of the fitting parameters, as shown in Fig. S4.

Graphical tests of the unfolding probability *F(t)*

Using the described methods for filtering and fitting of the experimental CDF $\widehat{P}(t)$, we assess the success of different models in explaining the ubiquitin data. The experimental time window is chosen to be between the time resolution of the experiment $t_{min} = 5$ ms and the cutting time $t_c = 5$ s, which ensures three decades over which to test the goodness of fit of the data. The same empirical $\widehat{P}(t)$ is then fit with Eq. 2 for the four functional forms of $F(t)$ proposed in the literature and listed in Table 1. The fitting can be done by least squares or maximum likelihood methods, which result in parameters that agree to within two decimal places. Because the fitting procedure self-consistently fixes $F(t_{max})$ and $F(t_{min})$ for each function, the resulting empirical $\widehat{F}(t) = (F(t_{max}) - F(t_{min}))\widehat{P}(t) + F(t_{min})$, obtained by solving Eq. 2 for $F(t)$, differ in their range, as shown in Fig. 2 *A*. For instance, the experimental window captures only 60% of the events in the case of the log normal distribution, whereas it covers almost all the events in the case of the exponential function. Moreover, the curves clearly show that the exponential fitting is inaccurate, although the other three models are all in good agreement with the data on the linear scale and exhibit comparable χ^2 values. To zoom into the two decades of fast unfolding times, the inset shows the data plotted as the conditional $P(t)$ on a log-log scale that emphasizes deviations from the fits. Here, it can be seen that the Weibull distribution performs better than all others on timescales below 0.1 s. Note that the Weibull distribution plotted as the $F(t)$ would be a straight line on the scales of the inset, but the $P(t)$ distribution is conditional on the time window of the experiment and thus exhibits curvature. Even though the Weibull distribution fits this data set most accurately, the statistical error in the experiment precludes the determination of the correct model by this graphical test alone.

Indeed, many functional forms (particularly those with several parameters) can be successful in fitting a particular time window chosen for the analysis, but it is a greater challenge to assess how robust the fitting function and its parameters are against filtering the same data over different time

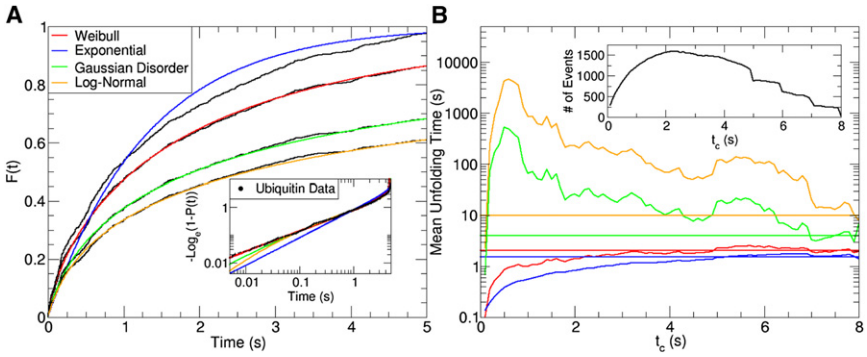

FIGURE 2 (*A*) The unfolding probability $F(t)$ for four models proposed in the literature is used to fit the same empirical CDF of dwell times. The normalization of each $F(t)$ leads to different timescales on which the data unfold. The inset shows the corresponding conditional $P(t)$ on a log-log plot to emphasize the goodness of fit at short times. (*B*) Changing the time window from 5 s in (*A*) to t_c shows the variability in the characteristic unfolding time between the different models. They span more than two orders of magnitude and only the Weibull and the exponential distribution settle to a given value. The inset shows how the number of data points changes as the time window is expanded.

TABLE 1 Parameter values from previous and our study for different distributions applied to a data set of ubiquitin pulled at 110 pN of constant force

Distribution $F(t)$	Previous Studies	MLE Parameters
Exponential $1 - e^{-at}$	a ~ 0.67 s^{-1} (9)	a = 0.66 ± 0.02 s^{-1}
Log-Normal $\frac{1}{2}\mathrm{erfc}\left[-\frac{\ln(t/t_0)}{\sigma\sqrt{2}}\right]$	σ = 3.0 (12) t_0 = 0.005 s	σ = 2.04 ± 0.05 t_0 = 1.26 ± 1.08 s
Gaussian Disorder (GD) $1 - \int_{\mathbb{R}} e^{-k_F t e^{-\beta r}} \frac{e^{-\frac{r^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dr$	k_F = 0.73 ± 0.03 s^{-1} (10) σ = 3.47 ± 1.16 pNnm	k_F = 0.57 ± 0.05 s^{-1} σ = 5.32 ± 0.72 pNnm
Force noise = GD With σ = σ_FΔx	Δx = 0.23 nm σ_F = 15.09 ± 5.04 pN	Δx = 0.23 nm σ_F = 23.13 ± 3.13 pN
Weibull $1 - e^{-(at)^b}$	a ~ 0.9 s^{-1} (8) b = γ − 1 = 0.8	a = 0.59 ± 0.04 s^{-1} b = 0.73 ± 0.02

windows. The shorter the time window, the more data points are needed to obtain the same statistical accuracy in the fitting, as shown in the synthetic example in Fig. S5. Nevertheless, there exists a range of cutting times t_c over which the fitted parameters should converge to the same values given a large enough pool of data. As a test of robustness of the parameters, we calculate the first moment of $F(t)$ (i.e., the mean unfolding time) fitted at different values of t_c, shown in Fig. 2 *B*. It can be seen that filtering the data at any time above 2.5 s has little effect on the mean unfolding time for the Weibull distribution, whereas the Gaussian disorder and log normal distributions vary greatly with t_c. The mean unfolding time is plotted on a logarithmic scale to capture the two orders of magnitude span that is predicted by the different experimental time windows of the same pool of data. This result shows that fitting with different physical models leads to dramatic consequences on biological function, because the characteristic protein unfolding time varies from 1 s to 3 min.

Although it can be argued that the statistical pool of filtered data shown in the inset is insufficient to fit $F(t)$ at short cutting times t_c, the lack of convergence over any significant filtering range for the Gaussian disorder and log normal functions questions their validity in describing the data. On the other hand, the exponential distribution does exhibit a range of stability after $t_c \approx 3$ s, but its poor performance in fitting the data invalidates its use for a different reason. This analysis shows that a successful model must not only fit the data with fidelity over a range of t_c, but also predict parameters that are stable over that range.

Instead of using the least squares method to assess the goodness of fit and extrapolate variance in the parameters by bootstrapping, other approaches work equally well. One such method is maximum likelihood estimation (MLE) (20), which computes the most likely parameters of a distribution using a set of variables—in our case the dwell times. The variance in the parameters is then obtained by Bayesian sampling of the data set (21), as shown in Fig. S6. The mean values of parameters a and b in the Weibull distribution and k_f and σ in the Gaussian disorder distribution are shown as a function of the experimental time window t_c in Fig. 3, *A* and *B*, respectively. The inset shows that the root mean standard deviation in the fitting parameters decreases as a function of t_c, which is consistent with the concomitant increase in the number of data points and the wider time window of the fit. Although the fluctuations observed in the Weibull parameters converge to stable values above $t_c \approx 3$ s, those of the Gaussian disorder model do not settle to any given values before $t_c \approx 7$ s, which is also reflected in the broad fluctuations of the mean unfolding time shown in Fig. 2 *B*. Note that filtering at long t_c

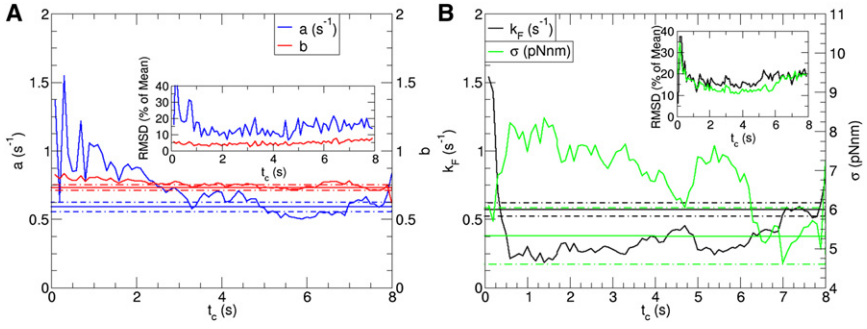

FIGURE 3 Estimate of the fitting parameters in the Weibull in (*A*) and the Gaussian disorder distribution in (*B*) as a function of the experimental time window. Bayesian sampling shows that the fluctuations around the mean of the parameters diminish as the time window increases. The constant solid lines are the parameter values obtained from the maximum likelihood function in Eq. 3 and the dashed lines are their standard deviation.

uses as little as 10% of the data collected, as shown in the inset in Fig. 2. Disregarding the majority of the data set is never desirable to an experimentalist. Nevertheless, having access to only those trajectories that unfold the entire polyprotein chain is an experimental way to bypass the need of data filtering.

Maximum likelihood function includes all collected data

An alternative MLE function to fitting the CDF of the dwell times as a function of t_c is one that takes into account experimental features of force-clamp trajectories and thus uses the whole data set to estimate parameters in the unfolding model. In a typical pulling experiment, the cantilever picks up a polyprotein chain of N domains with a probability $p(N)$. These domains subsequently unfold at dwell times t_1, t_2, …, t_k, where k corresponds to the last observed step in the staircase with a minimum $k_* = 3$ required as the signature of the single molecule. Finally, the molecule detaches either from the tip or the surface at time t_d. Assuming that the dwell times are independent of one another (3,11) and identically distributed (19,22,23) we calculate the probability of observing k unfolding events, multiplied by the probability of $N - k$ domains remaining folded up to the detachment time t_d, for every polyprotein chain:

$$\frac{1}{G_*}\sum_{N=k}^{N_*} p(N)\frac{N!}{(N-k)!k!}f(t_1)\cdots f(t_k)[1-F(t_d)]^{N-k}, \quad (3)$$

where $f(t) = dF/dt$ is the probability density associated with $F(t)$, $N_* = 12$ is the number of domains in the expressed protein construct, and G_* accounts for the probability of not including staircases with less than $k_* = 3$ steps

$$G_* = \sum_{N=k_*}^{N_*} p(N)\sum_{l=k_*}^{N}\frac{N!}{(N-l)!l!}[F(t_d)]^l[1-F(t_d)]^{N-l}. \quad (4)$$

Taking the product of the likelihoods for each polyprotein chain in Eq. 3 gives the overall likelihood function. The binomial prefactor takes into account the fact that the dwell times increase with a decreasing number of folded domains in the polyprotein chain. The parameters in the unfolding probability $F(t)$ as well as those defining $p(N)$ (assumed to be a power law with a decay coefficient γ in this case) are obtained by maximizing this likelihood function, whereas the uncertainties are estimated using Bayesian sampling.

The maximum value of the likelihood function from the ubiquitin data set ranks the four proposed unfolding distributions in the following order from highest to lowest likelihoods: Weibull, Gaussian disorder, log normal, and exponential distribution. Given that the actual values of the likelihoods depend on the size of the data set and $F(t)$, this rank test only estimates which distribution is more consistent with the data, however it cannot assess the accuracy of the fits themselves. Nevertheless, the fact that the Weibull parameters from the likelihood function, also shown in Fig. 3 *A*, are in good agreement with the parameter convergence of the fits of the $F(t)$ above $t_c \approx 3$ s gives further support to this model. Conversely, the lack of such an agreement in the runner-up Gaussian disorder model suggests that this is not the correct functional form for the unfolding probability.

A statistical test that quantitatively assesses whether a set of observables originates from a given distribution is the Kolmogorov-Smirnov approach with a modification by Kuiper (16). We therefore compare the empirical CDFs at different t_c from the data set with those generated from the four functional forms of $f(t)$ using the parameters estimated by the above likelihood function. Denoting as before by $P(t)$ the postulated distribution and by $\widehat{P}(t)$ the experimental distribution, the Kuiper statistic is defined as

$$U = \sqrt{N}\max_{j=1,\ldots,N}\left(P(t_j)-\widehat{P}(t_j)\right)-\sqrt{N}\min_{j=1,\ldots,N}\left(P(t_j)-\widehat{P}(t_j)\right), \quad (5)$$

where the maximum and the minimum are taken over all the N dwell times $t_1, \ldots, t_N$ in the data set. $U = 1$ signifies a perfect match. The results in Fig. 4 show that the Weibull distribution is closest to 1 over almost the entire range of t_c. Although the Gaussian disorder model is slightly closer to 1 between $7.2<t_c<8$ s, this narrow range is based on <10% of the collected data and is likely to be fortuitous.

Comparison with synthetic and other data sets

To further test the consistency of the ubiquitin data with the Weibull and Gaussian disorder models, we generate two

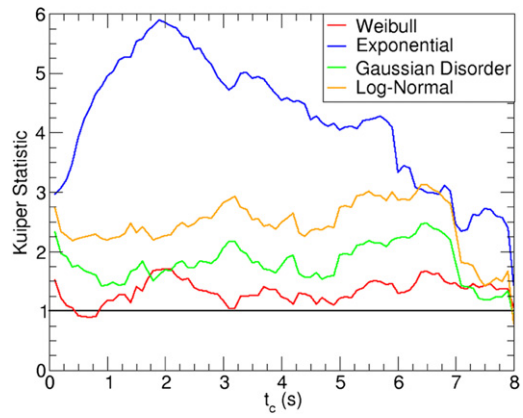

FIGURE 4 A Kuiper statistic of 1 signifies a perfect match between the experimental data and the proposed distribution. Deviations from the line at 1 quantify the disagreement between the maximum likelihood function estimate for the four models and the experimental data set as a function of the experimental time window t_c.

synthetic data sets using the parameters obtained from the maximum likelihood function in Eq. 3 that mimic the size of the experimental data. We then filter the synthetic and experimental data sets by t_c and compare the values of their fitting parameters using the Weibull distribution in Fig. 5, *A* and *B*, and the Gaussian disorder model in Fig. 5, *C* and *D*. In all cases, we find that the experimental and the synthetic Weibull data are in good agreement with each other above $t_c = 3$ s, whereas the synthetic Gaussian disorder data exhibits significant deviations. The two data sets are similar in that they not only exhibit comparable fluctuations in the fitting parameters of the Weibull arising from statistical errors, but they also follow similar trends in their discrepancy from the fitting parameters of the Gaussian disorder model. All these results are consistent with the hypothesis that the unfolding of ubiquitin data at 110 pN is most likely to originate from a Weibull distribution.

The fact that the Gaussian disorder model does not agree with the data contradicts theories of static disorder (10) and force noise (24), since they imply the same fitting function. Although the former places the Gaussian noise in the barriers to unfolding, the latter does so in the constant force applied by the cantilever. Given that $\sigma = \sigma_F \Delta x$, where σ_F is the noise in the applied force and $\Delta x = 0.23$ nm is the estimated distance to the transition state, σ in the barriers obtained from the MLE function translates to $\sigma_F = 21\%$ or $\sigma_F = 32\%$ error in the force calibration, depending on whether one takes the value of (σ) in Table 1 that is most likely or most stable against filtering, respectively. If this functional form had fit the data well, the estimated error in the force calibration would be much higher than the measured error of $\approx 5\%$ (25) and would thus give validity to the scenario of static disorder in the ubiquitin free energy landscape rather than that of the force noise.

It is worth noting that there are several reasons for which our results are not in agreement with those published in the literature and why the results in the literature disagree between each other. First, a common mistake in fitting force-clamp data is to introduce a normalization constant as an extra fitting parameter. Instead, care must be taken to fit the conditional unfolding probability distribution over the experimental window with $P(t)$ and obtain $F(t)$ using Eq. 2. Second, binning the distribution of unfolding times should be avoided because it effectively introduces an extra parameter into the fitting and loses resolution at short unfolding times. Third, filtering the data by accepting those trajectories that last a set minimum detachment time t_d and including events that occur after that t_d into the $P(t)$ biases the resulting distribution at long times, which in turn skews the fitting parameters. Instead, one must only include those dwell times that occur within exactly the same time window. Finally, plotting and fitting data on log-log scales can be useful if the data have been shown to fit well with a stretched exponential function to use a straight line fit. Otherwise, the compression of the data may obscure deviations from view and requires further assessment of the fits using MLE and Bayesian sampling.

CONCLUSIONS

Numerous force-clamp analysis methods, such as the fitting of filtered cumulative dwell time distributions, convergence

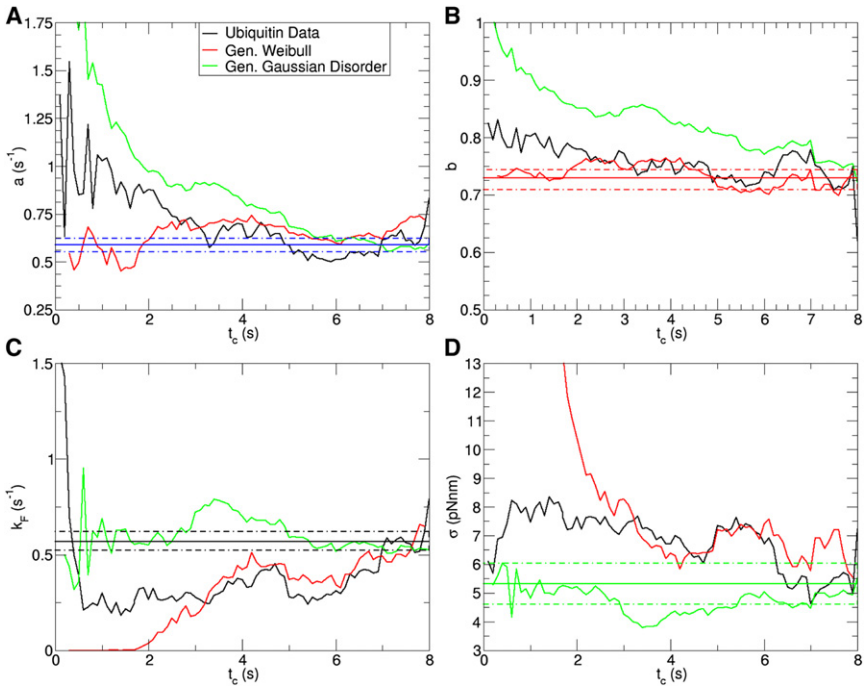

FIGURE 5 Comparison between synthetic data sets generated using the parameters in the maximum likelihood function for the Weibull and Gaussian disorder distribution and the experimental data set of ubiquitin. The constant solid lines are the parameter values obtained from the maximum likelihood function in Eq. 3 and the dashed lines are their standard deviation. Fitting the three data sets using the Weibull distribution gives the fluctuations in parameter a in (*A*) and b in (*B*) and using the Gaussian disorder distribution gives k_F in (*C*) and σ in (*D*). The ubiquitin data and the synthetic Weibull distribution behave similarly above $t_c = 3$ s in all cases, but the synthetic Gaussian distribution is significantly different.

of fitting parameters with an expanding time window of the experiment, the prediction of the maximum likelihood function for the whole data set, the Kuiper test, as well as the comparison with synthetically generated data sets, ubiquitously show that the data are most likely to arise from an underlying Weibull distribution, otherwise known as the stretched exponential distribution. This type of kinetics has been observed in the case of DNA relaxation (26), thermally induced protein folding (27,28), protein binding (29), and conformational dynamics in solution (30). Microscopically, the stretched exponential has been attributed to multiple pathways in the protein landscape (31) or memory effects (32). Our results show that such complexities may also play a role in the protein's response to a pulling force at the single molecule level.

One possible interpretation is that the unfolding events can occur via many (random) pathways, each with a different rate α, and the distribution of unfolding times is obtained via superposition of the exponential decays in each of these pathways. For example, the stretched exponential corresponds to rates that are distributed according to the Lévy distribution, because its probability is defined implicitly via

$$\int_0^\infty \left(1 - e^{-\alpha t}\right)\rho(\alpha)d\alpha \;=\; 1 - e^{-(at)^b} \equiv F(t), \qquad (6)$$

where $\rho(\alpha)$ cannot be written in closed analytical form but it exhibits a power law $\propto \alpha^{-\gamma}$ at large α. Therefore, the stretched exponential fitting function is in agreement with the theoretical model used in (8) to fit the ubiquitin unfolding kinetics. It remains an open question how the stretching exponent varies with the constant force in ubiquitin and whether it also captures the unfolding kinetics in other mechanically stable proteins. By contrast, the Gaussian distribution of energies or force noise proposed in (10) corresponds to a log-normal distribution in the rates in Eq. 6 via the Arrhenius assumption.

These methods invite previous studies to verify the accuracy of their results and provide a statistical toolbox for the analysis of future force-clamp studies. Moreover, it is possible to build on these techniques to take into account the particularities of a given experiment. For example, it is possible to introduce correlations between the domains within the likelihood function or assume a known $p(N)$ in the case of prepulled proteins. More generally, this type of analysis can be applied to other types of force-clamp measurements, such as the disulfide bond rupture kinetics (33) or the disassociation of quaternary interactions between individual domains (34).

SUPPORTING MATERIAL

Six figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(12)01127-7.

We thank Jin Montclare for the expression of the ubiquitin polyprotein and Maxime Clusel for useful discussions. J.B. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This work was supported partially by the MRSEC Program of the National Science Foundation under award No. DMR-0820341 and the National Science Foundation Career Award 0955621.

REFERENCES

1. 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.
2. Fernandez, J. M., and H. Li. 2004. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science.* 303:1674–1678.
3. Cao, Y., R. Kuske, and H. Li. 2008. Direct observation of markovian behavior of the mechanical unfolding of individual proteins. *Biophys. J.* 95:782–788.
4. Liu, R., S. Garcia-Manyes, …, J. M. Fernández. 2009. Mechanical characterization of protein L in the low-force regime by electromagnetic tweezers/evanescent nanometry. *Biophys. J.* 96:3810–3821.
5. Perez-Jimenez, R., S. Garcia-Manyes, …, J. M. Fernandez. 2006. Mechanical unfolding pathways of the enhanced yellow fluorescent protein revealed by single molecule force spectroscopy. *J. Biol. Chem.* 281:40010–40014.
6. Bullard, B., T. Garcia, …, A. F. Oberhauser. 2006. The molecular elasticity of the insect flight muscle proteins projectin and kettin. *Proc. Natl. Acad. Sci. USA*. 103:4451–4456.
7. Schlierf, M., H. Li, and J. M. Fernandez. 2004. The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. *Proc. Natl. Acad. Sci. USA*. 101:7299–7304.
8. Brujić, J., R. Hermans, …, J. Fernandez. 2006. Single-molecule force spectroscopy reveals signatures of glassy dynamics in the energy landscape of ubiquitin. *Nat. Phys.* 2:282–286.
9. Garcia-Manyes, S., L. Dougan, …, J. M. Fernández. 2009. Direct observation of an ensemble of stable collapsed states in the mechanical folding of ubiquitin. *Proc. Natl. Acad. Sci. USA*. 106:10534–10539.
10. Kuo, T. L., S. Garcia-Manyes, …, J. M. Fernández. 2010. Probing static disorder in Arrhenius kinetics by single-molecule force spectroscopy. *Proc. Natl. Acad. Sci. USA*. 107:11336–11340.
11. Brujić, J., R. I. Hermans, …, J. M. Fernandez. 2007. Dwell-time distribution analysis of polyprotein unfolding using force-clamp spectroscopy. *Biophys. J.* 92:2896–2903.
12. Garcia-Manyes, S., J. Brujić, …, J. M. Fernández. 2007. Force-clamp spectroscopy of single-protein monomers reveals the individual unfolding and folding pathways of I27 and ubiquitin. *Biophys. J.* 93: 2436–2446.
13. Bell, G. I. 1978. Models for the specific adhesion of cells to cells. *Science.* 200:618–627.
14. Zwanzig, R. 1990. Rate processes with dynamical disorder. *Acc. Chem. Res.* 23:148–152.
15. Zwanzig, R. 1992. Dynamical disorder: passage through a fluctuating bottleneck. *J. Chem. Phys.* 97:3587–3589.
16. Kuiper, N. 1960. Tests concerning random points on a circle. *Nederl. Akad. Wetensch. Proc. Ser. A.* 63:38–47.
17. Tygert, M. 2010. Statistical tests for whether a given set of independent, identically distributed draws comes from a specified probability density. *Proc. Natl. Acad. Sci. USA*. 107:16471–16476.
18. Koster, D. A., C. H. Wiggins, and N. H. Dekker. 2006. Multiple events on single molecules: unbiased estimation in single-molecule biophysics. *Proc. Natl. Acad. Sci. USA*. 103:1750–1755.
19. Bura, E., D. K. Klimov, and V. Barsegov. 2007. Analyzing forced unfolding of protein tandems by ordered variates, 1: Independent unfolding times. *Biophys. J.* 93:1100–1115.

20. Edwards, A. 1992. Likelihood, Expanded ed. The John Hopkins University Press, Baltimore, MD.

21. Howson, C., and P. Urbach. 2005. Scientific Reasoning: The Baysian Approach, 3rd ed. Open Court, Chicago, IL.

22. Bura, E., D. K. Klimov, and V. Barsegov. 2008. Analyzing forced unfolding of protein tandems by ordered variates, 2: dependent unfolding times. *Biophys. J.* 94:2516–2528.

23. Cao, Y., and H. Li. 2011. Single-molecule force-clamp spectroscopy: dwell time analysis and practical considerations. *Langmuir.* 27: 1440–1447.

24. Clusel, M., and E. I. Corwin. 2011. Unfolding proteins with an atomic force microscope: force-fluctuation-induced nonexponential kinetics. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 84:041920.

25. Ohler, B. 2007. Cantilever spring constant calibration using laser Doppler vibrometry. *Rev. Sci. Instrum.* 78:063701.

26. Biancaniello, P. L., A. J. Kim, and J. C. Crocker. 2008. Long-time stretched exponential kinetics in single DNA duplex dissociation. *Biophys. J.* 94:891–896.

27. Leeson, D. T., F. Gai, …, R. B. Dyer. 2000. Protein folding and unfolding on a complex energy landscape. *Proc. Natl. Acad. Sci. USA.* 97:2527–2532.

28. Chung, H. S., M. Khalil, …, A. Tokmakoff. 2005. Conformational changes during the nanosecond-to-millisecond unfolding of ubiquitin. *Proc. Natl. Acad. Sci. USA.* 102:612–617.

29. Hagen, S. J., J. Hofrichter, and W. A. Eaton. 1995. Protein reaction kinetics in a room-temperature glass. *Science.* 269:959–962.

30. Yang, H., G. Luo, …, X. S. Xie. 2003. Protein conformational dynamics probed by single-molecule electron transfer. *Science.* 302:262–266.

31. Hagen, S., and W. Eaton. 1996. Nonexponential structural relaxations in proteins. *J. Chem. Phys.* 104:3395–3398.

32. Kou, S. C., and X. S. Xie. 2004. Generalized Langevin equation with fractional Gaussian noise: subdiffusion within a single protein molecule. *Phys. Rev. Lett.* 93:180603–180607.

33. Wiita, A. P., S. R. Ainavarapu, …, J. M. Fernandez. 2006. Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques. *Proc. Natl. Acad. Sci. USA.* 103:7222–7227.

34. Xu, T., H. Lannon, S. Wolf, F. Nakamura, and J. Brujić. 2012. Filamin A (16–23) reveals a hierarchy of unfolding forces arising from domain-domain interactions in the polyprotein chain. *ArXiv e-prints*.