Analyzing Forced Unfolding of Protein Tandems by Ordered Variates, 1: Independent Unfolding Times

E. Bura,* D. K. Klimov,† and V. Barsegov‡

*Department of Statistics, George Washington University, Washington, DC; †Department of Bioinformatics & Computational Biology, George Mason University, Manassas, Virginia; and ‡Department of Chemistry, University of Massachusetts, Lowell, Massachusetts

ABSTRACT Most of the mechanically active proteins are organized into tandems of identical repeats, $(D)_N$, or heterogeneous tandems, D_1–D_2–. . .–D_N. In current atomic force microscopy experiments, conformational transitions of protein tandems can be accessed by employing constant stretching force f (force-clamp) and by analyzing the recorded unfolding times of individual domains. Analysis of unfolding data for homogeneous tandems relies on the assumption that unfolding times are independent and identically distributed, and involves inference of the (parent) probability density of unfolding times from the histogram of the combined unfolding times. This procedure cannot be used to describe tandems characterized by interdomain interactions, or heteregoneous tandems. In this article, we introduce an alternative approach that is based on recognizing that the observed data are ordered, i.e., first, second, third, etc., unfolding times. The approach is exemplified through the analysis of unfolding times for a computer model of the homogeneous and heterogeneous tandems, subjected to constant force. We show that, in the experimentally accessible range of stretching forces, the independent and identically distributed assumption may not hold. Specifically, the uncorrelated unfolding transitions of individual domains at lower force may become correlated (dependent) at elevated force levels. The proposed formalism can be used in atomic force microscopy experiments to infer the unfolding time distributions of individual domains from experimental histograms of ordered unfolding times, and it can be extended to analyzing protein tandems that exhibit interdomain interactions.

INTRODUCTION

Many modular proteins perform their biological function in linear tandems of identical or nonidentical repeats. There are numerous examples of polyproteins that consist of ''head-to-tail'' (C-terminal-to-N-terminal) connected homogeneous domains $(D)_N = D$–D–. . .–D, as well as heterogeneous domains D_1–D_2–. . .–D_N. Titin is a giant polyprotein that contains tandem arrays of immunoglobulins (Ig) domains, separated by short linker sequences (1,2). The 27th and 28th domains of human cardiac titin (*I*27 and *I*28) have been extensively studied (3,4). Actin cross-linking filamins play an important role in cellular locomotion (5,6). The mechanical stability of filamin domains provide the necessary flexibility to actin cross-links (7,8). The most studied filamins are *ddFLN*, a dimeric filamin from *Dictyostelium discoideum*, and human filamin A protein. Both proteins have two subunits, one of which contains an actin-binding domain, whereas the other is a rodlike tandem of several Ig (*ddFLN*) or Ig-like domains (filamin A) (9). The extracellular matrix (ECM), which determines the elasticity and tensile strength of tissues, contains fibronectin (Fn) tandems. Fibronectin type III (FnIII) contains binding sites for the components of the ECM assembly triggered by mechanical stretching (10). Ubiquitin (Ub), a naturally occurring multimeric protein $(\mathrm{Ub})_N$ of $N = 9$ identical Ub repeats (11,12), is involved in protein degradation and several signaling pathways.

It has become possible to unfold individual domains in protein tandems by applying either relatively constant pulling force f (force-clamp), or time-dependent force (force-ramp) (4,13–15). Recently, the constant force-clamp technique was used to study the unfolding kinetics of single polyubiquitin and titin molecules (16–19). The forced unraveling of a polyprotein is identified by a series of stepwise increases of its end-to-end distance X, observed at times $\{t_1, t_2, \ldots, t_m\}$, which mark the unfolding of individual modules. For example, at $f \approx 120$ pN, the polyubiquitin chain elongates in steps of $\Delta X \approx 20$ nm, marking the unfolding of individual modules, or integer multiples of ΔX, which corresponds to the simultaneous unfolding of several modules (16,17).

Two approaches are currently used to analyze forced unfolding times of homogeneous protein tandems. The first is based on first-order kinetics of step-by-step unraveling of identical domains,

$$\{n\} \rightarrow \{n-1\} \rightarrow \ldots \rightarrow \{1\} \rightarrow \{0\},$$

with equal rates for each step, $k_{n,\, n-1} = \ldots = k_{1,0}$ (20,21). In the second approach, the average relative extension of the chain $x \equiv \langle X/L \rangle$ $(x \leq 1)$, where L is the total contour length of the tandem, is identified with the unfolding probability $P(t)$. The global unfolding rate, K, is then obtained from the fit of the theoretical probability $P(t) = 1 - \exp\,[-Kt]$ to x (17–19,22), assuming single-step exponential unfolding kinetics for each domain.

Both approaches are applicable when the unfolding times are 1), independent, 2), exponentially distributed, and 3), are realizations of the same probability density function (pdf). In

Submitted February 5, 2007, and accepted for publication April 10, 2007.

Address reprint requests to V. Barsegov, Tel.: 978-934-3661; Fax: 978-934-3013; E-mail: valeri_barsegov@uml.edu.

Editor: Feng Gai.

0006-3495/07/08/1100/16 $2.00

doi: 10.1529/biophysj.107.105866

statistical terms, these approaches are applicable to protein tandems characterized by independent identically distributed (iid) unfolding times with exponential parent pdf. The ''exponentiality condition'' holds when the forced unraveling of the individual proteins is described by a single-step transition $F \rightarrow U$, from the folded state (F) to the unfolded state (U). However, when unfolding involves the formation of intermediate species, the kinetics becomes nonexponential. The ''independence hypothesis'' is satisfied when the proteins are connected by long linkers that smooth out the effect of sudden chain elongation resulting from the unfolding of a particular domain, say D_m, on the immediate neighbor proteins (D_{m+1} and D_{m-1}), or when the cantilever tip picks up tandems of short length, i.e., $N = 2$–5, so that the unfolding transitions are separated by long waiting periods that ''decorrelate'' the consecutive unfolding times. However, the unfolding ''staircase'' $X(t)$ may involve a variable number of steps ranging from few to few tens, as well as simultaneous unraveling of more than one protein, resulting in correlated unfolding events (17). Whether the condition that unfolding times are identically distributed is satisfied depends on the tandem composition and the magnitude of applied force. The terminal proteins that are close to the point of force application could be subjected to stretching force for longer times as compared with domains in the interior of the tandem. This could make the unfolding times of otherwise identical repeats sample nonidentical distributions.

An alternative approach to forced unfolding data analysis can be based on analyzing ordered variates, i.e., first, second, third, etc., unfolding times. This approach has been used in the past in lifetime testing (23,24) and system reliability studies (25,26). In the context of forced rupture of hormone-receptor complexes, order statistics has been recently used by Tees et al. to describe the lifetimes of parallel bonds (27). In this article, we use order-statistics-based methodology to analyze forced unfolding that is characterized by independent (uncorrelated) unfolding times. This is the simplest case and computational formulas are straightforward. It also serves as the basic paradigm for understanding the ideas underlying order statistics and to further guide the extension of the methodology to dependent (correlated) unfolding times.

First, we present the order statistics formalism for independent identically distributed and independent nonidentically distributed (inid) unfolding times, which can be used to analyze forced unfolding data for homogeneous and heterogeneous tandems, respectively. To illustrate the use of the proposed approach, we simulate the force-induced unraveling of the homogeneous tandem $S2$–$S2$–$S2$, and the heterogeneous tandem $S2$–$S1$–$S2$ formed by model β-barrels $S2$ and $S1$. These protein models were used in the past to study the basic principles of protein folding (28). Next, we use order statistics to infer the (parent) unfolding time pdfs for the individual $S2$ and $S1$ domains. To validate the use of the order statistics formalism, we compare the parameters of the pdfs for domains $S2$ and $S1$, inferred from the order statistics analysis of unfolding data for the tandems, with the corresponding parameters obtained for single domains $S2$ and $S1$. We conclude in a discussion of our results and of future directions.

METHODS

Order statistics for independent random variables

Motivation

In a typical force-clamp atomic force microscopy (AFM) experiment, a protein tandem $(D)_N$ is subjected to the external constant stretching force f. The cantilever tip randomly picks up a tandem of any length n, $1 \leq n \leq N$. The unfolding trajectories of the tandem end-to-end distance X show characteristic stepwise increases $\{\Delta X(t_1), \Delta X(t_2), \ldots, \Delta X(t_n)\}$ recorded at times $\{t_1, t_2, \ldots, t_n\}$ (16–19), which mark the unfolding transition of individual domains. The main goal of unfolding time data analysis is to obtain the unfolding time distributions of individual proteins. However, since any number of n proteins can unfold at any given time, it is not possible to determine which specific domain unfolded at time t_i, $t_i \in \{t_1, t_2, \ldots, t_n\}$.

An alternative approach is based on recognizing that, by the very method of instrumentation used in force-clamp AFM, the recorded unfolding times of individual proteins are ordered, i.e., $t_1 < t_2 < \ldots < t_n$. That is, the observed unfolding times comprise a set of ordered time variates or order statistics characterized by cumulative distribution functions (cdfs) $\Phi_{r:n}(t)$, and probability density functions (pdfs) $\phi_{r:n}(t)$ of the r-th unfolding time, $r = 1, 2, \ldots, n$, in a tandem of length $n \leq N$. The r-th order statistic cdf $\Phi_{r:n}(t)$ is defined to be the probability that the r-th unfolding time t_r does not exceed t, i.e., $\Phi_r(t) = Prob(t_r \leq t)$, and the r-th order statistic pdf is given by $\phi_{r:n}(t) = d\Phi_{r:n}(t)/dt$. Both $\Phi_{r:n}(t)$ and $\phi_{r:n}(t)$ depend on the parent cdf $\Psi(t)$ and pdf $\psi(t)$, which represent the distributions of the individual proteins in the tandem. Hence, theory based on order statistics can be employed to infer the parent cdf $\Psi(t)$ and parent pdf $\psi(t)$, from the cdf $\Phi_{r:n}(t)$ and pdf $\phi_{r:n}(t)$ of the r-th order statistic, respectively.

The order statistical measures can be obtained by grouping the unfolding times into sets,

$$(\{t_{1:n_1}\}, \{t_{2:n_1}\}, \ldots, \{t_{n_1:n_1}\}), \ldots, (\{t_{1:N}\}, \{t_{2:N}\}, \ldots, \{t_{N:N}\}), \tag{1}$$

of ordered times, for each tandem length n_i with $n_1 < n_2 < \ldots < N$. The notation $t_{r:n}$ denotes the r-th smallest unfolding time (from below) among n times, and will be used in the rest of the article to identify ordered variates. Simply indexed variables, such as t_i, will indicate unordered variates. The sets in Eq. 1 can then be binned into histograms of ordered unfolding times, which can be subsequently used to estimate the order statistics pdfs,

$$(\phi_{1:n_1}, \phi_{2:n_1}, \ldots, \phi_{n_1:n_1}), \ldots, (\phi_{1:N}, \phi_{2:N}, \ldots, \phi_{N:N}). \tag{2}$$

In the pdfs in Eq. 2, the r-th and r'-th unfolding times ($r' \neq r$) in a tandem of length n have different distributions, i.e., $\phi_{r,\,n}(t) \neq \phi_{r',\,n}(t)$, and the r-th unfolding times in samples of size n_1 and $n_2 \neq n_1$ do not belong to the same distribution, i.e., $\phi_{r,n_1}(t) \neq \phi_{r,n_2}(t)$. By construction, $\psi(t) = \phi_{1:1}$, i.e., the pdf for the first order statistic in a tandem of unit size ($n = 1$) is the parent density when all individual parent pdfs are identical.

Iid unfolding times

The definition of order statistics does not require the random variables be iid. In this article, we focus on order statistics of iid random variables. The formalism for iid unfolding times can be used to analyze uncorrelated unfolding times of homogeneous tandems $(D)_N$. The r-th unfolding time $t_{r:n}$ for a tandem of length n is such that $r-1$ values are below it, and $n-r$ values are above it. By taking into account the number of all possible permutations, we obtain (23,24)

$$\Phi_{r:n}(t) = \sum_{m=r}^{n} \binom{n}{m} \Psi(t)^{m}(1-\Psi(t))^{n-m}$$

$$\phi_{r:n}(t) = n\binom{n-1}{r-1}\Psi(t)^{r-1}(1-\Psi(t))^{n-r}\psi(t), \qquad (3)$$

where $\Psi(t)$ ($\psi(t)$) are the parent cdf (pdf) and $\binom{n}{m} = n!/m!(n-m)!$ (Appendix I). Special cases are the extreme order statistics, i.e., the minimum $t_{1:n}$ with pdf $\phi_{1:n}(t) = n(1-\Psi(t))^{n-1}\psi(t)$, and the maximum $t_{n:n}$ with pdf $\phi_{n:n}(t) = n\Psi(t)^{n-1}\psi(t)$ (24,29).

The forced unfolding times of the r-th order can also be described by the average time, $\mu_{r:n}$, and its higher moments, $\mu_{r:n}^{(k)}$. The k-th moment of the r-th order statistic of iid unfolding times in a sample (tandem) of size (length) n is computed by evaluating the integral,

$$\mu_{r:n}^{(k)} = \int_0^{\infty} dt\phi_{r:n}(t)t^{k}$$

$$= \frac{1}{B(r,n-r+1)}\int_0^{\infty} dt\Psi(t)^{r-1}(1-\Psi(t))^{n-r}\psi(t)t^{k}, \qquad (4)$$

where $B(r, n-r+1) = (r-1)!(n-r)!/n!$ is the β-function (30), and $\mu_{r:n}$ is obtained by setting $k = 1$ in Eq. 4.

The cdf and pdf of the r-th and $(r+1)$-st order statistics in a sample of size (tandem length) n are related to the cdf and pdf of the r-th order statistic in a sample of reduced size $n-1$ via the recurrence relations

$$n\Phi_{r:n-1}(t) = (n-r)\Phi_{r:n}(t) + r\Phi_{r+1:n}(t),$$

$$n\phi_{r:n-1}(t) = (n-r)\phi_{r:n}(t) + r\phi_{r+1:n}(t). \qquad (5)$$

Equation 5 can be used in unfolding time data analysis to relate unfolding time distributions for tandems of different length ($n_1 \neq n_2$). Applying the second Eq. 5 recursively, we obtain the parent pdf $\psi(t)$:

$$n\psi(t) \equiv n\phi_{1:1}(t) = \sum_{r=1}^{n}\phi_{r:n}(t). \qquad (6)$$

This property provides a means to infer the parent distribution for a single domain D as $\psi(t) \equiv \phi_{1:1}(t)$ from the r-th order statistics pdfs $\phi_{r:n}$ ($1 \leq r \leq n$), when the unfolding times are iid. The case of exponential iid unfolding times is reviewed in Appendix I.

The moments of order statistics also satisfy similar recurrence relations. By multiplying both sides of the second Eq. 5 by t taken to the k-th power and evaluating the integral from zero to infinity (as in Eq. 4), we obtain

$$n\mu_{r:n-1}^{(k)} = (n-r)\mu_{r:n}^{(k)} + r\mu_{r+1:n}^{(k)}, \qquad (7)$$

for $1 \leq r \leq n-1$. By using Eq. 7 recursively, we obtain:

$$\mu^{(k)} \equiv \mu_{1:1}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\mu_{i:n}^{(k)}. \qquad (8)$$

In Eq. 8, $\mu^{(k)}$ is the k-th moment of the unfolding time for a single domain. By performing integration by parts and using the definition of the k-th moment of the r-th order statistic (Eq. 4) we obtain another useful formula,

$$\mu_{r+1:n}^{(k)} = \mu_{r:n}^{(k)} + k\frac{n!}{r!(n-r)!}\int_0^{\infty} dt\Psi(t)^{r}(1-\Psi(t))^{n-r}\psi(t), \qquad (9)$$

which allows computation of the k-th moment of the unfolding times of order $r+1$ from the k-th moments of lower orders.

Inid unfolding times

Order statistics for independent nonidentically distributed (inid) random variables can be used to analyze uncorrelated unfolding times for heterogeneous tandems. The cdfs and pdfs are expressed in terms of the permanent of a matrix A, per$[A]$ (31). The permanent is defined as the determinant except that all signs are positive (32,33) (see also Appendix II). More specifically,

$$\Phi_{r:n}(t) = \sum_{m=r}^{n}\frac{1}{m!(n-m)!}\,\mathrm{per}\begin{array}{c}\begin{bmatrix} \Psi_1(t) & 1-\Psi_1(t) \\ \Psi_2(t) & 1-\Psi_2(t) \\ \vdots & \vdots \\ \Psi_n(t) & 1-\Psi_n(t) \end{bmatrix} \\ \begin{matrix} m & \quad m-r \end{matrix}\end{array}$$

$$\phi_{r:n}(t) = \frac{1}{(r-1)!(n-r)!}\,\mathrm{per}\begin{array}{c}\begin{bmatrix} \Psi_1(t) & 1-\Psi_1(t) & \psi_1(t) \\ \Psi_2(t) & 1-\Psi_2(t) & \psi_2(t) \\ \vdots & \vdots & \vdots \\ \Psi_n(t) & 1-\Psi_n(t) & \psi_n(t) \end{bmatrix} \\ \begin{matrix} r-1 & \quad n-r & \quad 1 \end{matrix}\end{array}. \qquad (10)$$

The numbers under the columns indicate number of copies of each column. Recurrence relations for inid variables read (34)

$$\sum_{i=1}^{n}\Phi_{r:n-1}^{[i]}(t) = r\Phi_{r+1:n}(t) + (n-r)\Phi_{r:n}(t)$$

$$\sum_{i=1}^{n}\phi_{r:n-1}^{[i]}(t) = r\phi_{r+1:n}(t) + (n-r)\phi_{r:n}(t), \qquad (11)$$

where $r = 1, \ldots, n-1$. In Eqs. 11, $\Phi_{r:n-1}^{[i]}(x)$ and $\phi_{r:n-1}^{[i]}(x)$ denote, respectively, the cdf and pdf of the r-th order statistic in the reduced sample of size $n-1$, which is obtained from the original sample by dropping the i-th variate t_i, $i = 1, \ldots, n$.

Langevin simulations of tandems *S*2–*S*2–*S*2 and *S*2–*S*1–*S*2

To test the use of ordered variates in analyzing forced unfolding of protein tandems, we performed Langevin simulations of forced unfolding using a coarse-grained model of the homogeneous protein tandem $S2$–$S2$–$S2$, and heterogeneous tandem $S2$–$S1$–$S2$ of linearly connected domains $S2$ and $S1$. The off-lattice C_α-based coarse grained models (CGM) (35–37) of protein tandems serve as conceptual representations of the wild-type multidomain proteins. The tandem length $n = 3$ allows us to analyze the first ($r = 1$), second ($r = 2$), and third ($r = 3$) order statistics of unfolding times.

Tandem construction

Tandem $S2$–$S2$–$S2$ is constructed by ''head-to-tail'' linking three β-barrel domains $S2$. Tandem $S2$–$S1$–$S2$ is constructed by linking two $S2$ domains and one β-sheet domain $S1$. Domains $S2$ and $S1$ are connected through flexible linkers of five neutral residues (Fig. 1). Both $S1$ and $S2$ consist of 46 hydrophobic (B), hydrophilic (L), and neutral (N) residues. Each bead is represented by a united atom at the position of the C_α-atom (Fig. 1). The distance between C_α-carbons is $a = 3.8$ Å. The contour length of single $S1$ and $S2$ is equal to $L = 46a$. The potential energy of the tandem conformations,

$$V = V_{\mathrm{BL}} + V_{\mathrm{BA}} + V_{\mathrm{DIH}} + V_{\mathrm{NB}}, \tag{12}$$

includes the bond length (V_{BL}) and bond angle (V_{BA}) potentials, the dihedral angle potential (V_{DIH}), and nonbonded potential (V_{NB}) introduced elsewhere (28,37). The nonbonded distance (R) dependent interaction between a pair of B residues is given by $V_{\mathrm{NB}}^{\mathrm{BB}}(R) = 4\lambda\varepsilon_{\mathrm{h}}\left[(a/R)^{12} - (a/R)^{6}\right]$, where λ accounts for variation in the strength of hydrophobic interactions and $\epsilon_{\mathrm{h}} = 1.25$ kcal/mol is the average strength of hydrophobic contacts. The interaction between all other residues is repulsive (28,37). The energy function includes attractive interactions but does not distinguish chiral states. The native structures of both $S1$ and $S2$ have $Q_0 = 106$ native contacts (with 6.8 Å cutoff) and potential energies of -85.5 kcal/mol and -88.0 kcal/mol, respectively.

Forced unfolding simulations

The forced unfolding kinetics are obtained by integrating the Langevin equations,

$$\eta \frac{d\mathbf{x}_{\mathrm{j}}}{dt} = -\frac{\partial V_{\mathrm{tot}}}{\partial \mathbf{x}_{\mathrm{j}}} + \mathbf{g}_{\mathrm{j}}(t), \tag{13}$$

for each residue coordinate $\mathbf{x}_{\mathrm{j}}$, subject to the total potential $V_{\mathrm{tot}} = V - \mathbf{fX}$. The stretching force $\mathbf{f} = f\mathbf{n}$ of magnitude $f = 66$ pN, is applied to both C- and N-terminals of the tandems in the fixed direction $\mathbf{n}$ (Fig. 1). In Eq. 13, η is the friction coefficient and $\mathbf{g}_{\mathrm{j}}$ is Gaussian white noise. Equation 13 was numerically integrated with step size $\delta t = 0.05\tau_{\mathrm{L}}$, where $\tau_{\mathrm{L}} = (ma^2/\epsilon_{\mathrm{h}})^{1/2} =$ 3 ps is the unit of time, and $m \approx 3 \times 10^{-22} g$ is the residue mass, at temperature $T_{\mathrm{s}} = 0.69\epsilon_{\mathrm{h}}/k_{\mathrm{B}}$. T_{s} is below the equilibrium folding temperature $T_{\mathrm{F}} \approx 0.79\epsilon_{\mathrm{h}}/k_{\mathrm{B}}$ for $S1$ and $S2$, and is the temperature at which the average fraction of native contacts $\langle Q(T_{\mathrm{s}})\rangle/Q_0 \approx 0.7$. Before generating the forced unfolding transitions, the initially folded structures of $S2$–$S2$–$S2$ and $S2$–$S1$–$S2$ were equilibrated for 100 ns at $T = T_{\mathrm{s}}$. The unfolding time of an individual domain $S2$ or $S1$ in the tandem was defined as the time at which all contacts (native and nonnative) were disrupted for the first time. Throughout the article the unfolding rates and times are expressed in terms of the number of integration steps N_{tot}. The transformation from N_{tot} to time t is given by $t = N_{\mathrm{tot}}\delta t$.

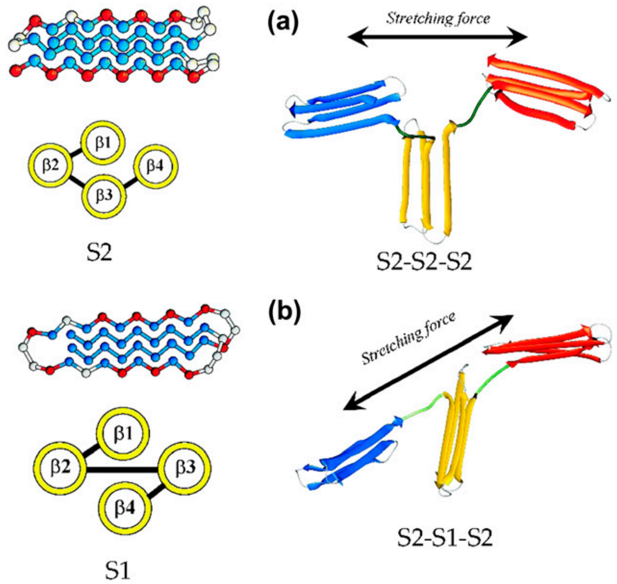

FIGURE 1 (a) ($left$) Model protein $S2$ (three-letter code $B_9N_3(LB)_3NBLN_2B_9N_3(LB)_5L$) formed by the hydrophobic residues (B, *blue spheres*), hydrophilic residues (L, *red*), and neutral residues (N, *gray*); (a, *right*) The heterogeneous tandem $S2$–$S2$–$S2$ of $S2$ domains ''head-to-tail'' connected by the flexible linkers of five neutral residues (*green*). (b, *left*) Model protein $S1$ (three-letter code $B_9N_3(LB)_3NBLN_2LBN(BL)_4N_2LB_9$); ($b$, *right*) the heterogeneous tandem $S2$–$S1$–$S2$ of $S2$ domains, separated by $S1$ domain, connected by the linkers as in tandem $S2$–$S2$–$S2$. The direction of constant force $\mathbf{f}$ is shown by the arrow.

RESULTS

Preliminary analysis of unfolding times for tandems *S*2–*S*2–*S*2 and *S*2–*S*1–*S*2

To prepare the stage for the use of order statistics, in this section we: a), analyze the unfolding times of single domains $S2$ and $S1$ and estimate their parent distributions $\psi(t)_{\mathrm{S2}}$ and $\psi(t)_{\mathrm{S1}}$, and b), test the unordered unfolding times for domains $S1$ and $S2$ in tandems $S2$–$S2$–$S2$ and $S2$–$S1$–$S2$ for independence and (in)equality of their (parent) distributions. We show that at $f = 66$ pN, the unfolding times for the $S2$–$S2$–$S2$ tandem are iid random variables, whereas the unfolding times for the $S2$–$S1$–$S2$ tandem form a set of inid variables.

Single domains S*2 and* S*1*

We generated 477 unfolding trajectories for single domain $S2$, and 300 for single domain $S1$, by stretching the protein at constant force $f = 66$ pN. The histograms of the unfolding times for $S2$ and $S1$ and the nonparametric density estimates are presented in Fig. 2. A density estimate, which provides a visual assessment of the distribution, attempts to estimate the density by locally weighting the observations (38–40). Kernel density estimates, such as those shown in Fig. 2, are a generalization of histograms (Appendix III). Both histograms and density estimates indicate that the unfolding times of $S2$ and $S1$ are nonidentically distributed. Quantile-quantile ($Q - Q$) plots of the unfolding times for $S2$ and $S1$ versus quantiles of the exponential distribution, k exp $[-kt]$, with rate $k = k_{\mathrm{S1}}$ for $S1$ and $k = k_{\mathrm{S2}}$ for $S2$, and the Gamma distribution (Eq. 14) are given in Fig. 3. The parameters of both the Gamma and the exponential distribution were computed using the maximum likelihood estimation (MLE) method (Appendix IV).

A Q-Q plot is a graphical technique for determining whether two data sets come from populations with a common distribution (41). For example, if the first sample has nine data points, say 2, 3, 0, 1, 6, 10, 7, 5, 20, the 0.1, 0.2, . . . , 0.9 quantiles are calculated and placed on the x axis. The quantiles of the second sample, say 1.5, 3, 2, 6, 13, 8, 10, 5,

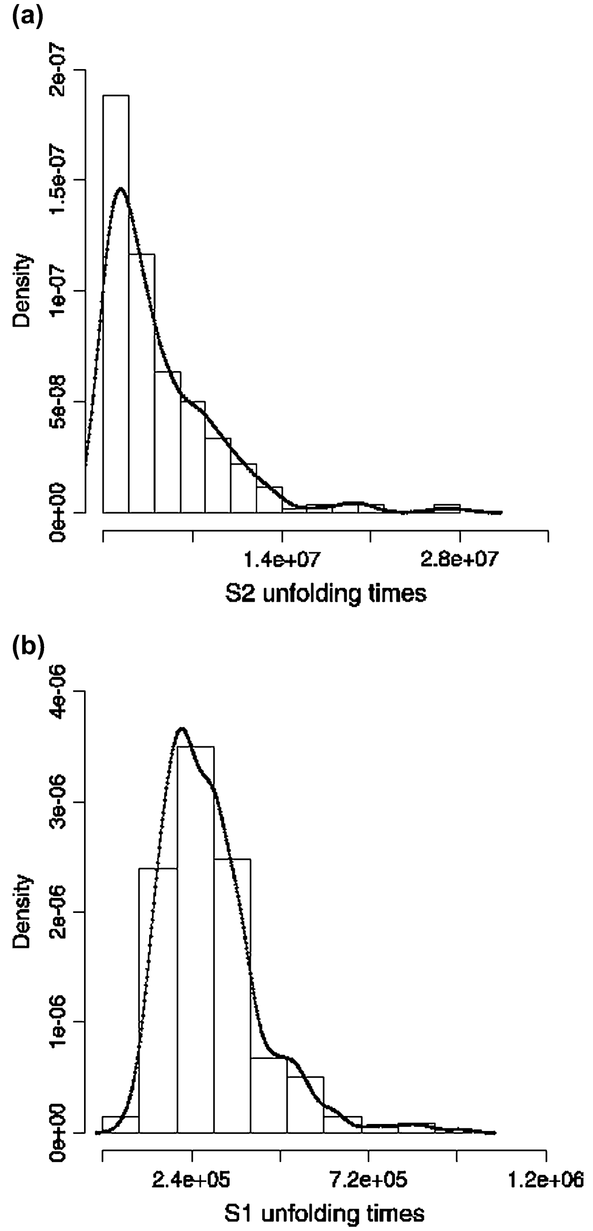

FIGURE 2 The distributions of forced unfolding times (in units of integration steps) for single *S*2 domain (*a*) and single *S*1 domain (*b*) obtained at $f = 66$ pN. The histograms of unfolding times are based on number of bins, n_b, computed using Sturge's formula $n_b = 1 + \log_2 n_p$, where n_p is the number of data points (39). The overlaid curve is the nonparametric density estimate (the bandwidth $bw = 0.9 \times \min(SD, IQR/1.34) n_p^{-1/5}$ is the default value used in the *R* software (40), where *SD* is the standard deviation, and *IQR* is the interquantile range (38)). The number of bins n_b and the bandwidth *bw* in Figs. 6–8, 10, and 12 are estimated as described above.

11, are placed on the *y* axis. A quantile is a number x_p such that $100 \times p\%$ of the data values are $\leq x_p$. For example, the 0.25 quantile (25-th percentile) of a variable is a value $x_{0.25}$ such that 25% of the data are less than or equal to that value. For the two data sets in this example, the quantiles of the first set are $x_{0.1} = 0, x_{0.2} = 1, x_{0.3} = 2, x_{0.4} = 3, x_{0.5} = 5, x_{0.6} = 6, x_{0.7} = 7, x_{0.8} = 10, x_{0.9} = 20$, and the quantiles of the second set are $y_{0.1} = 1.5, y_{0.2} = 2, y_{0.3} = 3, y_{0.4} = 5, y_{0.5} = 6, y_{0.6} = 8, y_{0.7} = 10, y_{0.8} = 11, y_{0.9} = 13$. The *Q-Q* plot is the scatterplot of the corresponding quantiles ($\{x_p, y_p\}$) of the two data sets. The 45° line (with slope 1) serves as the reference line. If the two sets come from a population with the same distribution, the points fall along the reference line. The greater the departure from the line, the greater the evidence that the two data sets come from populations with different distributions.

The *Q-Q* plots for single domains *S*2 and *S*1 show smaller deviations from the reference line for Gamma quantiles compared with the *Q-Q* plots with exponential quantiles (Fig. 3). This is particularly evident for the *S*2 unfolding times. We used the Gamma ansatz to parametrize the parent pdfs for *S*2 and *S*1,

$$\psi_{\mathrm{gamma}}(t) = \frac{k^{\alpha}}{\Gamma(\alpha)} t^{\alpha-1} e^{-\mathrm{kt}}, \qquad (14)$$

where α is the shape parameter and k is the unfolding rate. The Gamma ansatz is a generalization of the exponential distribution and can be used to model lifetime data (when $\alpha = 1$, $\psi_{\mathrm{gamma}}(t)$ is the exponential pdf with rate k). We parametrized the unfolding times for single *S*2 and *S*1 domains using Eq. 16 and ML estimation. The ML estimates of α and k for single domains *S*2 and *S*1 are given in Table 1.

Forced unfolding of S2–S2–S2 and S2–S1–S2

We generated 500 unfolding trajectories for *S*2–*S*2–*S*2 and *S*2–*S*1–*S*2. Fig. 4 displays the tandem end-to-end distance X as a function of the number of integration steps N_{tot} for a few runs. We extracted the unfolding times for the first (*S*2) domain t_1, second (*S*2 or *S*1) domain t_2, and the third (*S*2) domain t_3. The obtained data sets ($\{t_1\}$, $\{t_2\}$, and $\{t_3\}$) were tested for independence and equality of distributions.

Test for independence

The most widely used measure of dependence between random variables T_1 and T_2 is the (Pearson) correlation coefficient (42), which is an appropriate measure of dependence when the random variables have a multivariate normal or an elliptical distribution. However, protein unfolding times are not normally distributed (Fig. 2). More appropriate are nonparametric and scale-invariant measures (43), such as the Spearman's rank correlation coefficient defined by

$$R = \frac{12\sum_{\mathrm{i}=1}^{\mathrm{n}} R_{\mathrm{i}} S_{\mathrm{i}}}{n(n^2-1)} - \frac{3(n+1)}{n-1}, \qquad (15)$$

where $R_{\mathrm{i}} = \mathrm{rank}(T_{1\mathrm{i}})$, $S_{\mathrm{i}} = \mathrm{rank}(T_{2\mathrm{i}})$, $i = 1, \ldots, n$. R lies between -1 and 1 and is zero when T_1 and T_2 are independent

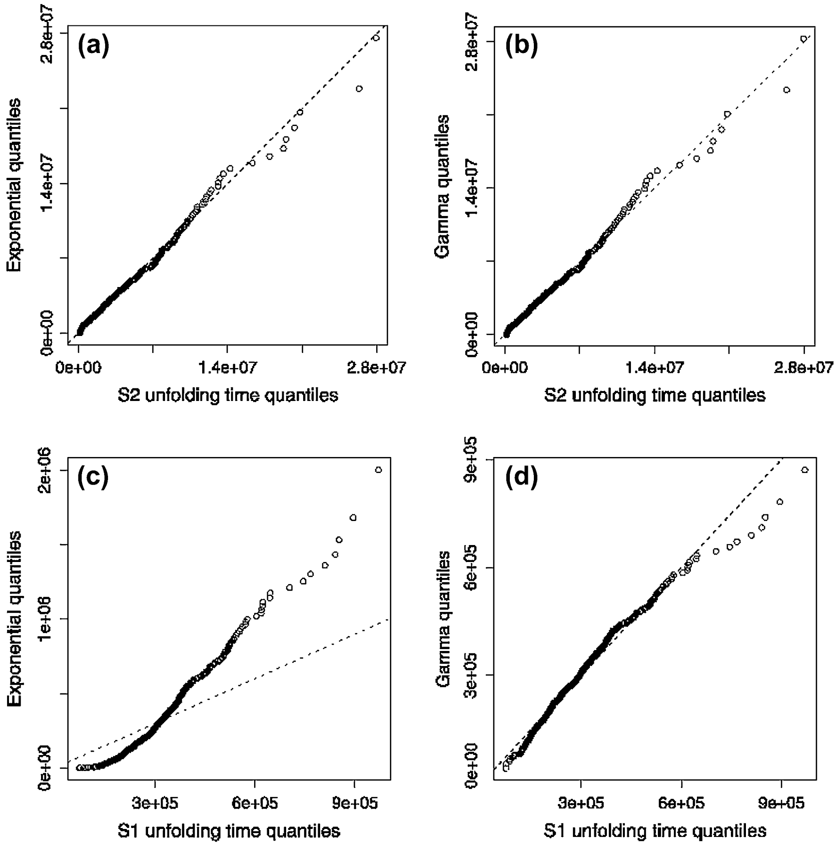

FIGURE 3 *Q-Q* plots of the forced unfolding times (in units of integration steps) for single *S*2 domain (*a*, *b*) and single *S*1 domain (*c*, *d*) for the data displayed in Fig. 2. (*a* and *c*) Quantiles (*open circles*) of unfolding times for single *S*2 and *S*1 domains versus exponential quantiles. (*b* and *d*) Quantiles of unfolding times for single *S*2 and *S*1 domains versus Gamma quantiles. The dashed line is the 45° reference line.

(42). If the sample size is large, the random variable $Z = R\sqrt{n-1}$ is approximately standard normal, when T_1 and T_2 are independent.

Results for S2–S2–S2. The values of $R(t_i, t_j)$ $(i, j = 1, 2, 3)$ for the unfolding times ($\{t_1\}$, $\{t_2\}$, $\{t_3\}$) of *S*2–*S*2–*S*2 are given in Table 2. We see that $R(t_1, t_2) = 0$, $R(t_1, t_3) = -0.11$, and $R(t_2, t_3) = -0.08$. The *p*-values for testing that $R(t_i, t_j) = 0$ (i.e., t_i and t_j are independent) were 0.94, 0.02, and 0.06, respectively. The *p*-values were computed using the $Z = R\sqrt{n-1}$ test statistic since the sample size is large enough (500) to use the standard normal approximation. The *p*-value of 0.02 means that there is a 2% chance of observing a Spearman rank correlation coefficient as large (in absolute value) as the observed even if the two unfolding times t_1 and t_3 are independent. The threshold *p*-value, which represents the level of tolerance for rejecting the independence hypothesis, was set to 0.01. In statistical hypothesis testing, the null hypothesis is rejected if the *p*-value is smaller than the threshold. At level 0.01, all pairs ($\{t_i\}$, $\{t_j\}$) are uncorrelated, implying that the *S*2 domains in the tandem *S*2–*S*2–*S*2 unfold independently. We also examined whether the unfolding times of the first *S*2 ($S2_1$), second *S*2 ($S2_2$), and third *S*2 domain ($S2_3$) are identically distributed. The *Q-Q* plots in Fig. 5 show the $S2_i$ unfolding time quantiles against

TABLE 1 The maximum likelihood estimates of α and *k* (in units of integration steps) with 95% confidence intervals for single domains *S*2 and *S*1 obtained at $f = 66$ pN

Single domain	Shape parameter α	Unfolding rate *k*
*S*1	4.92 ± 0.71	$(1.66 \pm 0.3) \times 10^{-5}$
*S*2	0.96 ± 0.14	$(2.22 \pm 0.5) \times 10^{-7}$

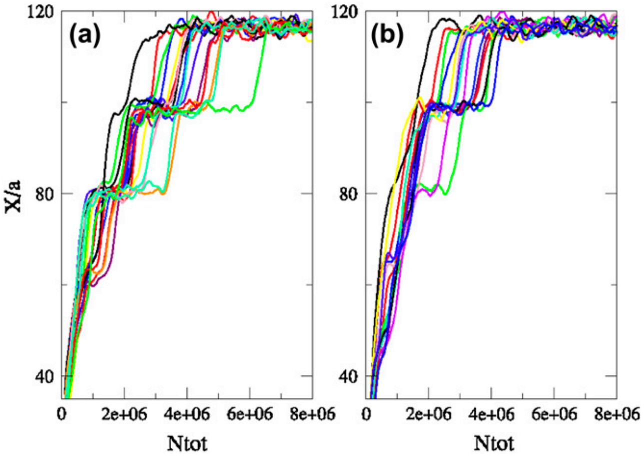

FIGURE 4 Trajectories of the tandem end-to-end distance, *X*/*a*, as a function of time (in units of integration steps, N_{tot}) for tandem *S*2–*S*2–*S*2 (*a*) and *S*2–*S*1–*S*2 (*b*) obtained at $f = 66$ pN. The progress of unfolding is witnessed as a series of stepwise increases in *X*/*a*.

TABLE 2 Spearman rank correlation coefficients for unfolding times for domains $S2_1$ (t_1), $S2_2$ (t_2), and $S2_3$ (t_3) in tandem $S2$–$S2$–$S2$, and for domains $S2_1$ (t_1), $S1_2$ (t_2), and $S2_3$ (t_3) in tandem $S2$–$S1$–$S2$, obtained at $f = 66$ pN

S2-S2-S2				*S2-S1-S2*			
Time	t_1	t_2	t_3	Time	t_1	t_2	t_3
t_1	1	0.0 (0.94)	−.11 (0.02)	t_1	1	−0.09 (0.04)	−.07 (0.13)
t_2	0	1	−.08 (0.06)	t_2	−0.09 (0.04)	1	−0.03 (0.44)
t_3	−0.11 (0.02)	−0.08 (0.06)	1	t_3	−0.07 (0.13)	−0.03 (0.44)	1

The p-values for testing $R = 0$ are in parentheses.

the corresponding $S2_j$ quantiles ($i, j = 1, 2, 3, i \neq j$). For $S2$–$S2$–$S2$, almost all data points fall on the reference line, indicating approximate equality of the parent distributions, i.e., $\psi_{S2_1}(t) \approx \psi_{S2_2}(t) \approx \psi_{S2_3}(t)$. The histograms of the unfolding times for $S2_1$, $S2_2$, and $S2_3$ are displayed in Fig. 6 along with kernel density estimates. Both histograms and density estimates agree with the *Q-Q* plots that the unfolding times of the three identical proteins $S2_1$, $S2_2$, and $S2_3$ are identically distributed.

Results for S2–S1–S2. The values of $R(t_i, t_j)$ for the unfolding times of $S2$–$S1$–$S2$ are given in Table 2. We see that $R(t_1, t_2) = -0.09$, $R(t_1, t_3) = -0.07$, and $R(t_2, t_3) = -0.03$ with p-values 0.04, 0.13, and 0.43, respectively. At level 0.01, none of the unfolding time pairs are correlated and we conclude that the $S2$ and $S1$ domains in the tandem *S2–S1–S2* unfold independently. The *Q-Q* plots (Fig. 5) indicate approximate equality of the distributions for terminal $S2$ domains, i.e., $\psi_{S2_1}(t) \approx \psi_{S2_3}(t)$. Large deviations from the reference line in the *Q–Q* plots for $S2_1$ vs. $S1_2$, and $S1_2$ vs. $S2_3$ (Fig. 5) indicate that the unfolding times of $S2$ and $S1$ have different distributions ($\psi_{S2_1}(t) \neq \psi_{S1_2}(t)$, $\psi_{S2_3}(t) \neq \psi_{S1_2}(t)$). The histograms of unfolding times and density estimates (Fig. 7) agree with the findings from the *Q-Q* plots.

Order statistics: from ordered variates to parent densities

To generate ordered time variates as observed in force-clamp experiments, the unfolding time triplets $\{t_1\}$, $\{t_2\}$, $\{t_3\}$ for domains $S2_1$ (t_1), $S2_2$ (t_2), and $S2_3$ (t_3) in $S2$–$S2$–$S2$, and domains $S2_1$ (t_1), $S1_2$ (t_2), and $S2_3$ (t_3) in $S2$–$S1$–$S2$, were rearranged in increasing time order. That is, $t_{\text{min}} < t_{\text{med}} < t_{\text{max}}$, where $t_{\text{min}} = \min(t_1, t_2, t_3)$, $t_{\text{med}} = \min\{(t_1, t_2, t_3) - t_{\text{min}}\}$, and $t_{\text{max}} = \max(t_1, t_2, t_3)$ are the minimum, median, and maximum unfolding time, respectively. The ordered variates from 500 runs were grouped into ordered sets of the first $\{t_{\text{min}}\} = \{t_{1:3}\}$, second $\{t_{\text{med}}\} = \{t_{2:3}\}$, and third $\{t_{\text{max}}\} = \{t_{3:3}\}$ unfolding times.

Iid unfolding times for S2–S2–S2. The forced unfolding times for *S2–S2–S2* form a set of iid random variables. The histograms and density estimates for the three order statistics $\{t_{1:3}\}$, $\{t_{2:3}\}$, and $\{t_{3:3}\}$ are displayed in Fig. 8. The parent pdf $\psi(t)$ can be obtained by ''summing over'' the order statistics pdfs (see Eq. 6), i.e., $\psi(t) \equiv \phi_{1:1}(t) = 1/n \sum_{r=1}^{n} \phi_{r:n}(t)$. This procedure is equivalent to binning all three variates $t_{r:3}$, $r = 1, 2, 3$, into a single histogram, i.e., all unfolding time data are combined into a single sample (of size $3 \times 500 = 1500$). This is common practice in statistical analyses of forced unfolding times obtained in force-clamp measurements on homogeneous tandems, such as $S2$–$S2$–$S2$. However, it is justified only when unfolding times are iid variables. The unfolding time histogram and density estimates for the single $S2$ domain (Fig. 2) and combined sample are superposed in Fig. 9.

The agreement between the corresponding histograms and between the density estimates is excellent, pointing to the fact that the two distributions (ψ_{S2}) for single $S2$ domain, and $\psi(t) = 1/n \sum_{r=1}^{n} \phi_{r:n}(t)$, constructed from the order statistics pdfs, are identical. This finding provides an empirical proof of Eq. 6 when it is applied to describing iid unfolding times. Using maximum likelihood estimation on the combined data, the shape was estimated to be $\hat{\alpha}_{S2} = 1.18$ and the rate was estimated to be $\hat{k}_{S2} = 2.24 \times 10^{-7}$, with 95% confidence intervals (1.09, 1.27) and (2.03, 2.49) × 10^{-7}, respectively. Comparing these estimates with the corresponding ML estimates for the single $S2$ domain, $\hat{\alpha}_{S2} = 0.96$ and $\hat{k}_{S2} = 2.22 \times 10^{-7}$ with respective 95% confidence intervals (0.83, 1.10) and (1.87, 2.72) × 10^{-7}, we see that the estimates for single $S2$ and the combined sample, obtained by using the order statistics pdfs, are fairly similar.

Inid unfolding times for S2–S1–S2. The unfolding times for the heterogeneous tandem *S2–S1–S2* form a set of inid random variables. The histograms and density estimates for the three order statistics $\{t_{1:3}\}$, $\{t_{2:3}\}$, and $\{t_{3:3}\}$ are displayed in Fig. 10. The pdf $\phi_{1:3}(t)$ of the first order statistic shows a bimodal profile with the peak at shorter (longer) times corresponding to unfolding of less (more) stable domain $S1$ ($S2$). We used the Gamma distribution (Eq. 14) to model the parent pdfs for domains $S2$ and $S1$. Thus, we set $\psi_{S2}(t) = k_{S2}^{\alpha_{S2}} t^{\alpha_{S2}-1} \exp[-k_{S2}t]/\Gamma(\alpha_{S2})$ and $\psi_{S1}(t) = k_{S1}^{\alpha_{S1}} t^{\alpha_{S1}-1} \exp[-k_{S1}t]/\Gamma(\alpha_{S1})$, where α_{S2}, α_{S1}, and k_{S2}, k_{S1}, are the shape and unfolding rates for $S2$ and $S1$, respectively, and $\gamma(\alpha, y) \equiv \int_0^y x^{\alpha-1^{-x}} dx$. The cdfs for $S2$ and $S1$ are $\Psi_{S2}(t) = \gamma(\alpha_{S2}, k_{S2}t)/\Gamma(\alpha_{S2})$ and $\Psi_{S1}(t) = \gamma(\alpha_{S1}, k_{S1}t)/\Gamma(\alpha_{S1})$, respectively. The expressions for the pdfs of the first ($\phi_{1:3}$), second ($\phi_{2:3}$), and third ($\phi_{3:3}$) order statistics for the $S2$–$S1$–$S2$ tandem are obtained by inserting ψ_{S2}, ψ_{S1}, Ψ_{S2}, and Ψ_{S1} into the following equations (see Eq. 10),

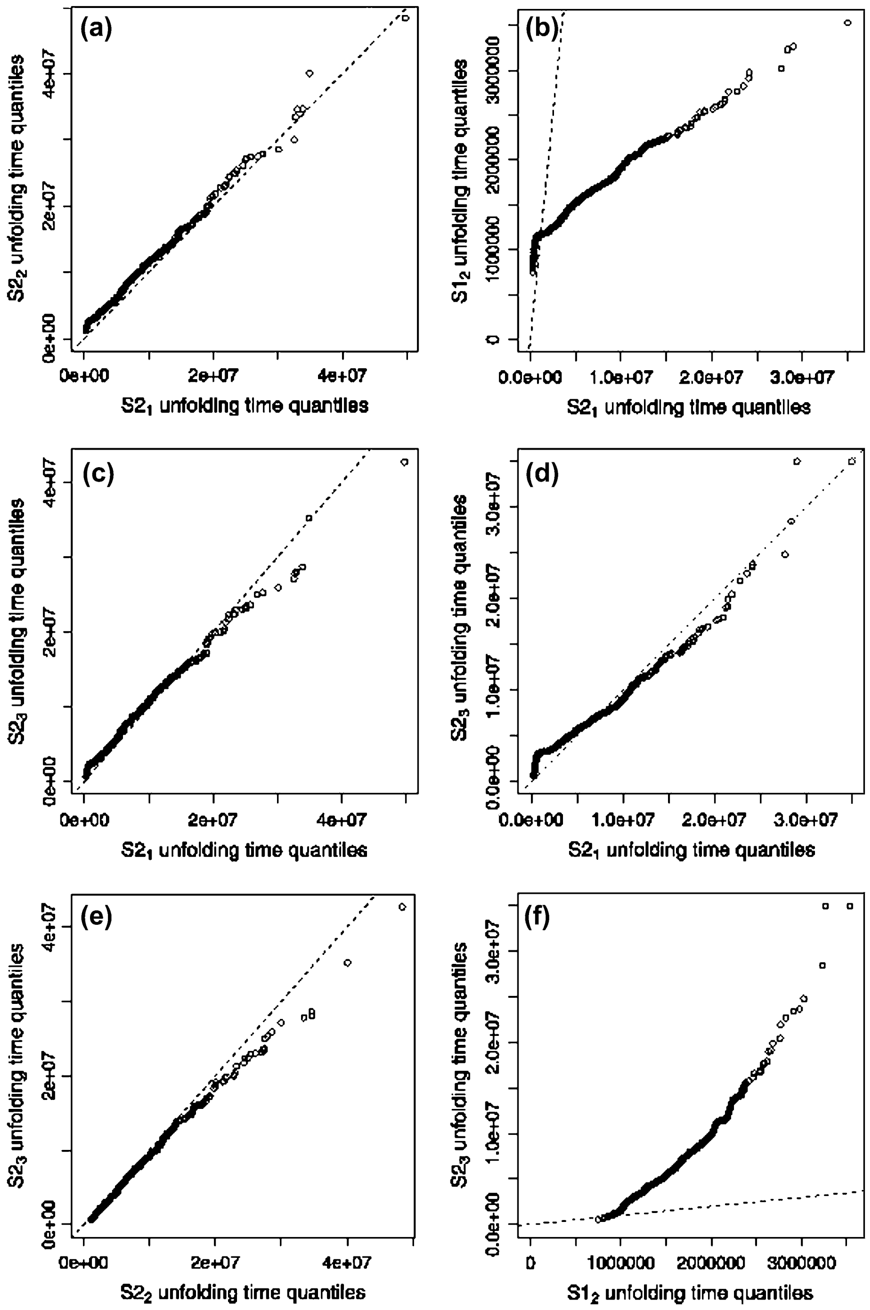

FIGURE 5 *Q-Q* plots of forced unfolding times (t_1, t_2) (*a*), (t_1, t_3) (*c*), and (t_2, t_3) (*e*) for domains $S2_1$ (t_1), $S2_2$ (t_2), and $S2_3$ (t_3) in tandem $S2$–$S2$–$S2$ obtained at $f = 66$ pN. *Q-Q* plots of forced unfolding times (t_1, t_2) (*b*), (t_1, t_3) (*d*), and (t_2, t_3) (*f*) for domains $S2_1$ (t_1), $S1_2$ (t_2), and $S2_3$ (t_3) in tandem $S2$–$S1$–$S2$ obtained at $f = 66$ pN.

$$\phi_{1:3}(t) = \frac{1}{2}\mathrm{per}\begin{bmatrix} 1-\Psi_1(t) & 1-\Psi_1(t) & \psi_1(t) \\ 1-\Psi_2(t) & 1-\Psi_2(t) & \psi_2(t) \\ 1-\Psi_3(t) & 1-\Psi_3(t) & \psi_3(t) \end{bmatrix}$$

$$\phi_{2:3}(t) = \mathrm{per}\begin{bmatrix} \Psi_1(t) & 1-\Psi_1(t) & \psi_1(t) \\ \Psi_2(t) & 1-\Psi_2(t) & \psi_2(t) \\ \Psi_3(t) & 1-\Psi_3(t) & \psi_3(t) \end{bmatrix}$$

$$\phi_{3:3}(t) = \frac{1}{2}\mathrm{per}\begin{bmatrix} \Psi_1(t) & \Psi_1(t) & \psi_1(t) \\ \Psi_2(t) & \Psi_2(t) & \psi_2(t) \\ \Psi_3(t) & \Psi_3(t) & \psi_3(t) \end{bmatrix}. \tag{16}$$

Order statistics pdfs, given by Eq. 16, were used to fit the histograms of the ordered first ($\{t_{1:3}\}$), second ($\{t_{2:3}\}$), and third ($\{t_{2:3}\}$) unfolding times (Fig. 10). The results, displayed in Fig. 11, show good agreement between the histograms of ordered variates and theoretical order statistics pdfs. The values of the fitting parameters are given in Table 3. The shape parameters α_{S2} and α_{S1}, and unfolding rates k_{S2} and k_{S1}, for $S2$ and $S1$ domains in the tandem $S2$–$S1$–$S2$, compare well with the ML estimates of the same quantities for the single $S2$ and $S1$ domains. However, the values for α_{S2} and α_{S1} for tandem $S2$–$S1$–$S2$ obtained from the numerical fit are

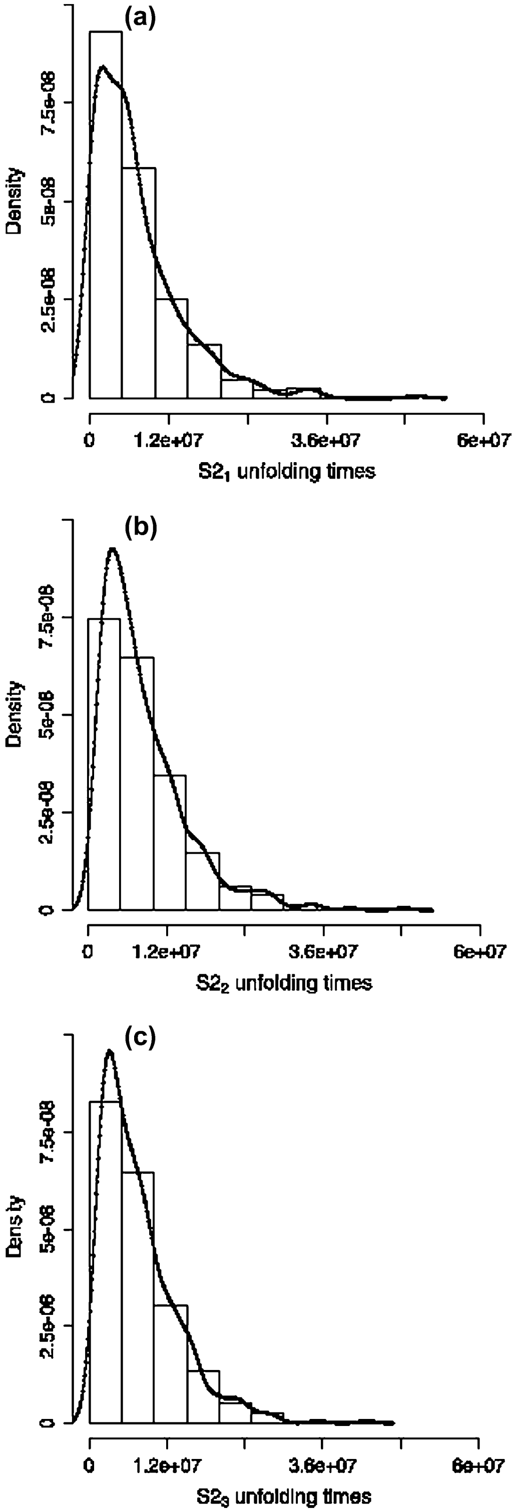

FIGURE 6 The distributions of forced unfolding times (in units of integration steps) for domains $S2_1$ (*a*), $S2_2$ (*b*), and $S2_3$ (*c*) in tandem *S*2–*S*2–*S*2 obtained at *f* = 66 pN. Histograms of the unfolding times are overlaid with corresponding nonparametric density estimates.

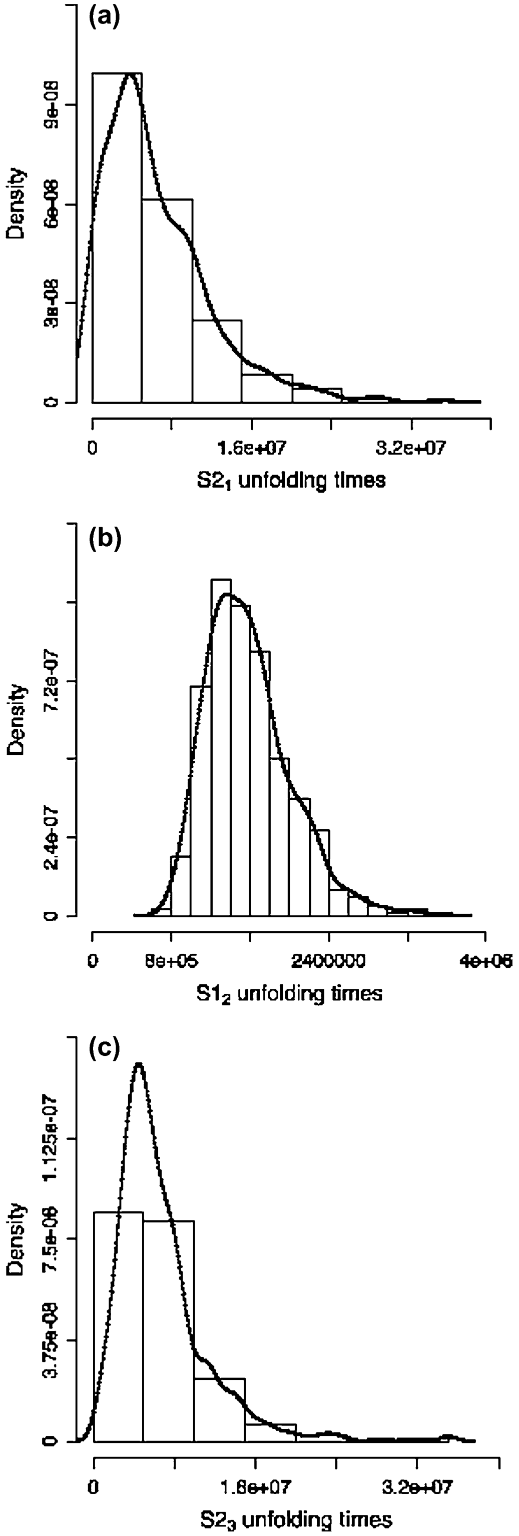

FIGURE 7 The distributions of forced unfolding times (in units of integration steps) for domains $S2_1$ (*a*), $S1_2$ (*b*), and $S2_3$ (*c*) in tandem *S*2–*S*1–*S*2 obtained at *f* = 66 pN. Histograms of the unfolding times are overlaid with corresponding nonparametric density estimates.

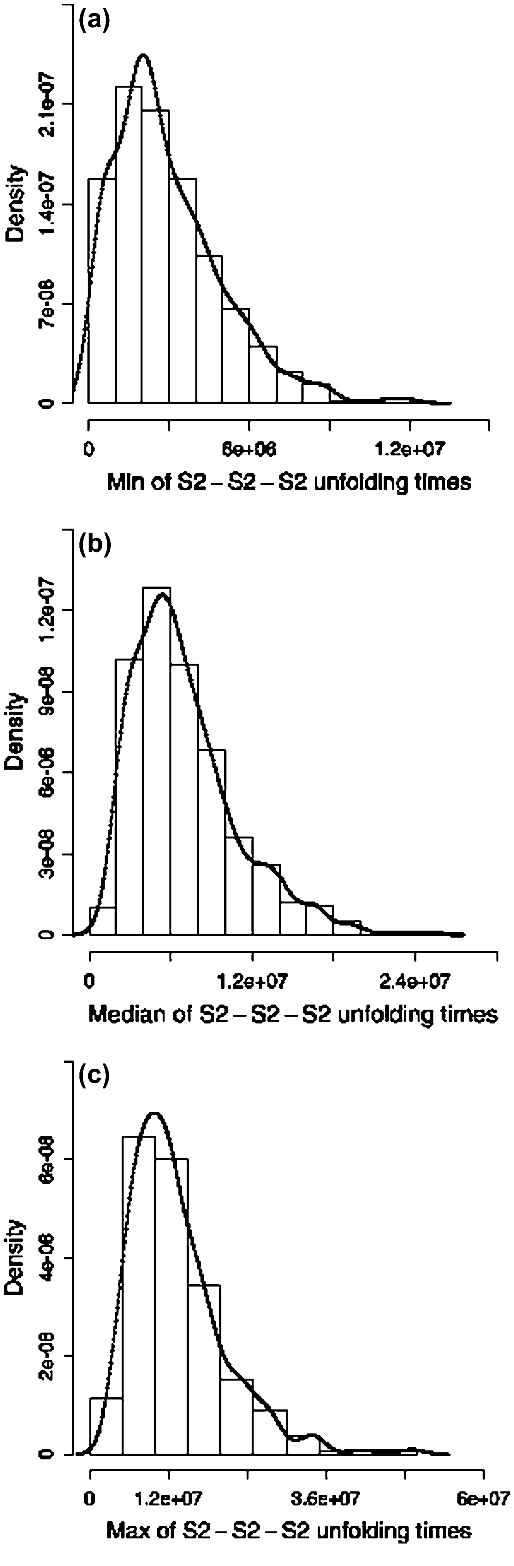

FIGURE 8 Histograms of ordered unfolding times (in units of integration steps) for tandem *S*2–*S*2–*S*2 obtained at $f = 66$ pN. Histograms of the minimum ($\{t_{1:3}\}$, *a*), median ($\{t_{2:3}\}$, *b*), and maximum ($\{t_{3:3}\}$, *c*) unfolding times are compared with the nonparametric density curves of the theoretical densities, $\phi_{1:3}(t)$ (*a*), $\phi_{2:3}(t)$ (*b*), and $\phi_{3:3}(t)$ (*c*).

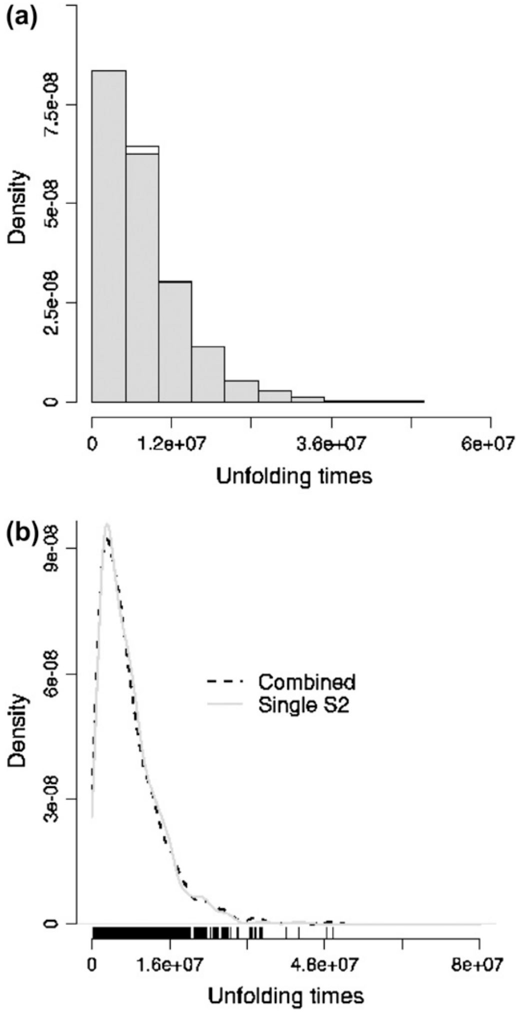

FIGURE 9 (*a*) Comparison of the histograms of forced unfolding times (in units of integration steps) for single *S*2 domain (*white bars*) and for combined data set for tandem *S*2–*S*2–*S*2 (*gray bars*), obtained by using Eq. 6. (*b*) Nonparametric density estimates of the theoretical probability density for single *S*2 domain (*solid curve*) and for combined data set (*dashed curve*).

larger than the corresponding ML estimates of the same quantities for single *S*2 and *S*1 domains.

DISCUSSION

Summary of main results

Conformational transitions in wild-type protein tandems and engineered polyproteins can be studied in force-clamp AFM experiments by employing constant stretching force and recording the unfolding times of individual domains (16–19). Up to now statistical analyses of forced unfolding times for homogeneous tandems ($(D)_N$) relied on the assumption that the recorded unfolding times form a set of independent identically distributed (iid) random variables with identical exponential parent pdfs for each domain *D*, i.e., $\psi_1(t) = \psi_2(t) = \ldots$ $\psi_N(t) = \psi(t)$. However, when the forced unfolding kinetics

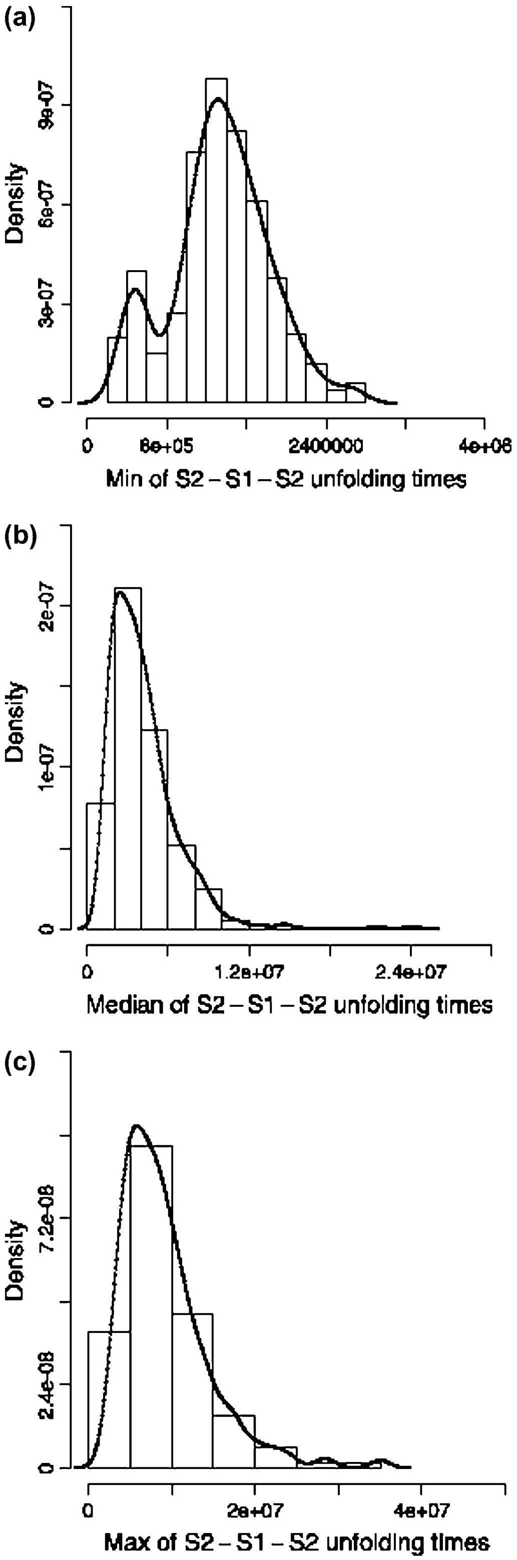

FIGURE 10 Histograms of ordered unfolding times (in units of integration steps) for tandem $S2$–$S1$–$S2$ obtained at $f = 66$ pN. Histograms of the minimum ($\{t_{1:3}\}$, *a*), median ($\{t_{2:3}\}$, *b*), and maximum ($\{t_{3:3}\}$, *c*) unfolding times are compared with the nonparametric density curves of the theoretical densities, $\phi_{1:3}(t)$ (*a*), $\phi_{2:3}(t)$ (*b*), and $\phi_{3:3}(t)$ (*c*).

competes with the dynamics of tension propagation along the tandem chain and/or when protein unfolding involves the formation of intermediate species, this assumption does not hold. To this day, no theoretical approaches have been developed for analyzing forced unraveling of heterogeneous tandems of nonidentical domains, D_1–D_2–. . .–D_N. For these tandems, forced unfolding times of nonidentical domains are expected to be nonidentically distributed, i.e., $\psi_1(t) \neq \psi_2(t) \neq \ldots \neq \psi_N(t)$.

The main goal of this work is to introduce a novel data analysis of unfolding data for protein tandems suited for studying both homogeneous as well as heterogeneous tandems, characterized by independent (uncorrelated) forced unfolding times. The presented approach is based on order statistics, i.e., statistical analysis of the ordered first, second, third, etc. unfolding times. This is quite natural since in force-clamp AFM probes the recorded unfolding times are ordered, i.e., $t_1 < t_2 < \ldots < t_N$. The methodology can be used to describe unfolding times characterized by an arbitrary parent pdf, $\psi(t)$. To introduce the reader to order statistics and to build the foundation for future extensions of the methodology to describing interdomain interactions (3), we focused on independent unfolding times.

We employed computer models of the homogeneous tandem ($S2$–$S2$–$S2$), and the heterogeneous tandem ($S2$–$S1$–$S2$) to exemplify application of the formalism. Langevin simulations were used to generate the forced unfolding trajectories of $S2$–$S2$–$S2$ and $S2$–$S1$–$S2$ at constant force $f = 66$ pN (Fig. 4). Preliminary analysis of unfolding times for single $S2$ and $S1$ domains and the same domains in the tandems, using nonparametric density estimation (Fig. 2) and quantile-quantile (or Q-Q) plots (Fig. 3), revealed that at $f = 66$ pN the unfolding transitions in $S2$–$S2$–$S2$ and $S2$–$S1$–$S2$ are described by independent identically distributed (iid) (Table 2; Figs. 5 and 6) and independent nonidentically distributed (inid) unfolding times (Table 2, Figs. 7 and 8), respectively.

Using a recurrence property for iid unfolding times (Eq. 6), we reconstructed the parent pdf $\psi_{S2}(t)$ for the $S2$ domain (Fig. 9). This is statistically equivalent to pooling all ordered unfolding time variates into a single data set and binning the obtained set into a histogram of unfolding times, which is how current analyses of unfolding time data are carried out. However, this procedure can be used only when the unfolding times are iid random variables. We modeled the $S2$ parent pdf using the Gamma ansatz (Eq. 14) with shape α_{S2}, which determines the position of the peak of the density (the most probable unfolding time), and unfolding rate k_{S2}, which quantifies the average unfolding time. The maximum likelihood (ML) estimates of these parameters for the single $S2$ domain (Table 1) are in agreement with the estimates of the same parameters for the $S2$ domain in the $S2$–$S2$–$S2$ tandem, obtained via fitting the order statistics pdfs (Fig. 9).

To illustrate the use of order statistics for describing inid forced unfolding times, we fitted numerically the theoretical order statistics pdfs (Eqs. 10 and 16) to the histograms of

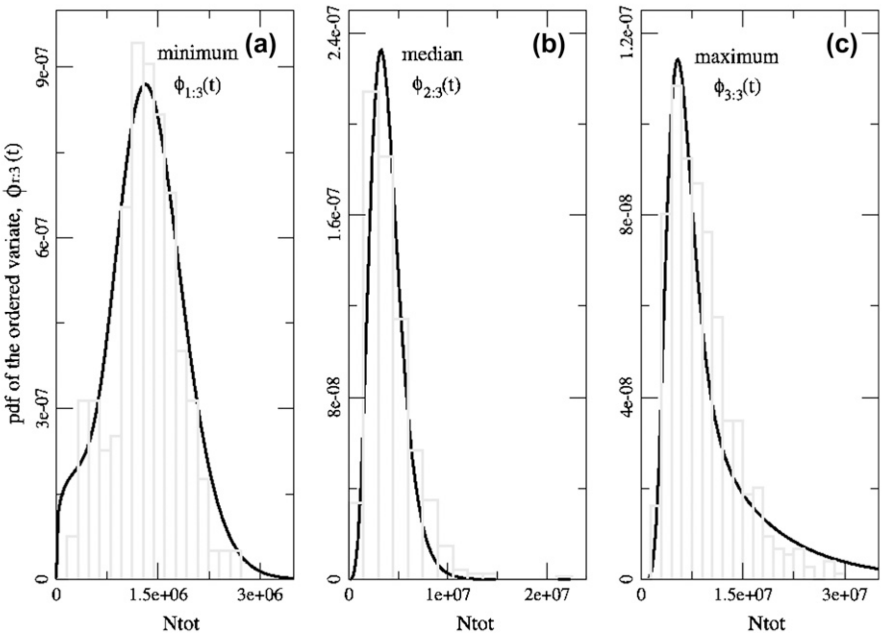

FIGURE 11 The fit of the histograms for the minimum $\{t_{1:3}\}$ (*a*), median $\{t_{2:3}\}$ (*b*), and maximum $\{t_{3:3}\}$ (*c*) unfolding times (in units of integration steps) for tandem *S*2–*S*1–*S*2 (*bars*). Theoretical order statistics pdfs $\phi_{r:3}(t)$ (r = 1, 2, 3, (see Eq. 16) are overlaid.

ordered unfolding times (Fig. 10) for the *S*2–*S*1–*S*2 tandem assuming the Gamma ansatz for the parent pdfs for domains *S*2 and *S*1 (Eq. 14). The results are presented in Fig. 11. The obtained shape parameters, α_{S2} and α_{S1}, and unfolding rates, k_{S2} and k_{S1} (Table 3), agree well with the ML estimates of these quantities for the single *S*2 and *S*1 domains (Table 1). However, due to flexible linkers, the values of α_{S2} (1.55, 2.05, and 1.75, Table 3) and α_{S1} (7.45, 4.95, Table 3), obtained for the *S*2–*S*1–*S*2 tandem, are slightly larger than the corresponding ML estimates of these quantities for single domains (0.96 for *S*2 and 4.92 for *S*1, Table 1). These findings imply that statistical analyses and theoretical modeling of unfolding data on wild-type protein tandems and polyproteins must take into account the effect of flexible linkers as their presence results in prolonged forced unfolding times.

Dependent forced unfolding times

In a separate set of Langevin simulations of forced unfolding for tandems *S*2–*S*2–*S*2 and *S*2–*S*1–*S*2, we increased the applied constant force from f = 66 pN to f = 88 pN to explore the effect of stretching force on the distribution and the dependence structure of the unfolding times. For each tandem, 500 unfolding trajectories were generated. The histograms of forced unfolding times for the first ($S2_1$), second ($S2_2$), and third ($S2_3$) *S*2 domain of tandem *S*2–*S*2–*S*2, obtained at f = 88 pN, are displayed in Fig. 12. These can be directly compared with the corresponding histograms and density estimates obtained for the same tandem at f = 66 pN (Fig. 6). As evidenced by both histograms and kernel density estimates as well as *Q*-*Q* plots, the nonidentical unfolding time distributions (Fig. 12) for individual *S*2 domains obtained at higher force f = 88 pN, are starkly different from the corresponding identical distributions for the same domains obtained at f = 66 pN (Figs. 5 and 6).

TABLE 3 Numerical values of α and k (in units of integration steps) for domains *S*2 and *S*1 in tandem *S*2–*S*1–*S*2, obtained from the fit of the order statistics pdfs $\phi_{r:3}$, r = 1, 2, 3, to the histograms of the ordered forced unfolding times $\{t_{1:3}\}$, $\{t_{2:3}\}$, and $\{t_{3:3}\}$

Order statistics pdf	α_{S2}	α_{S1}	k_{S2}	k_{S1}
$\phi_{1:3}$	1.55	7.45	3.04×10^{-7}	0.92×10^{-5}
$\phi_{2:3}$	2.05	4.95	3.82×10^{-7}	1.92×10^{-5}
$\phi_{3:3}$	1.75	4.85	2.85×10^{-7}	1.56×10^{-5}

Aside from the overall shift toward shorter unfolding times, due to Δf = 22 pN force increase, we witness a drastic change in the unfolding pattern. Specifically, at f = 88 pN the unfolding time distribution for the first *S*2 domain ($S2_1$) is bimodal with the first peak at shorter times ($\approx 0.4 \times 10^6$) and the second peak at longer times ($\approx 1.3 \times 10^6$). The unfolding time distributions for the second ($S2_2$) and third domain ($S2_3$) are peaked around 2.0×10^6 and 1.3×10^6, respectively (Fig. 12). These results imply that at f = 88 pN, on average, the forced unfolding of tandem *S*2–*S*2–*S*2 proceeds through initial unraveling of the terminal *S*2 domains ($S2_1$ first and $S2_3$ second), followed by the unfolding of the middle domain ($S2_2$). In contrast, the identical unfolding time distributions for all three *S*2 domains, obtained at f = 66 pN, are peaked at about the same time point $t \approx 4.0 \times 10^6$ (Fig. 6), which implies that at lower force any of the three *S*2 domains in the tandem can unfold first, second, or last.

We also observed that increased force induces dependence between the unfolding transitions in *S*2 domains. Whereas at f = 66 pN the individual unfolding times for all three domains in the *S*2–*S*2–*S*2 and *S*2–*S*1–*S*2 tandems were found

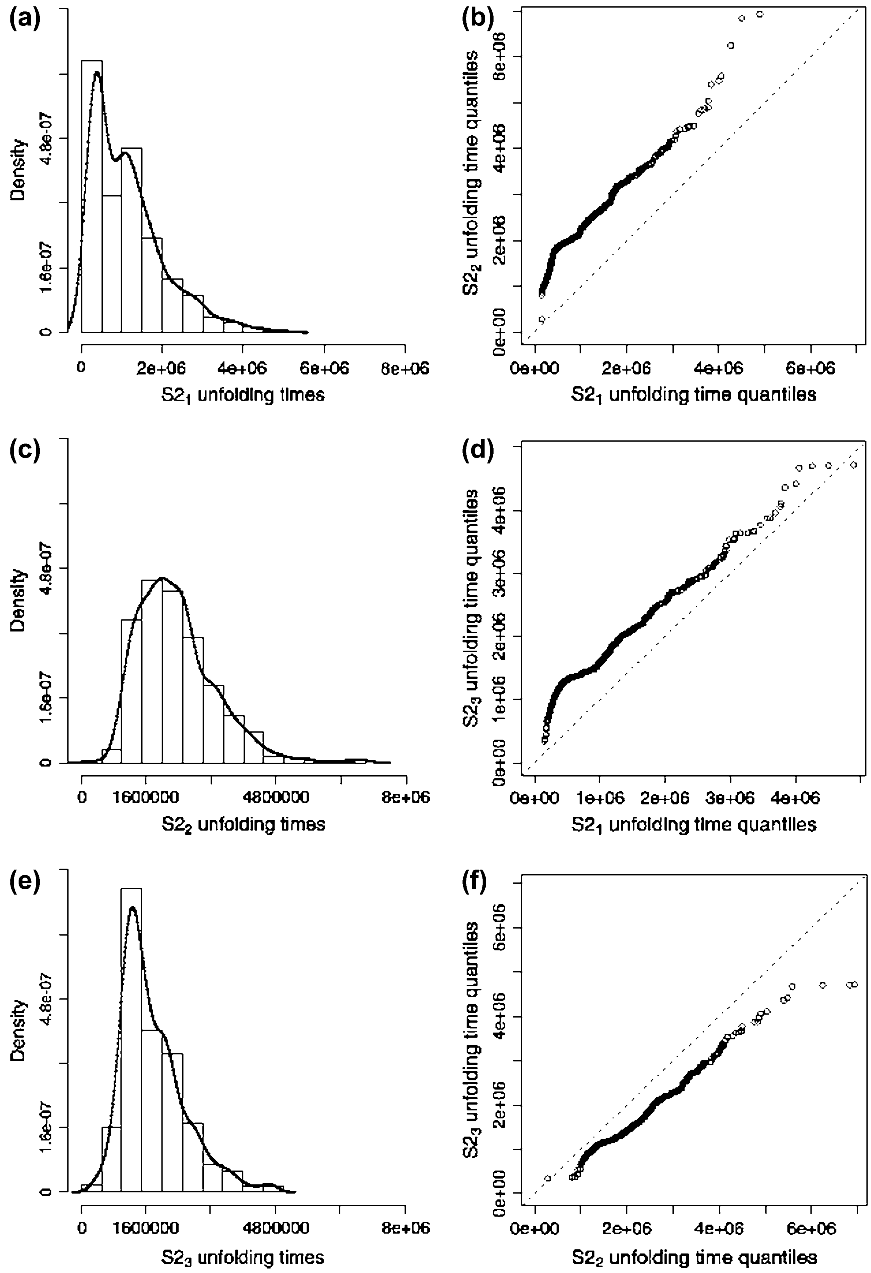

FIGURE 12 (*a*, *c*, and *e*) Histograms of forced unfolding times (*bars*) and nonparametric density estimates of the theoretical probability densities (*curve*) for domains $S2_1$ (*a*), $S2_2$ (*c*), and $S2_3$ (*e*) for tandem $S2$–$S2$–$S2$ at $f = 88$ pN. (*b*, *d*, and *f*) *Q-Q* plots of forced unfolding times (t_1, t_2) (*b*), (t_1, t_3) (*d*), and (t_2, t_3) (*f*) for domains $S2_1$ (t_1), $S2_2$ (t_2), and $S2_3$ (t_3).

to be pairwise independent (Table 2), this is no longer the case at $f = 88$ pN. Table 4 reports the Spearman rank correlation coefficients for three pairs of unfolding times (t_2, t_1), (t_3, t_1), and (t_3, t_2), in tandems $S2$–$S2$–$S2$ and $S2$–$S1$–$S2$ recorded at $f = 88$ pN. For the homogeneous tandem $S2$–$S2$–$S2$, unfolding times for domains $S2_1$ (t_1) and $S2_3$ (t_3) have a Spearman rank correlation of -0.17 indicating statistically significant dependence (correlation) with a p-value of 0; similarly, unfolding times for domains $S2_2$ (t_2) and $S2_3$ (t_3) have a Spearman rank correlation of -0.19 that is highly statistically significant with a p-value of 0 (Table 4). Yet, unfolding times for domains $S2_1$ (t_1) and $S2_2$ (t_2) have zero Spearman rank correlation coefficient with a high p-value of 0.95, which indicates statistically significant independence of the forced unfolding times between domains $S2_1$ and $S2_2$ (Table 4). These findings are particularly striking given that the domains $S2_1$ and $S2_3$ are spacially separated by the middle domain $S2_2$, whereas $S2_1$ and $S2_2$ domains are nearest neighbors. Similar analysis for the heterogeneous tandem $S2$–$S1$–$S2$ revealed statistically significant dependence between the forced unfolding times for the first $S2_1$ and the second domain $S1_2$ ($R = -0.16$ and associated 0 p-value, Table 4), and the second $S1_2$ and the third $S2_3$ domain ($R = -0.23$ and associated 0 p-value, Table 4).

The observed pairwise dependence of unfolding events is not unexpected. Indeed, when the timescale of tension

TABLE 4 Spearman rank correlation coefficients for the unfolding times for domains $S2_1$ (t_1), $S2_2$ (t_2), and $S2_3$ (t_3) in tandem $S2$–$S2$–$S2$, and for domains $S2_1$ (t_1), $S1_2$ (t_2), and $S2_3$ (t_3) in tandem $S2$–$S1$–$S2$ obtained at $f = 88$ pN

	*S*2-*S*2-*S*2				*S*2-*S*1-*S*2		
Time	t_1	t_2	t_3	Time	t_1	t_2	t_3
t_1	1	0.0 (0.95)	−.17 (0.00)	t_1	1	−.16 (0.00)	0.01 (0.83)
t_2	0.00 (0.95)	1	−0.19 (0.00)	t_2	−0.16 (0.00)	1	−0.23 (0.00)
t_3	−0.17 (0.00)	−0.19 (0.00)	1	t_3	0.01 (0.83)	−0.23 (0.00)	1

The p-values for testing $R = 0$ are in parentheses.

propagation along the tandem chain is comparable with the average forced unfolding time of domains $S2$ and $S1$, the recorded forced unfolding times $\{t_i\}$ and $\{t_j\}$ ($i \neq j = 1, 2, 3$) are separated by ''waiting periods,'' $\{\Delta t_{ij}\}$ ($\Delta t_{ij} = t_i - t_j$), that are not long enough to decorrelate the consecutive unfolding transitions. Uncorrelated at lower force unfolding transitions of individual domains in the tandem may thus become correlated (dependent) at elevated force levels. Even for homogeneous tandems, such as $S2$–$S2$–$S2$, whether the unfolding times of identical repeats are identically distributed depends on the magnitude of applied force. By comparing the unfolding time pdfs for $S2$ domains obtained at $f = 88$ pN (Fig. 12) with the corresponding pdfs obtained at $f = 66$ pN (Figs. 5 and 6), we see that at higher force unraveling of the domains closer to the point of force application (terminal domains $S2_1$ and $S2_3$) is characterized by shorter unfolding times compared with those for the more remote middle $S2_2$ domain.

It is likely that forced unfolding times of wild-type protein tandems, analyzed in force-clamp AFM experiments, also involve a complicated force-induced dependence structure. Yet, current approaches to statistical analysis of unfolding times rely on the simplifying assumption that the recorded unfolding times form a set of iid random variables. Because applied stretching force may couple the unfolding transitions of the individual tandem domains, novel theory and new approaches to statistical data analysis of correlated forced unfolding times that do not use the iid assumption are much needed.

CONCLUSIONS

In this article we introduced and tested an order statistics based methodology for analyzing the forced unfolding times of protein tandems. We used this approach to analyze computer simulated unfolding transitions, characterized by independent identically distributed (iid) and independent nonidentically distributed (inid) unfolding times. The order statistics based methodology presented here can now be used in force-clamp AFM experiments to study force-induced unfolding in both homogeneous and heterogeneous protein tandems.

An intriguing finding of this work is that the iid assumption, currently used in statistical analyses of forced unfolding data, that the unfolding transitions in the protein tandems of identical repeats can be characterized by independent identically distributed unfolding times may not hold at elevated piconewton levels of applied force. We showed that a moderate increase in the magnitude of applied force may result in correlated unfolding transitions, characterized by dependent unfolding times. In the case of tandems $S2$–$S2$–$S2$ and $S2$–$S1$–$S2$ formed by noninteracting domains $S2$ and/or $S1$, correlated unfolding events are due to competition between the unfolding kinetics and the dynamics of force propagation along the tandem chain, which couples consecutive unfolding transitions. The dependence among unfolding times may also be mediated by the interdomain interactions. For example, heterogeneous tandems of $I27$–$I28$ repeats show enhanced domain stabilization against applied force, which results in the increase of the average unfolding force from $f \approx 260$ pN (for the tandem of repeated Ig domains $I27$ of titin) to $f \approx 300$ pN (for $I27$–$I28$ repeats) (3). Similar domain stabilization effect has been observed in heterogeneous tandems of $Fn3$ domains (44).

These experimental findings, in unison with our theoretical results, necessitate extending the statistical tools for independent forced unfolding times, presented in this article, to accommodate correlated unfolding transitions. Depending on the tandem composition, correlated unfolding transitions are described by dependent identically distributed (did) unfolding times (homogeneous tandems) and dependent nonidentically distributed (dnid) unfolding times (heterogeneous tandems). For the convenience of the reader, the possible types of unfolding transitions are classified in Table 5. The development of order statistics based methodology for describing correlated unfolding data will enable researchers to accurately probe protein-protein interactions. Order statistics type methodology can also be formulated for analyzing forced unfolding data available from force-ramp AFM measurements.

To take full advantage of the presented methodology, one needs to be able to determine: a), (in)dependence of forced unfolding times, and b), (in)equality of the (parent) unfolding time distributions of individual proteins in a tandem under study. Both simple empirical tools and rigorous statistical tests for assessing the (in)dependence of unfolding times and (in)equality of the (parent) distributions are needed. Because in force-clamp AFM experiments, ordered unfolding times are the only observable quantities, tests for deducing the dependence structure among the original but otherwise

TABLE 5 Classification of the forced unfolding times based on the tandem composition and the presence/absence of correlations between unfolding of individual domains

Tandem composition	Uncorrelated unfolding events	Correlated unfolding events
Homogeneous tandem	iid unfolding times	did unfolding times
Heterogeneous tandem	inid unfolding times	dnid unfolding times

unobservable unfolding times and (in)equality of the unfolding time distributions have to be based on the observed ordered data. Development of such statistical tests along with methodology for analyzing correlated forced unfolding data is in progress (45).

APPENDIX I: ORDER STATISTICS FOR IID UNFOLDING TIMES

Suppose $\{T_1, T_2, \ldots, T_n\}$ are n random variables with distribution function, $\Psi(t_1, \ldots, t_n) = \text{Prob}[T_1 \leq t_1, \ldots, T_n \leq t_n]$ (uppercase letters represent random variables whereas lowercase letters represent their values). If they are arranged in increasing order, $T_{1:n} \leq T_{2:n} \leq \ldots \leq T_{n:n}$, then $T_{r:n}$—the r-th order statistic—is equal to the r-th smallest value. When the random variables T_i, $i = 1, \ldots, n$, are independent with common cumulative distribution function (cdf), $\Psi(t) = \text{Prob}[T_r \leq t]$, their joint cdf is the product of their common marginals, $\Psi(t_1, \ldots, t_n) = \prod_{i=1}^{n} \text{Prob}[T_i \leq t_i] = \prod_{i=1}^{n} \Psi(t_i)$. Let $\Phi_{r:n}(t)$ denote the cdf of the r-th order statistic $T_{r:n}$, $r = 1, 2, \ldots, n$. Then,

$$\begin{aligned}\Phi_{r:n}(t) &= \text{Prob}(\text{at least } r \text{ of the } T'\text{s are} \leq t) \\ &= \sum_{i=r}^{n} \binom{n}{i} \text{Prob}[T_1 \leq t]^{i} (1 - \text{Prob}[T_1 \leq t])^{n-i} \\ &= \sum_{i=r}^{n} \binom{n}{i} \Psi(t)^{i} (1 - \Psi(t))^{n-i}. \qquad \text{(A1)}\end{aligned}$$

The pdf of $T_{r:n}$, $\phi_{r:n}(t)$, is obtained by differentiating Eq. A1,

$$\phi_{r:n}(t) \equiv \frac{d\Phi_{r:n}(t)}{dt} = n\psi(t) \binom{n-1}{r-1} \Psi(t)^{r-1} (1 - \Psi(t))^{n-r}. \qquad \text{(A2)}$$

Equations A1 and A2 are the same as Eq. 3 for the r-th order statistic given in the main text.

In many applications, unfolding times of proteins are modeled as iid exponential random variables, and the unfolding times are characterized by the rate of unfolding K. Then, $\Psi_{\exp}(t) = 1 - \exp[-Kt]$, $\psi_{\exp}(t) = K \exp[-Kt]$, and

$$\begin{aligned}\Phi^{\exp}_{r:n}(t) &= \sum_{m=r}^{n} \binom{n}{m} (1 - e^{-Kt})^{m} e^{-(n-m)Kt} \\ \phi^{\exp}_{r:n}(t) &= n \binom{n-1}{r-1} (1 - e^{-Kt})^{r-1} e^{-(n-r)Kt} K e^{-Kt}. \qquad \text{(A3)}\end{aligned}$$

The minimum and maximum order statistics pdfs are given by $\phi^{\exp}_{1:n}(t) = n \exp[-Kt(n-1)]K \exp[-Kt]$ and $\phi^{\exp}_{n:n}(t) = n(1 - \exp[-Kt])^{n-1} K \exp[-Kt]$. The minimum unfolding times with exponential parent pdf are also exponentially distributed, i.e. $\phi^{\exp}_{1:n}(t) = nK \exp[-nKt]$, with the average unfolding time, $\mu_{1:n} = 1/nK$. This implies that the average first unfolding time in the tandem of length n, $\mu_{1:n}$, is exactly n times shorter than the average first unfolding time in the tandem of unit length ($n = 1$), $\mu_{1:1}$, i.e. $\mu_{1:n} = \mu_{1:1}/n$.

APPENDIX II: PERMANENTS

Let S_n denote the set of permutations of 1, 2, …, n. The permanent of the $n \times n$ matrix A is defined as

$$\text{per}A = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i\sigma(i)}.$$

The permanent remains unchanged if the rows or columns of the matrix are permuted and admits a Laplace expansion along any row or column of the matrix. That is, if we denote by $A(i, j)$ the matrix resulting by deleting row i and column j of matrix A, then

$$\text{per}A = \sum_{j=1}^{n} a_{ij}\, \text{per}A(i,j), \quad i = 1, \ldots, n, \qquad \text{(B1)}$$

$$\text{per}A = \sum_{i=1}^{n} a_{ij}\, \text{per}A(i,j), \quad j = 1, \ldots, n. \qquad \text{(B2)}$$

For example, if $A = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 2 & 1 \\ 1 & 0 & 0 \end{bmatrix}$, application of Eq. B1 yields

$$\begin{aligned}\text{per}A &= 2 \times \text{per}A(1,1) - 1 \times \text{per}A(1,2) + 0 \times \text{per}A(1,3) \\ &= 2(2 \times 0 + 0 \times 1) - 1(0 \times 0 + 1 \times 1) + 0(0 \times 0 + 1 \times 2) \\ &= -1. \qquad \text{(B3)}\end{aligned}$$

APPENDIX III: HISTOGRAMS AND KERNEL DENSITY ESTIMATORS

A histogram estimate of a probability density function $g(t)$ is defined to be

$$\hat{g}_h(t) = \frac{\hat{G}(t+h) - \hat{G}(t)}{h} = \frac{1}{nh} \sum_{i=1}^{n} 1\{t < t_i \leq t + h\}, \qquad \text{(C1)}$$

where h is the bandwidth (bin size), $1(A)$ is the indicator of the set A, and $\hat{G}(t) = 1/n \sum_{i=1}^{n} 1\{t_i \leq t\}$ is the empirical distribution function of the random variable T. In the histogram estimate of g, given by Eq. C1, all points in a bin are assigned equal weights. The idea behind kernel density estimation is that data closer to the point where the density is estimated should be given larger weight. The corresponding expression for a kernel density estimate of g, $\hat{g}_k$, is given by

$$\hat{g}_k(t) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{t - t_i}{h}\right), \qquad \text{(C2)}$$

where K is a symmetric kernel function ($K(t) = K(-t)$), satisfying $\int K(t)dt = 1$ with finite second moment. These conditions are imposed so that the estimate has desirable statistical properties, such as $\text{E}(\hat{g}_k(t)) = g(t) + O(h^2)$ (38,39). In this article, the Gaussian kernel, $K(t) = \exp(-t^2/2)/\sqrt{2\pi}$, was used to obtain the nonparametric density estimates.

APPENDIX IV: MAXIMUM LIKELIHOOD ESTIMATION

Suppose $T_1, T_2, \ldots, T_n$ are iid with density $\psi(x;\theta)$, where θ is a parameter vector. For example, if $\psi(t)$ is exponential, $\theta = \kappa$; if it is the Gamma density, $\theta = (\alpha, k)$ (Eq. 14). The likelihood function, $L(\theta) = \psi(t_1, t_2, \ldots, t_n|\theta)$, is the probability density, $\psi(t_1, t_2, \ldots, t_n|\theta)$, when considered as a function of θ for fixed data $t_1, t_2, \ldots, t_n$. The maximum likelihood estimator of θ is the value of θ that maximizes $L(\theta)$. For the Gamma density, the likelihood function is $L = \prod_{i=1}^{n} \psi(t_i|\alpha, k)$ with log-likelihood, $\log L = (\alpha - 1) \sum_{i=1}^{n} \log(t_i) - \sum_{i=1}^{n} kt_i + n\alpha \log(k) - n \log\Gamma(\alpha)$. Differentiating with respect to k and setting it equal to zero yields $\hat{k} = 1/\sum t_i \alpha n$. There is no closed-form solution for α, but a numerical solution can be obtained. We used the optimization algorithm in *R* (40) to find the ML estimates of α and k.

We thank two anonymous referees for valuable suggestions, which improved this article significantly.

This work was supported by National Science Foundation grant DMS-0204563 (E. Bura), and a start-up fund from the University of Massachusetts Lowell (V. Barsegov).

REFERENCES

1. Labeit, S., M. Gautel, A. Lakey, and J. Trinick. 1992. Towards a molecular understanding of titin. *EMBO J.* 11:1711–1716.
2. Trinick, J., P. Knight, and A. Whiting. 1984. Purification and properties of native titin. *J. Mol. Biol.* 180:331–356.
3. Li, H. B., A. F. Oberhauser, S. B. Fowler, J. Clarke, and J. M. Fernandez. 2000. Atomic force microscopy reveals the mechanical design of a modular protein. *Proc. Natl. Acad. Sci. USA.* 97:6527–6531.
4. Carrion-Vazquez, M., A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Proedel, J. Clarke, and J. M. Fernandez. 1999. Mechanical and chemical unfolding of a single protein: a comparison. *Proc. Natl. Acad. Sci. USA.* 96:3694–3699.
5. Stossel, T. P., J. Condeelis, L. Cooley, J. H. Hartwig, A. Noegel, M. Schleicher, and S. S. Shapiro. 2001. Filamins as integrators of cell mechanics and signalling. *Nat. Rev. Mol. Cell Biol.* 2:138–145.
6. Feng, Y., and C. A. Walsh. 2004. The many faces of filamin: a versatile molecular scaffold for cell motility and signalling. *Nat. Cell Biol.* 6: 1034–1038.
7. Schwaiger, I., A. Kardinal, M. Schleicher, A. A. Niegel, and M. Rief. 2004. A mechanical unfolding intermediate in an actin-crosslinking protein. *Nat. Struct. Biol.* 11:81–85.
8. Schwaiger, I., M. Schleicher, A. A. Niegel, and M. Rief. 2005. The folding pathway of a fast-folding immunoglobulin domain revealed by single-molecule mechanical experiments. *EMBO Rep.* 6:46–51.
9. Popowicz, G. M., R. Muller, A. A. Noegel, M. Schleicher, R. Huber, and T. A. Holak. 2004. Molecular structure of the rod domain of *Dictyostelium* filamin. *J. Mol. Biol.* 342:1637–1646.
10. Schwarzbauer, J. E., and J. L. Sechler. 1999. Fibronectin fibrillogenesis: a paradigm for extracellular matrix assembly. *Curr. Opin. Struct. Biol.* 11:622–627.
11. Pickard, C. M. 2001. Mechanisms underlying ubiquitination. *Annu. Rev. Biochem.* 70:503–533.
12. Weissman, A. M. 2001. Themes and variations on ubiquitylation. *Nat. Rev. Mol. Cell Biol.* 2:169–178.
13. Rief, M., M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub. 1997. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science.* 276:1109–1112.
14. Zinober, R. C., D. J. Brockwell, G. S. Beddard, A. W. Blake, P. D. Olmsted, S. E. Radford, and D. A. Smith. 2002. Mechanically unfolding proteins: the effect of unfolding history and the supramolecular scaffold. *Protein Sci.* 11:2759–2765.
15. Steward, A., J. L. Toca-Herrera, and J. Clarke. 2002. Versatile cloning system for construction of multimeric proteins for use in atomic force microscopy. *Protein Sci.* 11:2179–2183.
16. Brujić, J., R. I. Z. Hermans, K. A. Walther and. J. M. Fernandez. 2006. Single-molecule force spectroscopy reveals signatures of glassy dynamics in the energy landscape of ubiquitin. *Nature Physics.* 2:282–286.
17. Oberhauser, A. F., P. K. Hansma, M. Carrion-Vazquez, and J. M. Fernandez. 2001. Stepwise unfolding of titin under force-clamp atomic force microscopy. *Proc. Natl. Acad. Sci. USA.* 98:468–472.
18. 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.
19. Fernandez, J. M., and H. Li. 2004. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science.* 303:1674–1678.
20. Zhang, B., G. Xu, and J. S. Evans. 1999. A kinetic molecular model of the reversible unfolding and refolding of titin under force extension. *Biophys. J.* 77:1306–1315.
21. Hummer, G., and A. Szabo. 2003. Kinetics from nonequilibrium single-molecule pulling experiments. *Biophys. J.* 85:5–15.
22. Brujić, J., R. I. Z. Hermans, S. Garcia-Manyes, K. A. Walther, and J. M. Fernandez. 2007. Dwell-time distribution analysis of polyprotein unfolding using force-clamp spectroscopy. *Biophys. J.* 92: 2896–2903.
23. David, H. A., and H. N. Nagaraja. 2003. Order Statistics. Wiley Interscience, New York.
24. Gumbel, E. J. 2004. Statistics of Extremes. Dover Publications, New York.
25. Beckmann, P. 1967. Probability in Communication Engineering. Harcourt, Brace & World, New York.
26. Singpurwalla, N. D., and S. P. Wilson. 1999. Statistical Methods in Software Engineering: Reliability and Risk. Springer Verlag, New York.
27. Tees, D. F. J., J. T. Woodward, and D. A. Hammer. 2001. Reliability theory for receptor-ligand bond dissociation. *J. Chem. Phys.* 114:7483–7496.
28. Klimov, D. K., and D. Thirumalai. 2000. Native topology determines force-induced unfolding pathways in globular proteins. *Proc. Natl. Acad. Sci. USA.* 97:7254–7259.
29. Arnold, B. C., and N. Balakrishnan. 1989. Relations, bounds and approximations for order statistics. *In* Lecture Notes in Statistics. 53. Springer-Verlag, New York.
30. Abramowitz, M., and I. A. Stegun. 1972. Handbook of Mathematical Functions. Dover Publications, New York.
31. Vaughan, R. J., and W. N. Venables. 1972. Permanent expressions for order statistics densities. *J. R. Statist. Soc. B.* 34:308–310.
32. Minc, H. 1983. Theory of permanents 1978–1981. *Lin. Mult. Algebra.* 12:227–263.
33. Minc, H. 1983. Theory of permanents 1982–1985. *Lin. Mult. Algebra.* 21:109–148.
34. Sathe, Y. S., and U. J. Dixit. 1990. On a recurrence relation for order statistics. *Stat. Probab. Lett.* 9:1–4.
35. Onuchic, J. N., and P. G. Wolynes. 2004. Theory of protein folding. *Curr. Opin. Struct. Biol.* 14:70–75.
36. Dill, K. A., S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. 1995. Principles of protein folding: a perspective from simple exact models. *Protein Sci.* 4:561–602.
37. Veitshans, T., D. K. Klimov, and D. Thirumalai. 1997. Protein folding kinetics: time scales, pathways, and energy landscapes in terms of sequence dependent properties. *Fold. Des.* 2:1–22.
38. Silverman, B. W. 1986. Density Estimation. Chapman and Hall, London, UK.
39. Scott, D. W. 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, New York.
40. R Development Core Team. 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (http://www.R-project.org).
41. Davison, A. C. 2003. Statistical Models. Cambridge University Press, Cambridge, UK.
42. Gibbons, J. D., and S. Chakraborti. 2003. Nonparametric Statistical Inference, 4th Ed. Marcel Dekker, New York.
43. Kendall, M. G., and J. D. Gibbons. 1990. Rank Correlation Methods, 5th Ed. Edward Arnold, London, UK.
44. Oberhauser, A. F., C. Badilla-Fernandez, M. Carrion-Vazquez, and J. M. Fernandez. 2002. The mechanical hierarchies of fibronectin observed with single-molecule AFM. *J. Mol. Biol.* 319:433–447.
45. Bura, E., D. K. Klimov, and V. Barsegov. Analyzing forced unfolding of protein tandems by ordered variates. 2. Dependent unfolding times. In press.