THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Statistical Learning from Single-Molecule Experiments: Support Vector Machines and Expectation−Maximization Approaches to Understanding Protein Unfolding Data

Published as part of The Journal of Physical Chemistry virtual special issue "Dave Thirumalai Festschrift".

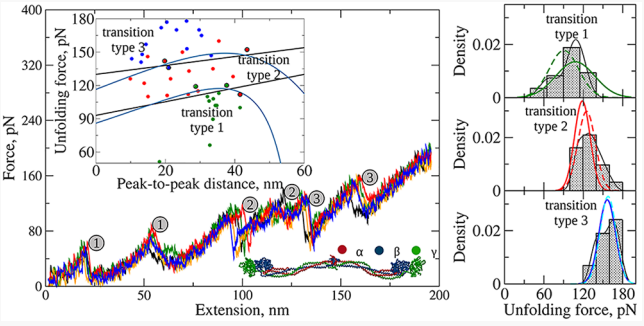
Farkhad Maksudov, Lee K. Jones, and Valeri Barsegov*

Cite This: *J. Phys. Chem. B* 2021, 125, 5794−5808

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Single-molecule force spectroscopy has become a powerful tool for the exploration of dynamic processes that involve proteins; yet, meaningful interpretation of the experimental data remains challenging. Owing to low signal-to-noise ratio, experimental force-extension spectra contain force signals due to nonspecific interactions, tip or substrate detachment, and protein desorption. Unravelling of complex protein structures results in the unfolding transitions of different types. Here, we test the performance of Support Vector Machines (SVM) and Expectation Maximization (EM) approaches in statistical learning from dynamic force experiments. When the output from molecular modeling *in silico* (or other studies) is used as a training set, SVM and EM can be applied to understand the unfolding force data. The maximal margin or maximum likelihood classifier can be used to separate experimental test observations into the unfolding transitions of different types, and EM optimization can then be utilized to resolve the statistics of unfolding forces: weights, average forces, and standard deviations. We designed an EM-based approach, which can be directly applied to the experimental data without data classification and division into training and test observations. This approach performs well even when the sample size is small and when the unfolding transitions are characterized by overlapping force ranges.



INTRODUCTION

Intra- and extracellular proteins use mechanical forces in diverse cellular processes, ranging from replication, transcription, and translation[1] to protein degradation,[2−4] to cytoskeleton support,[5,6] to cell adhesion and cell motility,[7,8] to formation of the extracellular matrix,[9] to muscle contraction and relaxation,[10−14] and to blood clotting.[15−20] Although considerable efforts have been expended by experimentalists and theoreticians to elucidate how proteins alter their shape and conformation in response to the external mechanical factors, understanding the interplay between the dynamics of proteins and their structural changes continues to be one of the major frontiers of research. Atomic Force Microscopy (AFM),[21−25] optical tweezers,[26−29] and magnetic tweezers[30−34] have been employed to access conformational transitions in proteins to mechanically unfold proteins and to rupture protein complexes. These experiments, in which the mechanical response of the protein is monitored by pulling a protein molecule with a constant force (force clamp) or constant force-loading rate (force ramp),[35] yield information about the dynamic transitions that occur on the nanometer length scale under the influence of pico-Newton forces.[35]

In single-protein assays, the one end of a protein molecule is tethered to a surface or microscopic tip while the other end interacts with the tip or a surface. In the force-ramp assays, the applied pulling force $f(t) = r_f t$ used to induce the mechanical unfolding reactions in proteins is linearly increasing in time t with the force-loading rate $r_f = \kappa v_f$ that depends on the pulling velocity v_f. Each unfolding transition in protein domains is accompanied by a tension drop in the polypeptide chain, and so the force-extension profile (FX curve) exhibits the characteristic sawtooth-like pattern. The FX curve displays multiple force maxima (peak forces), $f_1(X_1)$, $f_2(X_2)$, ..., $f_n(X_n)$, each peak marking the unfolding transition in a particular protein domain. Therefore, the force-extension spectra can be viewed as proteins' mechanical fingerprints in dynamic force spectroscopy experiments on proteins. For example, the peak-to-peak distances,

Received: March 16, 2021
Revised: May 16, 2021
Published: June 2, 2021

which correspond to the force-induced elongation of a polyprotein, carries information about the size and structure of unfolded domains. The peak forces provide information about the mechanical stability of protein domains and unfolding energies. The interpeak distances and peak forces can be combined to illuminate the details of free-energy landscape.[25]

Although single-molecule experiments are low throughput, their main advantage over the more traditional bulk measurements is that these experiments reveal the entire probability distributions of molecular characteristics, such as the distributions of unfolding forces, rather than the average quantities. This is important given that the protein folding and unfolding are stochastic processes. The single-molecule experiments have been used to provide insights into the mechanical stability and unfolding pathways for a wide range of proteins involved, *e.g.*, in force generation in molecular motors,[36,37] cell signaling,[38−40] formation of cell adhesion complexes,[41] and protein degradation.[2,42,43] Nevertheless, accurate interpretation of the single-molecule force experiments remains challenging. For example, it is difficult to distinguish the unfolding transitions from the "non-specific events" due to sample contamination and tip−surface interactions, contributions to the force signal from several unfolding transitions (rather than single transition) that occur simultaneously, *etc.* An attractive option is to use protein tandems of head-to-tail connected (identical or different) protein domains, $D_1 - D_2 - ... - D_n$, but there are challenges associated with the Order Statistics nature of the unfolding forces.[44−47]

To improve the statistical significance of the protein unfolding data, the force-extension spectra gathered together from different single-molecule force measurements are filtered to include only those spectra that (i) look appealing with minimal tip−surface interactions, (ii) show a large number of unfolding events (force peaks), and (iii) display the strong first and (or) last detachment peak(s) due to protein desorption from the substrate surface.[48] These multiple quality assessments are at most qualitative and subjective, which reduces the scope of potential information gain. Experimentalists gather many hundreds of the force-extension spectra, which contain thousands of data points (peak forces and peak-to-peak distances). However, due to data selection described above, nearly 90−95% of experimental data are discarded, and only 5−10% of data are analyzed, which creates a bias and introduces a human error. In recent years, supervised and unsupervised Statistical Learning (or Machine Learning) has emerged as a collection of powerful quantitative tools, both for interpretation and modeling of complex data sets.[49] In the past two decades, Statistical Learning has been increasingly more used in a variety of scientific disciplines, including biology,[50,51] material science,[52−55] and chemistry.[56,57]

Here, we explore the applicability of several powerful approaches to unsupervised and supervised learning to classify and characterize the experimental forced unfolding data from single-molecule force-ramp assays. We propose an approach, in which the results of mechanical testing experiments *in silico* (or other studies) are used as training data. We compare the performance of Support Vector Machines (SVM)- and Expectation−Maximization (EM)-based methods to understand complex unfolding force data for multidomain proteins and polyproteins. As a prototype of a polyprotein formed by connected protein repeats, we use a dimer $(WW)_2$ formed by the all-β-sheet domain WW.[58−61] As a model of large multidomain protein with complex structure, we use human fibrinogen (Fg).[62] To carry out various case studies, we use the output from dynamic force experiments *in silico* for the dimer $(WW)_2$ and for the Fg monomer, as well as the experimental forced unfolding data for the Fg monomer.[63] We show that SVM and EM can be used with success to understand the experimental forced unfolding data. Tests of performance and accuracy reveal that the SVM and EM approaches are suitable statistical learning tools for describing the experimental FX spectra even when the sample size is small. The developed SVM and EM approaches perform well even when protein unfolding data are complex, *i.e.*, characterized by multiple unfolding transitions of different types, and when the FX spectra are noisy. We also propose a simple EM-based approach to understanding the experimental forced unfolding data. This approach does not involve traditional data classification and data division into the training and test observations, and it can be applied even when the sample size is small and when the unfolding transitions of different types are characterized by overlapping force ranges. Taken together, the results obtained demonstrate that the SVM and EM method allow for accurate interpretation of the unfolding force data. Given their conceptual simplicity, the SVM and EM based approaches can be easily implemented in single-molecule experimental setting to model the single-protein unfolding data.

MATEARIALS AND METHODS

Molecular Modeling. *Coarse-Grained Models for $(WW)_2$ and Fg.* We used the C_α-based Self-Organized Polymer (SOP) models[64] for the dimer $(WW)_2$ formed by two WW domains (Figure 1) and for fibrinogen monomer (Figure 2). The WW domain (34 amino acids) was studied experimentally[58,59] and computationally[60,61] to describe folding and unfolding of the β-sheet proteins.[58,65] The dimer $(WW)_2$ is constructed by connecting the N- and C-termini of the adjacent WW domains using linkers of four neutral residues (Figure 1). Fibrinogen Fg (1925 residues) is a blood plasma protein, which consists of pairs of Aα chains, Bβ chains, and γ chains, linked by the disulfide bonds[66] (Figure 3b). The two distal globular regions and the central globular region of fibrinogen are connected by the α-helical coiled coils; each globular region at both ends of the molecule contains the β-nodule and the γ-nodule. The forced unfolding transitions in human fibrinogen molecule have studied both experimentally and theoretically.[63]

In the SOP models of $(WW)_2$ and Fg, each amino acid is represented by its C_α-atom with the $C_\alpha - C_\alpha$ bond distance of a = 3.8 Å, which corresponds to the length of a peptide bond. The potential energy for protein conformation V_{MOL} is given by[64,67,68]

$$\begin{aligned} V_{MOL} &= V_{FENE} + V_{NB}^{ATT} + V_{NB}^{REP} \\ &= -\sum_{i=1}^{Q-1} \frac{k}{2} {R_0}^2 \log\left(1 - \frac{(r_{i,i+1} - r_{i,i+1}^0)^2}{{R_0}^2}\right) \\ &\quad + \sum_{i=1}^{Q-3} \sum_{j=i+3}^{M} \varepsilon_h \left[\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^0}{r_{ij}}\right)^6\right] \Delta_{ij} \\ &\quad + \sum_{i=1}^{Q-2} \varepsilon_l \left(\frac{\sigma}{r_{i,i+2}}\right)^6 + \sum_{i=1}^{Q-3} \sum_{j=i+3}^{M} \varepsilon_l \left(\frac{\sigma}{r_{ij}}\right)^6 (1 - \Delta_{ij}) \end{aligned} \quad (1)$$

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

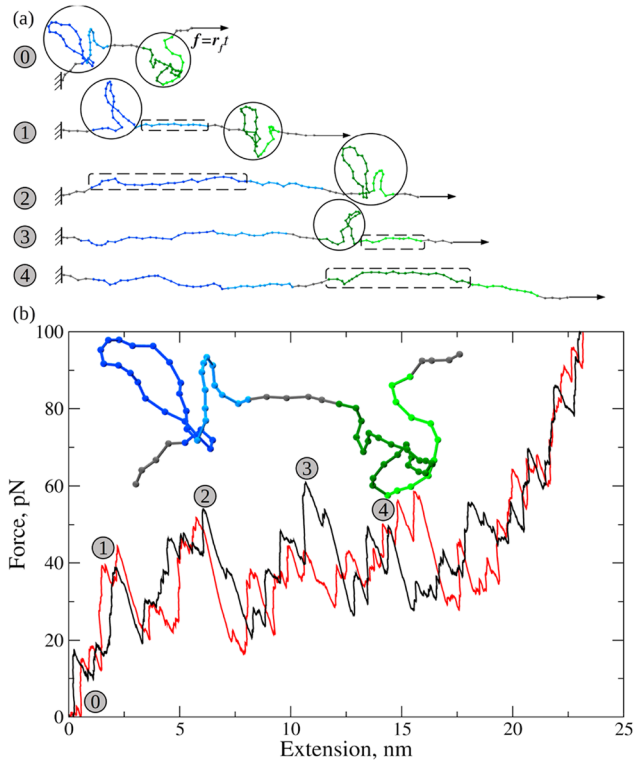


Figure 1. Forced unfolding transitions in dimer $(WW)_2$. Panel a: Schematic representation of $(WW)_2$ formed by two C-terminal-to-N-terminal connected WW domains (in blue and green colors) through flexible linkers (in gray) in the folded state (snapshot 0), partially unfolded conformations (snapshots 1−3), and unfolded state (snapshot 4) are shown. In mechanical testing *in silico* on $(WW)_2$, the time-dependent force $f(t) = r_f t$ is ramped up at the C-terminal end of the second WW domain (right), while the N-terminus of the first WW domain (left) is constrained. The forced unfolding transitions in WW domains occur in two steps: unfolding of the small loop (residues Thr29−Gly39; shown in light blue and light green) in WW domains (transition type 1; snapshots 1 and 3) and unfolding of the large loop (residues Lys6−Ile28; in blue and green) in WW domains (transition type 2; snapshots 2 and 4). See Table S1 in the Supporting Information for more detail. In snapshots 0−3, the intact (unfolded) structures are shown in solid circles (dashed rectangles). Panel b: Representative FX spectra for dimer $(WW)_2$ (shown in black and red color), obtained from dynamic force experiments *in silico*, are overlaid to demonstrate the stochastic nature of unfolding transitions. The force peaks for unfolding transitions of types 1 and 2, numbered 0−4, correspond to the accordingly numbered snapshots in panel a.

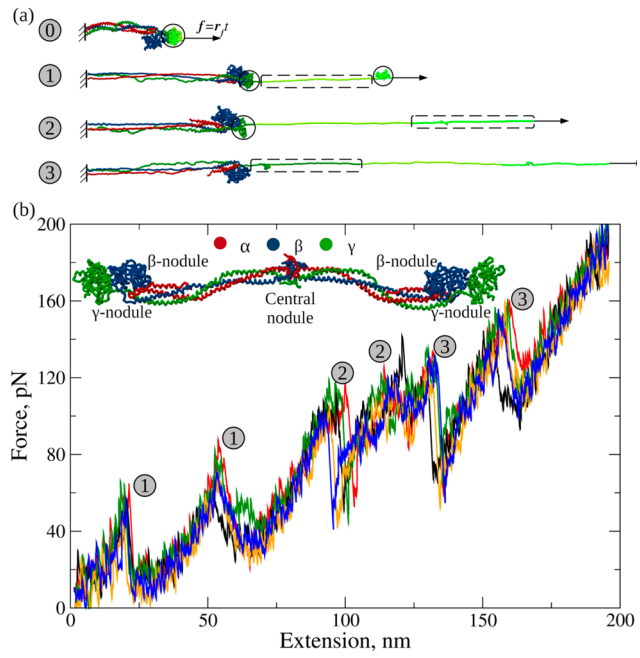


Figure 2. Forced unfolding transitions in Fg monomer. Panel a: Unfolding transitions of types 1−3 in Fg (summarized in Table SII) shown for the right half of Fg molecule. The time-dependent force $f(t) = r_f t$ is applied at the (right) C-terminal part of the γ chain (tagged residue γGly395), while residue γIle394 is constrained in the other (left) C-terminal part of the γ chain. The N-terminal (γ139−γ234) and C-terminal (γ311−γ381) parts of the γ-nodule are shown in dark and light green, respectively; the central region (γ234−γ311) is shown in yellow. The folded state (snapshot 0) becomes destabilized, which results in a series of unfolding transitions displayed using snapshots 0−3. These snapshots, which correspond to the accordingly numbered force peaks in the FX spectra in panel b, show the following unfolding transitions: unraveling of the central part of the γ-nodule (transition type 1; snapshot 1), unfolding of the C-terminal part of the γ-nodule (transition type 2; snapshot 2), and unfolding of the N-terminal part of the γ-nodule (transition type 3; snapshot 3). Panel b: Representative FX spectra for Fg monomer (displayed in different color for clarity) obtained from mechanical testing *in silico*. The force peaks numbered as 1−3 corresponding to the transitions of types 1−3 in panel a. The inset shows structural details of Fg molecule: the central nodule, γ-nodules, β-nodules, disulfide rings, and the γ−γ-cross-linking sites.

where r_i represent the coordinates of residues $i = 1, 2, ..., Q$. The distance between interacting residues i and $i + 1$ is $r_{i,i+1}$, and $r^0_{i,i+1}$ is its value in the native (PDB) structure. The first energy term in eq 1 is the finite extensible nonlinear elastic (FENE) potential V^{ATT}_{NB}, which describes the backbone chain connectivity; $R_0 = 2$ Å is the tolerance for the bond length change, and $k = 14$ N/m is the force constant. The second term in eq 1 is the Lennard-Jones potential V^{ATT}_{NB}, which accounts for the native interactions that stabilize the folded state. If the noncovalently linked residues i and j ($|i - j| > 2$) are within the cutoff distance R_C in the native state, *i.e.*, $r_{ij} < R_C$, then $\Delta_{ij} = 1$, and zero otherwise. All the non-native interactions described by the third term in eq 1 are treated using the repulsive Lennard-Jones potential V^{REP}_{NB}. Additional constraint was imposed on the bond angles formed by residues i, $i + 1$, and $i + 2$ by including the repulsive potential with parameters $\varepsilon_l = 1.0$ kcal/mol and $\sigma = 3.8$ Å, which quantify, respectively, the strength and the range of repulsion. To ensure the self-avoidance of the polypeptide chain, we set $\sigma = 3.8$ Å in third and fourth potential energy terms in eq 1.

Parameterization of SOP Models for $(WW)_2$ and Fg. The SOP model of $(WW)_2$ was derived from the atomic structure (Protein Data Bank (PDB) entry: 1PIN[69]). Two different parametrizations were used. In model M1, we used $\varepsilon_h = 1.5$ kcal/mol to specify the strength of nonbonded interactions (see eq 1). In model M2, we set $\varepsilon_h = 2.4$ kcal/mol for the small loop (Thr29−Gly39), and $\varepsilon_h = 0.6$ kcal/mol for the large loop (Lys6−Ile28). The SOP model of human Fg was derived from the atomic structure (PDB entry 3GHG[62]) and was parametrized as described in ref 63. The native contacts were divided into the following groups: (1) contacts in the central β-sheet of the γ-nodule, including residues γ189−197, γ243−284, and γ380−389 in the C-terminal β-strand (group 1); (2) contacts in the C-terminal part of the γ-nodule (γ284−380; group 2); (3) contacts in the N-terminal part of the γ-nodule (γ139−189 and γ197−243; group 3); (4) contacts in the α-helical regions in the

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

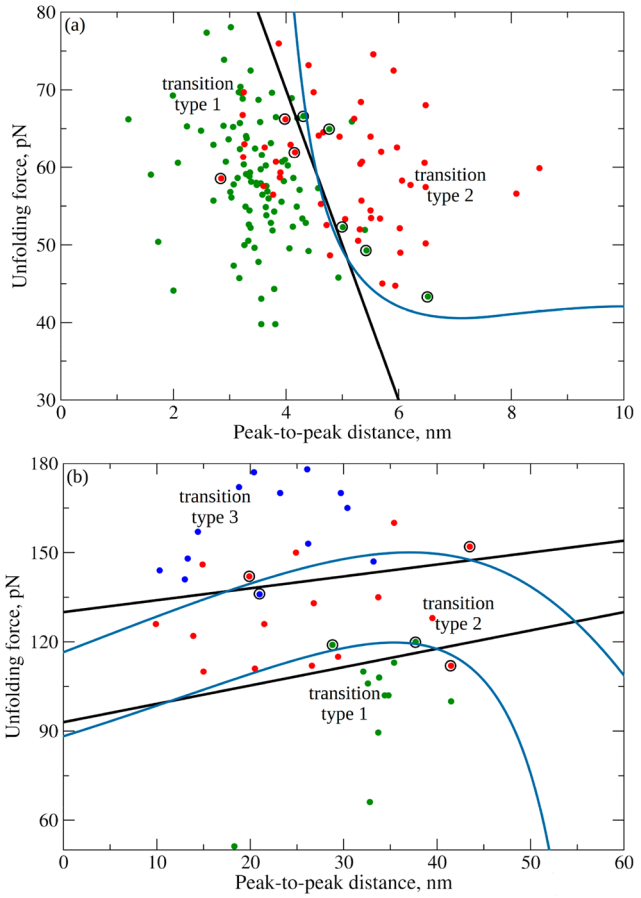


Figure 3. Characterizing unfolding transitions in dimer $(WW)_2$ and Fg monomer with maximum margin classifier (Case Study 1). Shown are scatterplots of the unfolding forces versus the peak-to-peak distances from test sets D_{test} characterizing the unfolding transitions in $(WW)_2$ (panel a) and Fg (panel b). The separating hyperplanes (black lines) divide the data into different classes: unfolding transitions of type 1 (green data points) and type 2 (red) for $(WW)_2$ (panel a; see Figure 1) and the unfolding transitions of type 1 (green data points), type 2 (red), and type 3 (blue data points) for the Fg monomer (panel b; see Figure 2). Some of the misclassified data points are shown in solid circles. Also shown are the separating hypersurfaces (blue curves) obtained using SVM classifier with the polynomial kernel of degree 5 (see the Supporting Information for more detail).

coiled-coil connectors (α45−200, β76−197, and γ19−139; group 4); and (5) contacts in the central nodule (α27−44, β58−75, and γ14−18; group 5) and in the β-nodules (β198−461; group 5). The following values of ε_h were used to describe these contacts: ε_h = 0.7, 1.2, 1.6 kcal/mol for groups 1, 2, and 3, respectively, and ε_h = 1.3 kcal/mol for groups 4 and 5.

Single-Molecule Dynamic Force Experiments in Silico. The unfolding dynamics of dimer $(WW)_2$ and Fg monomer was obtained by integrating the Langevin equations of motion for each amino−acid residue position $\mathbf{r}_i$ in the overdamped limit, $\vartheta\frac{d\mathbf{r}_i}{dt} = -\frac{\partial V}{\partial \mathbf{r}_i} + g_i(t)$, where $V = V_{MOL} - fX$ is the total potential energy V_{MOL} due to polypeptide chains (*i.e.*, molecular potentials for $(WW)_2$ or Fg) and the potential energy fX due to applied pulling force f, $g(t)$ is the Gaussian distributed zero-centered random force, which describes random collisions of amino acids with solvent molecules, and ϑ is the friction coefficient. To mimic the dynamic force-ramp measurement for dimer $(WW)_2$ *in vitro*, the N-terminal C_α-atom of the left WW domain was constrained and a time-dependent force $f(t) = fn$ with the magnitude $f = r_f t$ was applied to the C-terminal C_α-atom of the right WW domain in the direction n coinciding with the direction of the end-to-end vector X (Figure 1). For the Fg monomer, the left end of the molecule (γIle394) was constrained and force $f(t)$ was applied at the right end of the molecule (γGly395) in the direction coinciding with the direction of the end-to-end vector. The Langevin equations of motion were propagated forward with the time step Δt = 20 ps. Pulling simulations were carried out at room temperature (T = 300 K) using the bulk water viscosity, which corresponds to the friction coefficient $\vartheta = 7.0 \times 10^5$ pN ps/nm. We used the experimental values of the cantilever spring constant κ = 35 pN/nm and the pulling speed v_f = 10 μm/s, which translates to the force-loading rate $r_f = \kappa v_f$ = 350 nN/s.

Analysis of Simulation Output. For models M1 and M2, WW domains undergo unfolding from the native (folded) state (F) to the unfolded state (U) in two steps, F → I → U, where I is the intermediate (partially unfolded) conformation. The unfolding transition of type 1 corresponds to unraveling of the small loop (Thr29−Gly39; Figure 1). The unfolding transition of type 2 corresponds to unraveling of the large loop (Lys6−Ile28; Figure 1). A summarized description of unfolding transitions in dimer $(WW)_2$ is presented in Table SI in the Supporting Information. The Fg molecule undergoes forced unfolding in three steps, F → I_1 → I_2 → U, with two intermediate structures I_1 and I_2, which correspond to the following types of unfolding transitions (see Table SII): (1) separation of the C- and N-terminal parts of γ-nodule (transition type 1), (2) unfolding of the C-terminal part of γ-nodule (transition type 2), and (3) unfolding of its N-terminal part of γ-nodule (transition type 3). The force peaks for $(WW)_2$ and Fg were sorted into two and three groups, respectively, according to the type of unfolding transition they represent. The peak forces and peak-to-peak distances were evaluated for each transition types 1 and 2 for $(WW)_2$ and for each transition types 1−3 for Fg. The histogram-based estimates of the probability density functions (pdfs) of unfolding forces for $(WW)_2$ and Fg were constructed using the Freedman−Diaconis rule for the bandwidth selection.[70,71] We used nonparametric density estimation[63,64] with the kernel density $\varphi_K(f) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{h}K\left(\frac{f-f_i}{h}\right)$, where h is the bandwidth and $K(f) = \exp\left(-\frac{f^2}{2}\right)/\sqrt{2\pi}$ is the normalized Gaussian kernel function. We set the bandwidth to $h = M^{-1/5}$.[72]

Statistical Learning. *Support Vector Classifier.* We employed the Support Vector Classifier (SVC) method to classify the unfolding force data into the unfolding transitions of types 1 and 2 for $(WW)_2$ and types 1−3 for Fg. For a d-dimensional space of input variables, one attempts to find a hyperplane of the dimension $d - 1$, which has the largest distance to the nearest data points, called the maximum margin, in the training set.[73] In the 2-dimensional case of the unfolding force separated by the peak-to-peak distance, $\{(f_1, x_1), (f_2, x_2), ..., (f_M x_M)\}$, a hyperplane is a one-dimensional subspace (line), which has the maximum distance between the data points from different classes (unfolding transition types). The maximal margin classifier is defined by the equation of a line:

$$b_0 + b_1 x + b_2 f = 0 \quad (2)$$

with constant parameters b_0, b_1, and b_2. Equation 2 divides the data into two halves, and so they define the SVC and the

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

decision rule, which assign a class (transition type) to which a particular data point (unfolding force) belongs. If for any pair of data points (f_i, x_i) and (f_j, x_j) the classifier $b_0 + b_1 x_i + b_2$ $(f_i > 0)$ while $b_0 + b_1 x_j + b_2$ $(f_j < 0)$, then these data points end up on different sides of the hyperplane (belong to different classes). We also implemented Support Vector Machines using higher order polynomials of a degree 2, 3, 4, and 5 (see the Supporting Information).

Expectation−Maximization Method. Assume a bivariate normal density $\mathcal{N}(f, x|\mu_j, \sigma_{fj}, \sigma_{xj}, \sigma_{fxj})$ and prior probability π_j for each unfolding transition type j for the vector (f, x) of unfolding force f and peak-to-peak distance x. Let p denote the mixture density for the vector (f, x). Application of the EM method[73,74] is based on specifying the initial values for the mean vectors μ_j, standard deviations σ_{fj} and σ_{xj}, covariances σ_{fxj}, and prior probabilities π_j. In the first (expectation) step, the log-likelihood of expectation

$$\ln[p(\mathbf{F}, \mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi})] = \sum_{i=1}^{M} \ln\left[\sum_{j=1}^{J} \pi_j \mathcal{N}(f, x|\mu_j, \sigma_{fj}, \sigma_{xj}, \sigma_{fxj})\right] \tag{3}$$

is calculated. Here, $\mathbf{F}, \mathbf{X} = \{(f_1, x_1), ...(f_M, x_M)\}$are pairs of unfolding forces and peak-to-peak distances, $\boldsymbol{\mu} = \{\mu_1, ..., \mu_J\} = \{(\mu_{f1}, \mu_{x1}), ..., (\mu_{fJ}, \mu_{xJ})\}$, $\boldsymbol{\sigma} = \{(\sigma_{f1}, \sigma_{x1}, \sigma_{fx1}), ..., (\sigma_{fJ}, \sigma_{xJ}, \sigma_{fxJ})\}$, and $\boldsymbol{\pi} = \{\pi_1, ..., \pi_J\}$ are, respectively, the average unfolding forces and average peak-to-peak distances, standard deviations and covariances, and prior probabilities for the unfolding transitions of types $j = 1, 2, ..., J$. The log-likelihood $\ln[p(\mathbf{F},\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\pi})]$ for test observations is calculated using the estimates for parameters describing the unfolding transitions of type j. Next, the posterior probabilities γ_i^j for each transition type j

$$\gamma_i^j = \frac{\pi_j \mathcal{N}(f_i, x_i|\mu_j, \sigma_{fj}, \sigma_{xj}, \sigma_{fxj})}{\sum_{k=1}^{J} \pi_k \mathcal{N}(f_i, x_i|\mu_k, \sigma_{fk}, \sigma_{xk}, \sigma_{fxk})} \tag{4}$$

are calculated. In the second (maximization) step, one finds new values of μ_j, σ_{fj}, σ_{xj}, σ_{fxj}, and π_j for each transition type j that maximize the log-likelihood:

$$\mu_j^{new} = \frac{\sum_{i=1}^{M} \gamma_i^j (f_i, x_i)}{\sum_{i=1}^{N} \gamma_i^j} \tag{5}$$

$$\sigma_{fj}^{new} = \frac{\sum_{i=1}^{M} \gamma_i^j (f_i - \mu_{j1}^{new})^2}{\sum_{i=1}^{N} \gamma_i^j} \tag{6}$$

$$\sigma_{xj}^{new} = \frac{\sum_{i=1}^{M} \gamma_i^j (x_i - \mu_{j2}^{new})^2}{\sum_{i=1}^{N} \gamma_i^j} \tag{7}$$

$$\sigma_{fxj}^{new} = \frac{\sum_{i=1}^{M} \gamma_i^j (f_i - \mu_{j1}^{new})(x_i - \mu_{j2}^{new})}{\sum_{i=1}^{N} \gamma_i^j} \tag{8}$$

$$\pi_j^{new} = \frac{\sum_{i=1}^{M} \gamma_i^j}{J} \tag{9}$$

The expectation and maximization steps are repeated until convergence is reached, *i.e.*, when the difference between the old and new values of the average unfolding force is less than 0.1 pN, *i.e.*, $\|\mu_j^{new} - \mu_j^{old}\| < 0.1$ pN, for all transition types $j = 1, 2, ..., J$. We used the following two versions of EM algorithm. In the first case, all parameters (i.e., π_j, μ_j, σ_{fj}, σ_{xj}, σ_{fxj}, and γ_i^j) for all transition types $j = 1, 2, ..., J$ are allowed to change. In the second case, all parameters except for the prior probabilities π_j are allowed to change (π_j are kept at their initial values). Although the EM method attempts to find π_j, μ_j, σ_{fj}, σ_{xj}, and σ_{fxj} for bivariate normal densities, in this work we are concerned with the (marginal) distributions of unfolding forces characterized by the univariate normal densities $\mathcal{N}(f|\mu_{fj}, \sigma_{fj})$ with parameters π_j, μ_{fj}, and σ_{fj}.

RESULTS

Dynamic Force Experiments on $(WW)_2$ and Fg. In model M1 for $(WW)_2$, we used $\varepsilon_h = 1.5$ kcal/mol for all native contacts, which sets the strength of the nonbonded interactions (eq 1). The unfolding transitions are the two low and two high peaks as observed in the FX profiles (Figure S1). Because one of the main objectives of this study was to describe unfolding data characterized by overlapping ranges of unfolding forces, we increased the strength of native contacts in the small loop and decreased the strength of contacts in the large loop (Figure 1). In model M2, we set $\varepsilon_h = 2.4$ kcal/mol and $\varepsilon_h = 0.6$ kcal/mol for the native contacts in the small and large loops, respectively. For model M2, the FX profiles show four force peaks of nearly equal height (two peaks per WW domain; Figure S1b). Using models M1 and M2, we carried out Langevin simulations of the forced unfolding for $(WW)_2$, and construct a total of 100 FX spectra (Figure S1c), each containing four unfolding forces separated by peak-to-peak distances. The FX profiles for model M2 for $(WW)_2$ along with a schematic of mechanical testing *in silico* are displayed in Figure 1, which also shows the native state, partially unfolded structures, and the unfolded state. A schematic of mechanical testing of Fg monomer is displayed in Figure 2a, which also shows the native conformation for the right half of fibrinogen monomer, stretched conformations with unraveled central domain of the γ-nodule (transition type 1), with unfolded C-terminal part of the γ-nodule (transition type 2), and with unfolded N-terminal part of the γ-nodule (transition type 3); see Table SII. The output from single-molecule experiments and pulling simulations for Fg[63] was used to characterize the FX spectra (Figure 2b). Typical simulated FX spectra for Fg monomer are displayed in Figure 2.

Constructing Data Sets for $(WW)_2$ and Fg. Here, we summarize all the data sets used in Case Studies 1−5 below. Data sets for $(WW)_2$-simulated FX curves (Figure 1b and Figure S1) were used to form combined data set D ($M = 400$), which was divided into data sets D_1 ($M_1 = 200$) and D_2 ($M_2 = 200$) for the unfolding transitions of types 1 and 2. We randomly divided the combined data set D into training set D_{train} ($M_{train} = 200$) and test set D_{test} ($M_{test} = 200$). Data from D_{train} were subdivided into data subsets $D_{train,1}$ ($M_{train,1} = 100$) and $D_{train,2}$ ($M_{train,2} = 100$) for transitions of types 1 and 2. The maximal margin classifier (SVC) was based on D_{train} and was applied to D_{test} (Case Study 1; Figures 3 and 4). $D_{train,1}$ and $D_{train,2}$ were used to estimate the average forces μ_{fj} and average peak-to-peak distances μ_{xj} and the standard deviations σ_{fj} and σ_{xj} (EM; Case Study 3). SVC classifier was used to divide D_{test} into two classes: $D_{test,1}^{SVM}$ and $D_{test,2}^{SVM}$. These were used to construct the marginal histograms and nonparametric densities of unfolding forces for transitions of types 1 and 2, $p_1(f)$ and $p_2(f)$ (Case Study 3; Figure 4). Using D_{train}, we classified unfolding forces in D_1 and D_2 for the unfolding transitions of types 1 and 2. These data sets were used to construct nonparametric densities of unfolding forces for

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

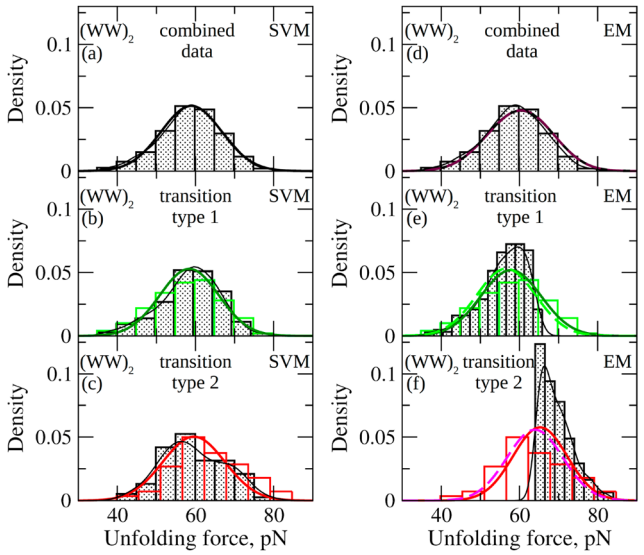


Figure 4. Performance of SVM and EM in modeling unfolding force data for dimer $(WW)_2$ (Case Studies 2 and 3): Compared for $(WW)_2$ are the histograms (bars), nonparametric densities (thin curves), and theoretical pdfs (thick curves) of unfolding forces (see Figure 1b and Figure S1) obtained from the pulling simulations for $(WW)_2$ using SVM (panels a-c) and EM (panels d−f). The histograms are overlaid with nonparametric density curves. The following data sets are shown as gray histograms: combined data set D_{test} (M_{test} = 200; panels a and d), and data sets $D^{SVM}_{test,1}$ ($M^{SVM}_{test,1}$ = 110; panel b), $D^{SVM}_{test,2}$ ($M^{SVM}_{test,2}$ = 90; panel c), $D^{ML}_{test,1}$ ($M^{ML}_{test,1}$ = 138; panel e), and $D^{ML}_{test,2}$ ($M^{ML}_{test,2}$ = 62; panel f). We compare the histograms and density estimates with theoretical pdfs of unfolding forces for weighted superposition $\sum_j \pi_j \mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$ obtained with SVM (panel a) and EM (panel d) and for the unfolding transition of types j = 1 and 2, $\mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$, obtained with SVM (panels b and c) and EM (panels e and f): j = 1 (panels b and e), j = 2 (panels c and f). In panels e and f, theoretical pdfs obtained with fixed prior probabilities (solid curves) and variable prior probabilities (dashed curves) are compared for unfolding transitions of types 1 and 2. In panels b and c and e and f, the colored transparent histograms represent actual unfolding force data for the transitions of type 1 (data set D_1, M_1= 200) and type 2 (data set D_2, M_2= 200).

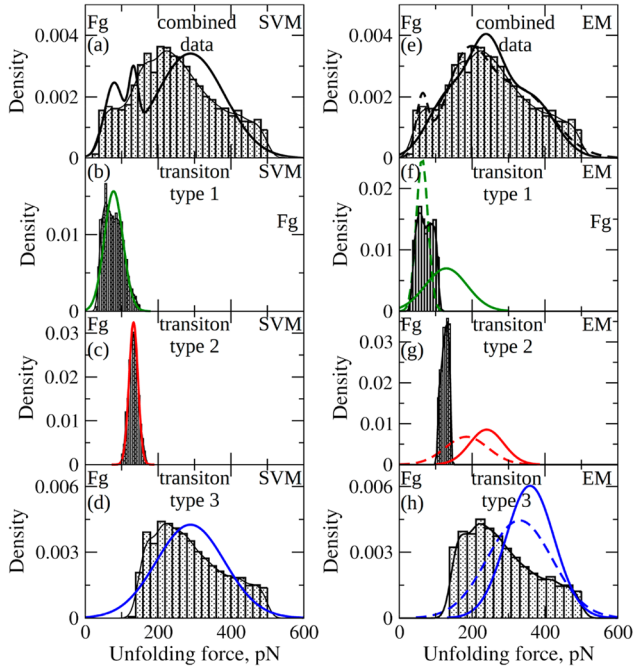


Figure 5. Performance of SVM and EM in modeling 0−500 pN unfolding force data for the Fg monomer (Case Study 4). Compared are the histograms (bars) of unfolding forces in the 500 pN range from single-molecule experiments for Fg,[63] nonparametric densities (thin curves), and theoretical pdfs (thick curves) of unfolding forces obtained with SVM (panels a−d) and EM (panels e−h). The following data sets were used: combined data set D_{500} (M_{500} = 15 023; panels a and e), and data sets $D^{SVM}_{500,1}$ ($M^{SVM}_{500,1}$ = 2165; panel b), $D^{SVM}_{500,2}$ ($M^{SVM}_{500,2}$ = 913; panel c), $D^{SVM}_{500,3}$ ($M^{SVM}_{500,3}$ = 11 495; panel d), $D^{ML}_{500,1}$ ($M^{ML}_{500,1}$ = 1862; panel f), $D^{ML}_{500,2}$ ($M^{ML}_{500,2}$ = 3837; panel g), and $D^{ML}_{500,3}$ ($M^{ML}_{500,3}$ = 9324; panel h). We compare the histograms and density estimates with theoretical pdfs of unfolding forces for weighted superposition $\sum_j \pi_j \mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$ obtained with SVM (panel a) and EM (panel e), as well as for the unfolding transition of types j = 1, 2 and 3, $\mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$, obtained with SVM (panels b−d) and EM (panels f−h): j = 1 (panels b and f), j = 2 (panels c and g), and j = 3 (panels d and h). In panels e−h, theoretical pdfs obtained with fixed (solid curves) and variable (dashed curves) prior probabilities are compared for the unfolding transitions of type 1, type 2, and type 3.

transitions of types 1 and 2 (Figure S3). We used $p_1(f)$ and $p_2(f)$ and the maximum likelihood estimation

$$B(f_i) = \ln p_{\hat{j}}(f_i) \quad \text{and} \quad \hat{j} = \underset{j}{\text{argmax}}\, p_j(f_i) \tag{10}$$

to subdivide D_{test} into data subsets $D^{ML}_{test,1}$ ($M^{ML}_{test,1}$ = 138) and $D^{ML}_{test,2}$ ($M^{ML}_{test,2}$ = 62) for the transitions of types 1 and 2 (Case Study 3; Figures 3 and 4). Using D_1 and D_2, we randomly selected 100 unfolding forces to construct combined data set $D_{(WW)_2}$, which was used to evaluate performance of EM algorithm (Figure 7). Using this set, we constructed data sets $D_{(WW)_2,1}$ ($M_{(WW)_2,1}$ = 50) and $D_{(WW)_2,2}$ ($M_{(WW)_2,2}$ = 50) for transitions of types 1 and 2.

Data sets for Fg-Experimental FX spectra for Fg[63] were used to form combined data set D_{exp} (M_{exp}= 20 193; Figure S3a). Using D_{exp}, we constructed data set D_{500} (M_{500}= 15 023), which excludes data that correspond to large unfolding forces >500 pN and long peak-to-peak distances >100 nm (Case Study 4; Figure 5). Next, we constructed data set D_{200} (M_{200} = 4572), which excludes forces below 30 pN and above 200 pN and distances shorter than 9 nm and longer than 62 nm (Case Study 5; Figure 6). Simulated FX spectra for Fg[63] were used to construct combined data set D_{sim} (M_{sim}= 82) and data subsets $D_{sim,1}$ ($M_{sim,1}$ = 24), $D_{sim,2}$ ($M_{sim,2}$ = 31), and $D_{sim,3}$ ($M_{sim,3}$ = 27) for unfolding transitions of types 1, 2, and 3. These were used to assess performance of the EM method (Figure 7). D_{sim} was randomly divided into training set D_{train} (M_{train} = 41) and test set D_{test} (M_{test} = 41). Using D_{train}, we created data sets $D_{train,1}$ ($M_{train,1}$ = 12), $D_{train,2}$ ($M_{train,2}$ = 15), and $D_{train,3}$ ($M_{train,3}$ = 13) for transitions of types 1, 2, and 3. These were used to obtain the maximal margin classifier (SVM; Case Study 1; Figure 3). Using D_{test}, we created data sets $D^{SVM}_{test,1}$ ($M^{SVM}_{test,1}$ = 12), $D^{SVM}_{test,2}$ ($M^{SVM}_{test,2}$ = 16), and $D^{SVM}_{test,3}$ ($M^{SVM}_{test,3}$ = 14) for transitions of types 1, 2, and 3. D_{train} was used to train the SVM classifier, which was applied to D_{test} (Case Studies 1 and 2; Figures 3, 5, and 6). The classifier was applied to observations in D_{500} and D_{200}. Using D_{500}, we constructed data sets $D^{SVM}_{500,1}$ ($M^{SVM}_{500,1}$ = 2165), $D^{SVM}_{500,2}$ ($M^{SVM}_{500,2}$ = 913), and $D^{SVM}_{500,3}$ ($M^{SVM}_{500,3}$ = 11 495) for transitions of types 1, 2, and 3 (Case Study 4; Figure 5). Using D_{200}, we constructed data sets $D^{SVM}_{200,1}$ ($M^{SVM}_{200,1}$ = 1571), $D^{SVM}_{200,2}$ ($M^{SVM}_{200,2}$ = 808), and $D^{SVM}_{200,3}$ ($M^{SVM}_{200,3}$ = 2110) for transitions of types 1, 2, and 3 (Case Study 5; Figure 6). When applying EM, we used D_{train} to estimate the

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

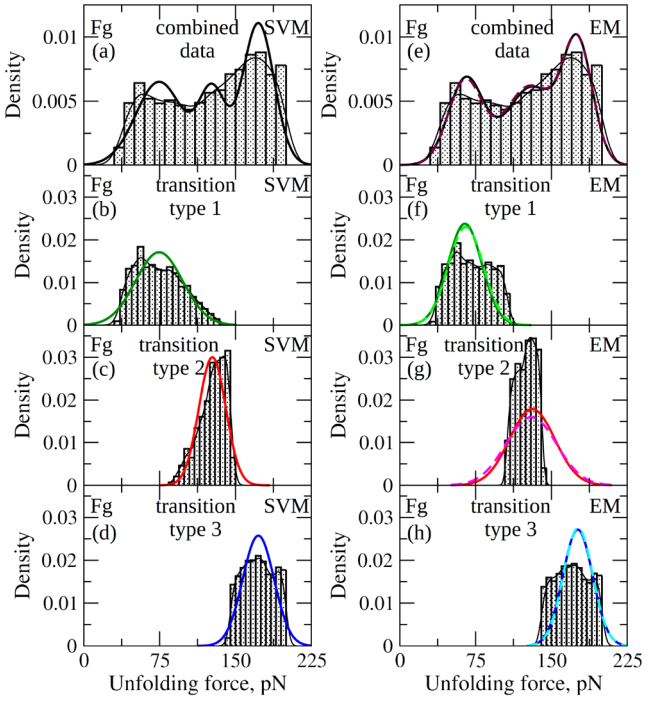


Figure 6. Performance of SVM and EM in modeling 30−200 pN unfolding force data for Fg monomer (Case Study 5). Compared are the histograms (bars) of unfolding forces in the 200 pN range from single-molecule experiments for Fg,[63] nonparametric densities (thin curves), and theoretical pdfs (thick curves) of unfolding forces obtained with SVM (panels a−d) and EM (panels e−h). The following data sets were used: combined data set D_{200} (M_{200} = 4572; panels a and e), and data sets $D_{200,1}^{SVM}$ ($M_{200,1}^{SVM}$ = 1691; panel b), $D_{200,2}^{SVM}$ ($M_{200,2}^{SVM}$ = 856; panel c), $D_{200,3}^{SVM}$ ($M_{200,3}^{SVM}$ = 1941; panel d), $D_{200,1}^{ML}$ ($M_{200,1}^{ML}$ = 1571; panel f), $D_{200,2}^{ML}$ ($M_{200,2}^{ML}$ = 808; panel g), and $D_{200,3}^{ML}$ ($M_{200,3}^{ML}$ = 2110; panel h). We compare histograms and density estimates with theoretical pdfs of unfolding forces for weighted superposition $\sum_j \pi_j \mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$ obtained with SVM (panel a) and EM (panel e), as well as for the unfolding transition of types j = 1, 2, and 3, $\mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$, obtained with SVM (panels b−d) and EM (panels f−h): j = 1 (panels b and f), j = 2 (panels c and g), and j = 3 (panels d and h). In panels e−h, theoretical pdfs obtained with fixed (solid curves) and variable (dashed curves) prior probabilities are compared for the unfolding transitions of type 1, type 2, and type 3.

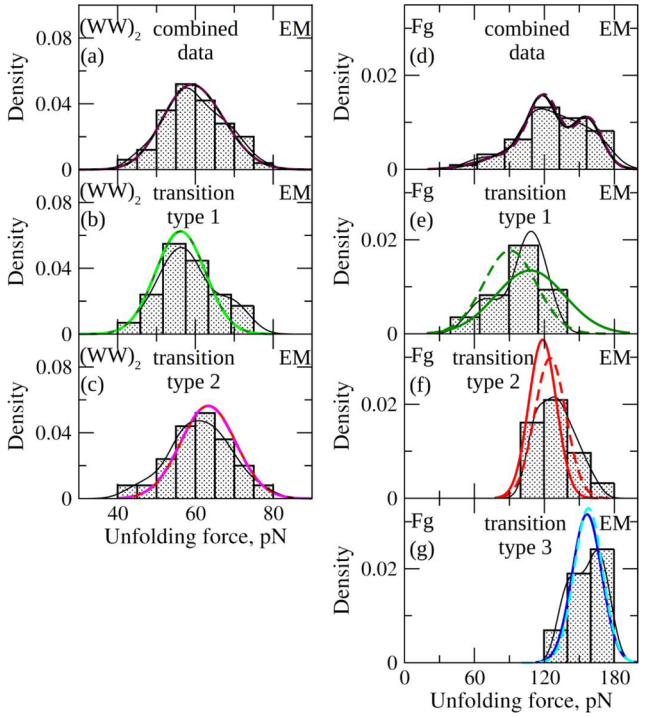


Figure 7. Performance of EM in resolving distributions of unfolding forces for $(WW)_2$ and Fg without data classification and data division (Case Study 6). Compared are the histograms (bars) of unfolding forces obtained from the pulling simulations for dimer $(WW)_2$ and the Fg monomer, nonparametric densities (thin curves), and theoretical pdfs (thick curves) of unfolding forces obtained with EM. The following data sets were used: for $(WW)_2$, combined data set $D_{(WW)_2}$ ($M_{(WW)_2}$ = 100; panels a and d) and data sets $D_{(WW)_2,1}$ ($M_{(WW)_2,1}$ = 50; panel b) and $D_{(WW)_2,2}$ ($M_{(WW)_2,2}$ = 50; panel c); for Fg, combined data set D_{Fg} (M_{Fg} = 82) and data sets $D_{sim,1}$ ($M_{sim,1}$ = 24), $D_{sim,2}$ ($M_{sim,2}$ = 31), and $D_{sim,3}$ ($M_{sim,3}$ = 27). We compare the force histograms and density estimates with theoretical pdfs of unfolding forces for weighted superposition $\sum_j \pi_j \mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$ (j = 1 and 2 for $(WW)_2$ and j = 1, 2, and 3 for Fg) obtained with EM (panels a and d), as well as for the unfolding transitions of types j = 1 and 2, $\mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$ obtained with EM for $(WW)_2$ (panel b for j = 1 and panel c for j = 2), and the unfolding transitions of types j = 1, 2, and 3, $\mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$ obtained with EM for Fg (panel e for j = 1, panel f for j = 2, and panel g for j = 3). Theoretical pdfs obtained with fixed (solid curves) and variable (dashed curves) prior probabilities are compared.

average forces μ_{fj} and average peak-to-peak distances μ_{xj}, and the standard deviations σ_{fj} and σ_{xj} (Case Study 4; Figures 5 and 6). We used $D_{sim,1}$, $D_{sim,2}$, and $D_{sim,3}$ to construct three nonparametric density estimations of unfolding forces of types 1, 2, and 3 and $p_1(f)$, $p_2(f)$, and $p_3(f)$ (see Figure S2) to classify data sets D_{500} and D_{200} with the maximum likelihood estimation (Case Study 4 and 5; Figures 5 and 6). Using D_{500}, we constructed data subsets $D_{500,1}^{ML}$ ($M_{500,1}^{ML}$ = 1862), $D_{500,2}^{ML}$ ($M_{500,2}^{ML}$ = 913), and $D_{500,3}^{ML}$ ($M_{500,3}^{ML}$ = 11 495) for transitions of types 1, 2, and 3 (Case Study 4; Figure 5). Using D_{200}, we constructed data subsets $D_{200,1}^{ML}$ ($M_{200,1}^{ML}$ = 1571), $D_{200,2}^{ML}$ ($M_{200,2}^{ML}$ = 808), and $D_{200,3}^{ML}$ ($M_{200,3}^{ML}$ = 2110) for transitions of types 1, 2, and 3 (Case Study 5; Figure 6).

Case Study 1. SVM-Based Classification of Unfolding Transitions for $(WW)_2$ and Fg. First, we assessed the performance of SVM in classification of unfolding force data. These can be characterized using the peak forces (f_i) and peak-to-peak distances (x_i) as input variables, *i.e.*, $\{(f_1, x_1), (f_2, x_2), ..., (f_M, x_M)\}$. From the simulations for $(WW)_2$ and Fg, we know exactly which unfolding transitions belong to which transition types for all M = 400 unfolding events for $(WW)_2$ and for all M = 82 unfolding events for Fg. We used training set D_{train} (M_{train}= 200) and test set D_{test} (M_{test}= 200) for $(WW)_2$, and training set D_{train} (M_{train}= 41) and test set D_{test} (M_{test}= 41) for Fg. The scatterplots of f_i versus x_i for $(WW)_2$ and Fg are displayed in Figure 3, which also shows the separating hyperplanes. Although the hyperplane is chosen to divide the training observations into two classes (unfolding transitions of types 1 and 2) for $(WW)_2$ and into three classes (transitions of types 1−3) for Fg, there are data points that are misclassified. In D_{test} for $(WW)_2$, 9 and 17 data points for transitions of type 1 (9% error) and type 2 (17% error) are misclassified. In D_{test} for Fg, 2, 6, and 1 data points for transition of type 1 (16% error), type 2 (38% error), and type 3 (7% error) are misclassified. In D_{train} for $(WW)_2$, 4 and 13 data points for transitions of type 1 (4% error) and type 2 (13% error)

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

Table 1. Performance of SVM and EM in Describing Unfolding Force Data for $(WW)_2$ (Case Studies 2 and 3)[a]

	SVM method		EM method	
methods/quantities	type 1	type 2	type 1	type 2
μ_{fj}, pN	58.4/57.4	59.6/62.0	57.8 (56.6)/57.4	65.3 (64.0)/62.0
σ_{fj}, pN	7.6/8.0	7.9/8.4	7.7 (7.5)/8.0	6.9 (7.2)/8.4
π_j	0.5/0.5	0.5/0.5	0.5 (0.52)/0.5	0.5 (0.48)/0.5
L_j^1	0.07	0.09	0.15 (0.20)	0.17 (0.20)

[a]Compared for unfolding transitions of types 1 and 2 in WW domains are the theoretical predictions for the average unfolding forces μ_{fj}, standard deviations σ_{fj}, and prior probabilities π_j, obtained with SVM and EM, and values of these quantities obtained from the pulling simulations (separated by the slash). For EM, values of μ_{fj}, σ_{fj}, and π_j obtained with fixed and variable (in parentheses) prior probabilities π_j are compared. Also shown are the values of L_j^1-norm obtained with fixed and variable (in parentheses) prior probabilities. The total error (L^1-norm for combined dataset D_{test} ($M_{\text{test}} = 200$)) is 0.06 for SVM, and it is 0.13 (0.13) for EM with fixed (variable) prior probabilities. The following data sets were used: $D_{\text{test},1}^{\text{SVM}}$ ($M_{\text{test},1}^{\text{SVM}} = 110$), $D_{\text{test},2}^{\text{SVM}}$ ($M_{\text{test},1}^{\text{SVM}} = 90$), $D_{\text{test},1}^{\text{ML}}$ ($M_{\text{test},1}^{ML} = 138$), and $D_{\text{test},2}^{\text{ML}}$ ($M_{\text{test},2}^{\text{ML}} = 62$).

are misclassified. In D_{train} for Fg, 2, 5, and 3 data points for transition of type 1 (16% error), type 2 (31% error), and type 3 (21% error) are misclassified. We also applied SVM to classify the unfolding transitions for $(WW)_2$ and Fg using higher order polynomials of degree 2, 3, 4, and 5 (see the Supporting Information). The results obtained are displayed in Figure 3. The test errors obtained using SVM with the polynomial kernel of degree 5 were roughly the same as the errors obtained with the linear classifier (eq 2). For this reason, we used the linear classifier in Case Studies 2−5 below.

Case Study 2. Resolving Distributions of Unfolding Forces for $(WW)_2$ with SVM. Next, we tested how accurately SVM resolves the distributions of unfolding forces for $(WW)_2$ using the simulated FX spectra. The Gaussian-like symmetric shape of the marginal histograms of unfolding forces for combined data set D (Figure 4a) reveals broad but unimodal distribution of unfolding forces due to the overlapping 37.4−77.8 and 42.6−82.2 pN force ranges, which characterize, respectively, the unfolding transitions of type 1 (data set D_1) and type 2 (data set D_2). We used the Gaussian ansatz $\mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$ to describe the marginal distributions of unfolding forces in terms of the average forces μ_{fj} and standard deviations σ_{fj} for the unfolding transitions of types $j = 1$ and 2. To perform classification, we applied SVM to the unfolding data in D_{train}, and used the maximal margin classifier to divide the data in D_{test} into data subsets for transition types 1 ($D_{\text{test},1}^{\text{SVM}}$) and 2 ($D_{\text{test},2}^{\text{SVM}}$). Using these subsets, we constructed nonparametric densities $p_j(f)$ and calculated the average forces μ_{fj} and standard deviations σ_{fj} for the transitions of types $j = 1$ and 2 (Table 1). Then, values of μ_{fj} and σ_{fj} were substituted into the Gaussian ansatz to obtain the theoretical curves of $\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1})$ and $\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$ and weighted superposition $\pi_1\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1}) + \pi_2\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$. We set the weights (prior probabilities) to be equal, *i.e.*, $\pi_1 = \pi_2 = 1/2$. The histograms and nonparametric densities, and theoretical pdf curves of the unfolding forces obtained with SVM for combined data set D_{test}, and data subsets $D_{\text{test},1}^{\text{SVM}}$ (transition type 1) and $D_{\text{test},1}^{\text{SVM}}$ (transition type 2) are compared in Figure 4.

To quantify the difference between theoretical pdfs (constructed for training data) and nonparametric densities (constructed for test observations) of the unfolding transition of type j, we used the L^1-norm:

$$L_j^1 = \sum_{i=1}^{M} |\pi_j y_i - w_j \psi_i| \Delta f \tag{11}$$

where y_i is the value of probability density corresponding to the unfolding force f_i, *i.e.*, $y_i = \mathcal{N}(f_i|\mu_j, \sigma_j)$ predicted by SVM (or EM), ψ_i is the value of kernel density estimate corresponding to the same force value f_i, *i.e.*, $\psi_i = \varphi_K(f_i)$, and Δf is the force interval. In eq 11, π_j is the prior probability estimated theoretically. For example, for $(WW)_2$ $\pi_j = 1/2$ for both transition types 1 and 2 (Table 1). In eq 11, w_j is the weight of jth transition type, which in the case of $(WW)_2$ is also equal to 1/2. The obtained values of L^1-norm were $L_1^1 = 0.07$ and $L_2^1 = 0.09$ for the unfolding transitions of types 1 and 2 (Table 1).

Case Study 3. Resolving Distributions of Unfolding Forces for $(WW)_2$ with EM. We tested how accurately EM resolves the distributions of unfolding forces for $(WW)_2$. First, using data sets D_1 and D_2 we constructed nonparametric densities of the unfolding forces for unfolding transitions of type 1 and type 2, $p_1(f)$ and $p_2(f)$, and then carried out classification of unfolding forces based on D_{train} using maximum likelihood estimation (eq 10). We constructed subsets $D_{\text{test},1}^{\text{ML}}$ ($M_{\text{test},1}^{\text{ML}} = 138$) and $D_{\text{test},2}^{\text{ML}}$ ($M_{\text{test},2}^{\text{ML}} = 62$) for transitions of types 1 and 2. Next, we applied the EM algorithm and used the bivariate Gaussian ansatz with the average forces $\boldsymbol{\mu}_j = (\mu_{fj}, \mu_{xj})$ and standard deviations $\boldsymbol{\sigma}_j = (\sigma_{fj}, \sigma_{xj})$. We set initial values of μ_{fj} for transitions of types $j = 1$ and 2 to be equal to the average unfolding forces estimated from $D_{\text{train},1}$ (57.4 pN) and $D_{\text{train},2}$ (62.0 pN); initial values of μ_{xj} were estimated from $D_{\text{train},1}$ (3.4 nm) and $D_{\text{train},2}$ (5.0 nm); initial values of standard deviations of unfolding forces σ_{fj} and peak-to-peak distances were set to be equal to $\sigma_f/\sqrt{2}$ and $\sigma_x/\sqrt{2}$ (σ_f = 7.8 pN and σ_x = 1.2 nm are the standard deviations of unfolding forces and peak-to-peak distances obtained for data from D_{test}). We considered cases of fixed and variable π_j for the unfolding transitions of types $j = 1$ and 2.

Fixed Prior Probabilities. We set $\pi_1 = \pi_2 = 0.5$ and varied μ_j and σ_j. The convergence was reached in 15 steps. Final values of μ_{fj} and σ_{fj} are compared with the values of these quantities from subsets D_1 and D_2 in Table 1. Because we used the results of simulations, we know the actual values of μ_{fj} and σ_{fj}, and so the predicted and actual values can be directly compared. We constructed the marginal histograms and nonparametric densities $p(f)$, $p_1(f)$, and $p_2(f)$ of unfolding forces using D_{test}, $D_{\text{test},1}^{\text{ML}}$, and $D_{\text{test},2}^{\text{ML}}$ for unfolding transitions of types $j = 1$ and 2. These are compared in Figure 4 with theoretical marginal pdfs of unfolding forces derived by substituting the final values of μ_{fj} and σ_{fj} obtained with EM (Table 1) into the superposition $\pi_1\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1}) + \pi_2\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$ (D_{test}; Figure 4d), into $\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1})$ ($D_{\text{test},1}^{\text{ML}}$ for transitions of type 1; Figure 4e), and into $\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$ ($D_{\text{test},2}^{\text{ML}}$ for transitions of type 2; Figure 4f). The agreement between the histograms and density estimates,

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

Table 2. Performance of SVM and EM in Describing 500 pN Experimental Unfolding Forces for Fg (Case Study 4)[a]

	SVM method			EM method		
methods/quantities	type 1	type 2	type 3	type 1	type 2	type 3
μ_{fj}, pN	77.8/98.0	132.6/124.0	288.8/164.0	136.9 (62.7)/71.2	239.9 (184.54)/125.0	358.6 (330.2)/282.3
σ_{fj}, pN	25.5/20.3	12.3/15.3	93.7/13.8	57.2 (16.0)/19.9	46.7 (58.6)/9.6	66.1 (89.6)/94.4
π_j	0.14/0.29	0.06/0.38	0.80/0.31	0.29 (0.07)/0.11	0.38 (0.454)/0.07	0.31 (0.49)/0.82
L_j^1	0.03	0.01	0.26	0.33 (0.07)	0.45 (0.47)	0.70 (0.37)

[a]Compared unfolding transitions of types 1−3 in Fg are theoretical predictions for the average unfolding forces μ_{fj}, standard deviations σ_{fj} and prior probabilities π_j obtained with SVM and EM, and values of these quantities obtained from the experimental data (separated by the slash). For EM, values of μ_{fj}, σ_{fj}, and π_j obtained with fixed and variable (in parentheses) prior probabilities are compared with the same quantities obtained from experiment (separated by the slash). Also shown are values of L_j^1-norm. For EM, values of L_j^1-norm obtained with variable prior probabilities are shown in parentheses. The total error (L^1-norm for combined dataset D_{500}) is 0.27 for the SVM method, and it is 0.17 (0.13) for the EM method with fixed (variable) prior probabilities. The following data sets were used: D_{500} (M_{500} = 15 023) and data sets $D_{500,1}^{SVM}$ ($M_{500,1}^{SVM}$ = 2165), $M_{500,2}^{SVM}$ ($M_{500,2}^{SVM}$ = 913), $D_{500,3}^{SVM}$ ($M_{500,3}^{SVM}$ = 11 945), $D_{500,1}^{ML}$ ($M_{500,1}^{ML}$ = 1862), $D_{500,2}^{ML}$ ($M_{500,2}^{ML}$ = 3837), and $D_{500,3}^{ML}$ ($M_{500,3}^{ML}$ = 9324)

and theoretical pdf curves is very good: $L_1^1 = 0.15$ and $L_2^1 = 0.17$ for the unfolding transitions of types 1 and 2 (Table 1).

Variable Prior Probabilities. We varied π_j, μ_j, and σ_j. We set initial values of μ_{fj} to be equal to the average unfolding forces from $D_{train,1}$ (57.4 pN) and $D_{train,2}$ (62.0 pN). Initial values of μ_{xj} were set to the peak-to-peak distances corresponding to the unfolding forces estimated from 3D histogram of bivariate data (not shown): 2.2 and 5.8 nm. Initial values of the prior probabilities were set to 0.5 and standard deviations were set to be equal to $\sigma_{f1} = \sigma_{f2} = \sigma_f/\sqrt{2}$ for unfolding forces and $\sigma_{x1} = \sigma_{x2} = \sigma_x/\sqrt{2}$ for peak-to-peak distances (σ_f = 7.8 pN and σ_x = 1.2 nm are the standard deviations for data from D_{test}). We applied the EM algorithm to bivariate data from D_{test}. The convergence was reached in 15 steps. Final values of π_j, μ_{fj}, and σ_{fj} are compared with the values of these quantities for data sets D_1 and D_2 in Table 1. The predicted and actual values of μ_{fj} and σ_{fj} show good agreement, but values $\pi_1 = 0.52$ and $\pi_2 = 0.48$ deviate from 0.5 (Table 1). The histograms and nonparametric densities for combined data set D and for data sets D_1 and D_2 for transitions of types j = 1 and 2, $p(f)$, $p_1(f)$, and $p_2(f)$, are compared in Figure 4 with theoretical pdf curves of weighted superposition $\pi_1 N(f_i|\mu_{f1},\sigma_{f1}) + \pi_2 N(f_i|\mu_{f2},\sigma_{f2})$ $\pi_1\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1}) + \pi_2\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$ (D_{test}; Figure 4d), $\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1})$ ($D_{test,1}^{ML}$ for transitions of type 1; Figure 4e), and $\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$ ($D_{test,2}^{ML}$ for transitions of type 2; Figure 4f). Theoretical curves were derived using final values of π_j, μ_{fj}, and σ_{fj} from Table 1. The agreement between the histograms and density estimates, and theoretical pdf curves of the unfolding forces is very good: $L_1^1 = 0.20$ and $L_2^1 = 0.20$ for the unfolding transitions of types 1 and 2 (see Table 1).

Case Study 4. Resolving Distributions of 500 pN Unfolding Forces for Fg with SVM and EM. Next, we compared the performance of SVM and EM using the experimental unfolding force data from combined data set D_{500}, which excludes large unfolding forces >500 pN that correspond to protein desorption and/or cantilever tip detachment.[63]

Classification. SMV classification of unfolding data for Fg was performed using the results of simulations as described in Case Studies 1 and 2. We obtained the maximal margin classifier based on D_{train} (Figure 3b) and applied the classifier to test observations in D_{500}, $\{(f_1, x_1), (f_2, x_2), ..., (f_M x_M)\}$, which contains the experimental peak forces in the 0−500 pN range and peak-to-peak distances in the 0-100 nm range. Combined data set D_{500} was subdivided into subsets for the unfolding transitions of types 1−3, $D_{500,1}^{SVM}$, $D_{500,2}^{SVM}$, and $D_{500,3}^{SVM}$. The results are displayed in Figure 5, which shows the histograms and nonparametric densities $p(f)$, $p_1(f)$, $p_2(f)$, and $p_3(f)$ for combined data set D_{500} (Figure 5a) and for subsets $D_{500,1}^{SVM}$ (transitions of type 1; Figure 5b), $D_{500,2}^{SVM}$ (type 2; Figure 5c), and $D_{500,3}^{SVM}$ (type 3; Figure 5d). EM classification for Fg was performed using the results of simulations as described in Case Study 3. Using data sets $D_{sim,1}$, $D_{sim,2}$, and $D_{sim,3}$ we constructed nonparametric densities $p_1(f)$, $p_2(f)$, and $p_3(f)$ for the unfolding transitions of types 1−3 (Figure S2). These were used in maximum likelihood estimation to classify data from combined data set D_{500} into data subsets $D_{500,1}^{ML}$, $D_{500,2}^{ML}$, and $D_{500,3}^{ML}$ for transitions of types 1−3. Figure 5 shows the histograms and nonparametric densities of unfolding forces obtained using data from D_{500} (Figure 5a), $D_{500,1}^{ML}$ (Figure 5b), $D_{500,2}^{ML}$ (Figure 5c), and $D_{500,3}^{ML}$ (Figure 5d).

Regression. Resolving the distributions of unfolding forces for Fg with SVM was performed as in Case Study 2. We used the Gaussian ansatz $\mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$ to approximate the pdfs of unfolding forces. For each subset $D_{500,1}^{SVM}$, $D_{500,2}^{SVM}$, and $D_{500,3}^{SVM}$ (transition types j = 1−3), we calculated the values of μ_{fj} and σ_{fj}. These were substituted into the Gaussian ansatz to obtain the theoretical pdfs of unfolding forces for each transition type, $\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1})$, $\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$, and $\mathcal{N}(f_i|\mu_{f3}, \sigma_{f3})$, and the pdf for weighted superposition, $\pi_1\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1}) + \pi_2\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2}) + \pi_3\mathcal{N}(f_i|\mu_{f3}, \sigma_{f3})$. The weights π_j were evaluated by dividing the number of data points M_j for transition type j in subset $D_{500,j}^{SVM}$ by the total number of data points in D_{500} (M_{500} = 15 023), $\pi_j = M_j/M_{500}$. The theoretical pdf curves are compared with the histograms and nonparametric densities of unfolding forces in Figure 5. We obtained $L_1^1 = 0.03$, $L_2^1 = 0.01$, and $L_3^1 = 0.26$ for the unfolding transitions of types 1, 2, and 3 (Table 2).

EM regression was carried out as in Case Study 3. We used EM to divide data in combined data set D_{500} into the unfolding transitions of types 1−3. We tested EM algorithm for two cases: with fixed prior probabilities, $\pi_1 = 0.29$, $\pi_2 = 0.38$, and $\pi_3 = 0.33$, and with variable prior probabilities (see Case Study 3). In the latter, initial values of π_j were set to $\pi_j = M_j/M_{500}$, where M_j is the number of data points in data sets $D_{500,1}^{ML}$, $D_{500,2}^{ML}$, and $D_{500,3}^{ML}$ for transitions of types j = 1−3 obtained with maximum likelihood. In both cases, we set initial values of μ_{fj} for transitions of type j to be equal to positions of three highest bins in the histogram of unfolding forces of combined data D_{500} (Figure 5), *i.e.*, 61, 173, and 207 pN, respectively; initial values of μ_{xj} were set to be equal to the peak-to-peak distances estimated from 3D histogram of

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

Table 3. Performance of SVM and EM in Describing 30−200 pN Experimental Unfolding Forces for Fg (Case Study 5)[a]

	SVM method			EM method		
methods/quantities	type 1	type 2	type 3	type 1	type 2	type 3
μ_{fj}, pN	74.5/98.0	127.1/124.0	172.6/164.0	65.3 (64.4)/70.9	136.7 (127.8)/124.7	177.7 (171.3)/170.4
σ_{fj}, pN	23.3/20.3	13.3/15.3	15.5/13.8	17.3 (16.8)/19.7	20.7 (15.2)/9.6	13.7 (17.1)/16.8
π_j	0.38/0.29	0.19/0.38	0.33/0.31	0.29 (0.27)/0.35	0.38 (0.40)/0.18	0.33 (0.33)/0.47
L_j^1	0.09	0.05	0.12	0.14 (0.16)	0.35 (0.37)	0.23 (0.22)

[a]Compared for unfolding transitions of types 1−3 in Fg are theoretical predictions of the same quantities as in Table 2 obtained with SVM and EM methods, but for the unfolding forces in the 30−200 pN range of forces. The total error (L^1-norm for combined dataset D_{200}) is 0.19 for SVM method, and 0.15 (0.14) for EM with fixed (variable) prior probabilities. The following data sets were used: D_{200} (M_{200} = 4572) and data sets $D_{200,1}^{SVM}$ ($M_{200,1}^{SVM}$ = 1691), $D_{200,2}^{SVM}$ ($M_{200,2}^{SVM}$ = 856), $D_{200,3}^{SVM}$ ($M_{200,3}^{SVM}$ = 1941), $D_{200,1}^{ML}$ ($M_{200,1}^{ML}$ = 1571), $D_{200,2}^{ML}$ ($M_{200,2}^{ML}$ = 808), and $D_{200,3}^{ML}$ ($M_{200,3}^{ML}$ = 2110).

bivariate data (not shown), which correspond to these forces, *i.e.*, 26, 22, and 27 nm, respectively. We set initial values of σ_{fj} to be equal to $\sigma_f/\sqrt{2}$ for unfolding forces and $\sigma_x/\sqrt{2}$ for peak-to-peak distances (σ_f = 115.7 pN and σ_x = 21.3 nm are obtained on the basis of data from D_{500}). The convergence of EM algorithm was reached in 37 and 34 steps, respectively. Final values of π_j, μ_j, and σ_j obtained with EM are compared with actual values of these quantities in Table 2. The histograms and nonparametric densities $p(f)$, $p_1(f)$, $p_2(f)$, and $p_3(f)$ for the combined data set D_{500} and for data sets $D_{500,1}^{ML}$, $D_{500,2}^{ML}$, and $D_{500,3}^{ML}$ for transitions of types j = 1−3 are compared in Figure 5 with the theoretical pdfs of unfolding forces for weighted superposition $\pi_1\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1}) + \pi_2\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2}) + \pi_3\mathcal{N}(f_i|\mu_{f3}, \sigma_{f3})$ for D_{500} (Figure 5e), $\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1})$ for $D_{500,1}^{ML}$ (transition of type 1; Figure 5f), $\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$ for $D_{500,2}^{ML}$ (type 2; Figure 5g), and $\mathcal{N}(f_i|\mu_{f3}, \sigma_{f3})$ for $D_{500,3}^{ML}$ (type 3; Figure 5h). The theoretical curves were derived using π_j, μ_{fj}, and σ_{fj} obtained with EM (Table 2). For fixed (variable) prior probabilities, we obtained L_1^1 = 0.33 (0.07), L_2^1 = 0.45 (0.47), and L_3^1 = 0.70 (0.37) for the unfolding transitions of types 1, 2, and 3 (Table 2).

Case Study 5. Resolving Distributions of 30−200 pN Unfolding Forces for Fg with SVM and EM. Last, we applied SVM and EM to resolve the distributions of unfolding forces for Fg using the experimental unfolding data from combined data set D_{200}, which excludes the unfolding forces below 30 pN, due to nonspecific interactions, and above 200 pN, which correspond to several unfolding transitions that occur simultaneously (Table SII).

Classification. SMV classification of unfolding force data for Fg was performed using the results of simulations as described in Case Study 4. The maximum margin classifier was based on data from D_{train} and applied to data in D_{200}. The results are displayed in Figure 6, which compares the histograms and nonparametric densities of unfolding forces for D_{200} (Figure 6a), and for data sets $D_{200,1}^{SVM}$ (Figure 6b), $D_{200,2}^{SVM}$ (Figure 6c), and $D_{200,3}^{SVM}$ (Figure 6d) for unfolding transitions of types 1−3. EM classification of unfolding forces was performed using data sets $D_{sim,1}$, $D_{sim,2}$, and $D_{sim,3}$ as described in Case Study 4. The nonparametric densities $p_1(f)$, $p_2(f)$, and $p_3(f)$ for transitions of types 1−3 (Figure S2) and maximum likelihood estimation were used to subdivide combined data set D_{200} into subsets $D_{200,1}^{ML}$, $D_{200,2}^{ML}$, and $D_{200,3}^{ML}$ for transitions of types 1−3. In Figure 6, the histograms and nonparametric densities are compared with the theoretical pdfs of unfolding forces for D_{200} (Figure 6a), and for $D_{200,1}^{ML}$ (transitions of type 1; Figure 6b), $D_{200,2}^{ML}$ (type 2; Figure 6c), and $D_{200,3}^{ML}$ (type 3; Figure 6d).

Regression. Resolving the distributions of unfolding forces for Fg with SVM was performed as described in Case Study 4. For each subset $D_{200,1}^{SVM}$, $D_{200,2}^{SVM}$, and $D_{200,3}^{SVM}$, we calculated the values of μ_{fj} and σ_{fj}, which were substituted into the Gaussian ansatz to obtain the theoretical pdfs of unfolding forces for transitions of types j = 1, 2, and 3, $\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1})$, $\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$, and $\mathcal{N}(f_i|\mu_{f3}, \sigma_{f3})$, and for weighted superposition $\pi_1\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1}) + \pi_2\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2}) + \pi_3\mathcal{N}(f_i|\mu_{f3}, \sigma_{f3})$, with the weights $\pi_j = M_j/M_{200}$ (M_{200} = 4572 and M_j is the number of data points in subsets $D_{200,1}^{ML}$, $D_{200,2}^{ML}$, and $D_{200,3}^{ML}$). The theoretical pdf curves are compared with the histograms and nonparametric densities of unfolding forces in Figure 6, which shows good agreement: L_1^1 = 0.09, L_2^1 = 0.05, and L_3^1 = 0.12 for the unfolding transitions of types 1, 2, and 3 (Table 3).

EM regression was carried out as described in Case Study 4. We applied EM to divide data in combined data set D_{200}. We tested EM with fixed and variable prior probabilities; in the first case, we set π_1 = 0.29, π_2 = 0.38, and π_3 = 0.33; in the second case, we varied π_j starting from $\pi_j = M_j/D_{200}$. In both cases, we set initial values of μ_{fj} for the unfolding transitions of types j = 1−3 to be equal to the positions of three highest bins in the histogram of D_{200} (Figure 6), 59, 145, and 172 pN, respectively; initial values of μ_{xj} were set to be equal to the peak-to-peak distances estimated from the 3D histogram of bivariate data (not shown), which correspond to these forces, *i.e.*, 22, 36, and 30 nm, respectively. We set initial values of σ_{fj} to be equal to $\sigma_f/\sqrt{2}$ for unfolding forces and $\sigma_x/\sqrt{2}$ for peak-to-peak distances (σ_f = 47.7 pN and σ_x = 12.6 nm are based on data from D_{200}). The convergence of EM algorithm was reached in 30 and 28 steps, respectively. Final values of π_j, μ_{fj}, and σ_{fj} obtained with EM are compared with the actual values of these quantities in Table 3. The force histograms and nonparametric densities for combined data set D_{200} and for subsets $D_{200,1}^{ML}$, $D_{200,2}^{ML}$, and $D_{200,3}^{ML}$ for the transitions of types j = 1−3 are compared in Figure 6 with the theoretical pdfs of unfolding forces for $\pi_1\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1}) + \pi_2\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2}) + \pi_3\mathcal{N}(f_i|\mu_{f3}, \sigma_{f3})$ (Figure 6e), $\mathcal{N}(f_i|\mu_{f1}, \sigma_{f1})$ (transition of type 1; Figure 6f), $\mathcal{N}(f_i|\mu_{f2}, \sigma_{f2})$ (type 2; Figure 6g), and $\mathcal{N}(f_i|\mu_{f3}, \sigma_{f3})$ (type 3; Figure 6h). The theoretical pdf curves were derived using π_j, μ_{fj}, and σ_{fj} obtained with EM (Table 3). For fixed (variable) prior probabilities, the L^1-norms were L_1^1 = 0.14 (0.16), L_2^1 = 0.35 (0.37), and L_3^1 = 0.23 (0.22) for the unfolding transitions of types 1, 2, and 3 (Table 3).

DISCUSSION

Over the past several decades, single-molecule spectroscopy has become a powerful approach to explore the dynamic processes that involve proteins at the single-molecule level of detail. A number of experimental techniques, including AFM[21−25] and

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

magnetic[30−34] and optical tweezers,[26−29] have been used by researchers to subject proteins to mechanical forces and to probe conformational transitions in proteins. The force-induced extension of a polypeptide chain leads to an increase in the restoring force, which results in the protein unraveling when the applied force exceeds the limits of protein mechanical stability and chain elongation, which in turn leads to a decrease in the restoring force due to loss of tension. This process of the gradual force increase followed by the sudden force drop is repeated over again until all protein domains have become unfolded, which results in a repeated sawtooth-like pattern of the unfolding force as a function of the end-to-end distance. For example, in the FX spectra for dimer $(WW)_2$ (Figure 1), the force maxima corresponding to the unfolding transitions of types 1 and 2 are due to unraveling of the small loop (Thr29−Gly39) and the large loop (Lys6−Ile28), respectively. In the FX spectra for Fg monomer (Figure 2), the peak forces are due to unfolding transitions in the left and right γ-nodules and elongation of the coiled coils in the Fg molecule.

Meaningful interpretation of the results of single-molecule experiments on biomolecules (RNA, DNA, and proteins) remains challenging. In addition, due to the nonspecific events that contribute to the force signal, the FX spectra show first and last force peaks due to protein desorption from the substrate and/or tip detachment. In experiments on complex multidomain proteins, such as fibrinogen oligomers $(Fg)_n$, the desorption peaks might appear in the middle portion of the FX spectrum. The cantilever tip might pick up and desorb one end of $(Fg)_n$, then stretch and unfold a half of the $(Fg)_n$ chain, and then desorb, stretch, and unfold the other half of $(Fg)_n$. Furthermore, several unfolding events might occur simultaneously (*i.e.*, in one step), and so the FX spectra might display force signals owing to several unfolding transitions (rather than single transition). For example, the unfolding transitions of types 1 and 2 in dimer $(WW)_2$ occur in the strongly overlapping ∼30−80 and ∼40−90 pN ranges of unfolding forces (see Figure S2), and so these unfolding transitions might occur simultaneously. Hence, it is difficult to extract the "physically relevant data" (due to single unfolding transitions), from the "raw experimental data", which always includes a large amount of noise (nonspecific interactions, protein desorption, tip detachment, simultaneous unfolding events, *etc.*) using the experimental force-extension spectra alone.

We tested several statistical modeling approaches to understanding the forced unfolding data based on Support Vector Machines and Expectation Maximization. SVM is widely used in supervised data classification,[49] whereas EM is broadly used in unsupervised learning to describe a mixture of several distributions. We proposed an approach, in which the results of mechanical testing experiments *in silico* are used as training data. Statistical models are, first, trained using the output from the pulling simulations, and then applied to understand the experimental data. In this context, "understanding experimental data" includes (i) correctly assigning an observation (peak force) to the type of unfolding transition it represents (classification) and (ii) accurately resolving the distribution of unfolding forces for each transition type (regression). On the one hand, outputs from the simulations have a relatively small size (from tens to a hundred of data points) because of a large computational cost, but trajectories of forced protein unfolding are free from noisy force signals discussed above. On the other hand, the experimental single-molecule data are characterized by a large sample size (thousands of data points), but the experimental raw data are noisy.

Here, we carried out Case Studies 1−5, in which we compared the accuracy of SVM- and EM-based methods in solving classification and regression problems for the forced unfolding data for proteins. We demonstrated that, although the experimental and simulated force-extension data might disagree, computer simulations can provide accurate information about the types of unfolding transitions (*e.g.*, prior probabilities for unfolding transitions of types 1 and 2 for $(WW)_2$, and types 1−3 for Fg) but might not be accurate in terms of the statistics of unfolding forces (average forces and standard deviations). Hence, results of computational molecular modeling can be used as a good starting point to train statistical models and to separate the unfolding data from noise. When the input from computational molecular modeling is not available, single-molecule forced unfolding experiments on proteins can be complemented by the results from thermal unfolding assays or analysis of structures of the protein in question. A good understanding of the number and types of unfolding transitions in proteins can be gathered by analyzing the differential scanning calorimetry data in conjunction with the tertiary structure of proteins available from the X-ray crystallography.

Although the distributions of unfolding forces are slightly skewed and asymmetric, in Case Studies 1−5 carried out in this work we assumed that the probability density functions of unfolding forces can be described by the normal distribution $\sum_j \pi_j \mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$ specified in terms of the prior probabilities π_j (weights), the average forces μ_{fj}, and the standard deviations σ_{fj}. Our use of the normal distribution is fully justified for these proof-of-concept studies, and the extension of the developed formalism to describing asymmetric skewed distributions of unfolding forces is possible, but beyond the scope of this work. In addition, the EM algorithm converges rapidly numerically when it is used in conjunction with the normal distribution (∼10−20 steps). In Case Studies 2−5, we used the L^1-norm to quantify the difference between the "theoretical predictions" (pdfs of unfolding forces) and the "actual data" (histograms and nonparametric densities of unfolding forces). The L^1-norm is an upper-bounded metric, and it is widely used to compare a pair of distributions. There are other ways to quantify the difference between any two distributions, such as the L^2-norm or KL divergence, but these quantities do not have an upper bound. Also, the L^2-norm is dominated by contributions from large values of a random variable (*e.g.*, strong unfolding force signals).

In Case Study 1, we assessed the performance of SVM method in data classification for $(WW)_2$ and Fg (Figure 3) using only the simulation data, because in the simulations we can associate the unfolding events with the types of transitions they represent, and so the SVM predictions can be tested to assess performance. In Case Studies 2 and 3, we tested SVM and EM in a simpler problem of characterizing the statistics of unfolding forces for dimer $(WW)_2$. Here, the unfolding data for the training and test sets were extracted from the simulations. In Case Studies 4 and 5, we took a step further; we compared performance of SVM and EM at describing the experimental unfolding data for Fg monomer. Here, theoretical models were trained on the basis of the unfolding data from the simulated FX curves; these models were then applied to the unfolding data from the experimental FX spectra. The other steps of SVM and EM algorithms were same as in Case Studies 2 and 3. Because most of the experimental data (∼90−95%) are noisy force signals and in

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

order to increase the signal-to-noise ratio, in Case Study 4 for the Fg monomer we analyzed the unfolding forces <500 pN and in Case Study 5 we analyzed the unfolding forces in the 30−200 pN range. The unfolding forces >500 pN (Figure S3), which correlate with the peak-to-peak distances >100 nm, are due to protein desorption and/or cantilever tip detachment. The 100 nm distance is much longer than the average elongations of Fg monomer due to unfolding transitions of types 1, 2, or 3 (Table SII; see also ref 63). The unfolding force >200 pN, which correlates with long extension >60 nm, corresponds to a situation when several unfolding transitions in Fg monomer occur simultaneously. Indeed, the 60 nm Fg extension is longer than the sum of extensions for the unfolding transitions of types 1 and 2 (Table SII). Also, the unfolding forces <30 pN (Figure S3) are due to weak nonspecific interactions.

The prior probability π_j is the likelihood that a particular unfolding transition corresponds to the transition of type j. This quantity defines the weight of this transition type in an ensemble of all observations. In experiment, some of the unfolding transitions in a protein might not occur before protein desorption from the substrate takes place. Also, different types of unfolding transitions might not occur an equal number of times due to cantilever tip detachment. For example, in the simulations for dimer $(WW)_2$ the unfolding transitions of type 2 almost always occur last. The unfolding transitions of type 3 in Fg monomer is the last unfolding event, which occurs only after the unfolding transitions of types 1 and 2 took place. Suppose $(WW)_2$ desorption and/or cantilever tip detachment takes place right before the last unfolding transition of type 2 in $(WW)_2$. Then, the prior probabilities are $\pi_1 = 2/3$ and $\pi_2 = 1/3$ for the unfolding transitions of types 1 and 2, respectively. Furthermore, the strength of the desorption peak decreases as the length of a protein fragment adsorbed on the substrate surface decreases, and so for short protein fragments the strength of desorption peak might become comparable with the strength of the unfolding force. Hence, prior probabilities of observing the unfolding transitions of different types may or may not be equal to $1/J$, where J is the total number of types of unfolding transitions (*i.e.*, $J = 2$ for $(WW)_2$ and $J = 3$ for Fg). Therefore, in Case Studies 4 and 5 we tested EM both with fixed and variable prior probabilities π_j.

SVM classification of the unfolding transitions for $(WW)_2$ and Fg (Case Study 1) showed that the maximal margin classifier performs well even when the distributions of unfolding forces for different types of transitions overlap (Figure S2). For Fg monomer, errors are only ∼16−30% and the number of misclassified data points is 2−5 for unfolding transitions of types 1−3. This is a good agreement given a small sample size (M_{train} = 47; Figure 3b). For $(WW)_2$, the errors are 9−17%, but the unfolding transitions of types 1 and 2 are characterized by the strongly overlapping 37.4−77.8 and 42.6−82.2 pN force ranges, respectively (Figure S2). The lower errors might be due to larger sample size for $(WW)_2$ compared to Fg. Given these overlapping force ranges and small sample size of the training set (M_{train} = 200), 41 misclassified data points is a good achievement (Figure 3a). The errors are $L_1^1 = 0.24$ and $L_2^1 = 0.39$ for the unfolding transitions of types 1 and 2 for $(WW)_2$, and $L_1^1 = 0.14$, $L_2^1 = 0.21$, $L_3^1 = 0.20$ for the unfolding transitions of types 1, 2, and 3 for Fg. We applied SVM to map the distributions of unfolding forces for $(WW)_2$ (Case Study 2); here, we used maximal margin classier to separate unfolding forces in the test set into the unfolding transitions of types 1 and 2. The statistics of unfolding forces (π_j, μ_{fj}, and σ_{fj}) for the training data and test observations compare well (Figure 4, panels a−c), and the errors are small: $L_1^1 = 0.07$ and $L_2^1 = 0.09$ (Table 1). We applied EM to map the distributions of unfolding forces for $(WW)_2$ (Case Study 3). Here, we used nonparametric densities for the transitions of types 1 and 2 from training sets, maximal likelihood estimation to classify test observations, and EM optimization of π_j, μ_{fj}, and σ_{fj}. The statistics of unfolding forces (π_j, μ_{fj}, and σ_{fj}) for training and test observations show worse agreement (compared to SVM; Figure 4, panels d−f), and the errors are larger: $L_1^1 = L_2^1 = 0.2$ (Table 1). In Case Study 4, SVM outperformed EM (Figure 5), and the errors were $L_j^1 = 0.01−0.26$ for SVM *vs* $L_j^1 = 0.30−0.70$ for EM (Table 2). EM works better with variable prior probabilities (Table 2). In Case Study 5, SVM outperformed EM (Figure 6), and the errors were $L_j^1 = 0.05−0.12$ for SVM *vs* $L_j^1 = 0.14−0.37$ for EM (Table 3). Comparing the errors in Case Studies 4 and 5, EM performs better with variable prior probabilities especially when noisy data are excluded.

The SVM- and EM-based approaches tested in Case Studies 2−5 require the total number of different types of unfolding transitions J as an input. We remind that $J = 2$ for dimer $(WW)_2$ and $J = 3$ for Fg monomer. When this information is available from computational molecular modeling or other studies (crystal structures, differential scanning calorimetry data), both SVM and EM can be used with success to understand the unfolding force data (Figures 4−6). Furthermore, when experimental unfolding data contain more meaningful force signals (due to protein unfolding) and less noise, SVM and EM perform very well (Tables 2 and 3). The question is can one develop an approach that uses a minimal input (prior knowledge) from molecular modeling *in silico* or other studies? This problem can be overcome by using an EM-based approach described below, where only the total number of unfolding transition types J is specified (no information about the force ranges is necessary). In a simple implementation, one can set initial values for the average unfolding forces μ_j for the unfolding transitions of type j to be equal to the few largest experimental unfolding forces. For example, from the force histogram for $(WW)_2$ dimer, the two largest forces are 62 and 59 pN (see locations of the two tallest bins in Figure 4a). In many situations, the force histograms for combined data sets reveal broad unimodal distributions (as for dimer $(WW)_2$; see Figure 4), or bimodal or multimodal distributions (as for Fg monomer; see Figures 5 and 6). For bimodal and multimodal shapes, selecting initial values of the average forces μ_{fj} is simple; one can set μ_{f1}, μ_{f2}, ... for $j = 1, 2$, *etc.*, to be equal to the force modes. Initial values of the prior probabilities can be taken as $\pi_j = 1/J$, and initial values of the standard deviations σ_{fj} can be set to be equal to $\sigma_f/\sqrt{2}$ (σ_f is the standard deviation for combined data set).

Case Study 6. Resolving Distributions of Unfolding Forces without Data Classification and Data Division into Training and Test Observations. We used the output from the pulling simulations for dimer $(WW)_2$ (Figure 1) and for the Fg monomer (Figure 2) to mimic curated experimental unfolding force data (rather than "raw experimental data"), but now we apply EM directly to the unfolding forces from combined data sets $D_{(WW)_2}$ for $(WW)_2$ and D_{Fg} for Fg without maximum likelihood classification. Here, we use visual inspection of the shapes of experimental force histograms. We tested EM algorithm with fixed and variable prior probabilities for Fg. In the first case, we used fixed values $\pi_1 = 0.29$, $\pi_2 = 0.38$, and $\pi_3 = 0.33$, and in the second case, we varied the prior probabilities starting from $\pi_1 = 0.29$, $\pi_2 = 0.38$, and $\pi_3 = 0.33$. For

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

Table 4. Performance of EM in Describing Unfolding Forces for $(WW)_2$ and Fg without Data Classification[a]

	$(WW)_2$		Fg		
methods/quantities	type 1	type 2	type 1	type 2	type 3
μ_{fj}, pN	56.2 (56.1)/58.3	63.3 (63.2)/61.2	98.0 (90.5)/98.0	126.8 (125)/124.0	157.2 (157.8)/164.0
σ_{fj}, pN	6.4 (6.4)/7.4	7.1 (7.1)/7.8	26.7 (22.5)/20.3	14.8 (13.4)/15.3	13.3 (12.1)/13.8
π_j	0.5 (0.49)/0.5	0.5 (0.51)/0.5	0.29 (0.25)/0.29	0.38 (0.40)/0.38	0.33 (0.35)/0.33
L_j^1	0.13 (0.14)	0.13 (0.12)	0.13 (0.15)	0.07 (0.10)	0.08 (0.10)

[a]Compared for unfolding transitions of types 1 and 2 for $(WW)_2$ and types 1−3 for Fg are theoretical predictions for the average unfolding forces μ_{fj}, standard deviations σ_{fj}, and prior probabilities π_j obtained with EM and values of these quantities obtained from the pulling simulations (separated by the slash). The total error, L^1-norm, for combined dataset $D_{(WW)_2}$ for $(WW)_2$ is 0.10. The total error, L^1-norm, for combined dataset D_{Fg} for Fg is 0.15 (0.20) for fixed (variable) prior probabilities. The following data sets were used: $D_{(WW)_2,1}$ ($M_{(WW)_2,1}$ = 50), $D_{(WW)_2,2}$ ($M_{(WW)_2,2}$ = 50), D_{Fg} (M_{Fg} = 82), $D_{sim,1}$ ($M_{sim,1}$ = 24), $D_{sim,2}$ ($M_{sim,2}$ = 31), and $D_{sim,3}$ ($M_{sim,3}$ = 27).

$(WW)_2$, we set initial values to $\pi_1 = 0.5$ and $\pi_2 = 0.5$. We set initial values of μ_{fj} to be equal to the locations of the highest bins in the force histogram for combined data sets (Figure 7): μ_{f1} = 56 pN and μ_{f2} = 62 pN for $(WW)_2$ ($D_{(WW)_2}$) and μ_{f1} = 99 pN, μ_{f2} = 122 pN, and μ_{f3} = 144 pN for Fg (D_{Fg}). Initial values of μ_{xj} were set to be equal to the peak-to-peak distances from the 3D histogram of bivariate data (not shown) corresponding to these forces: μ_{x1} = 2.5 nm and μ_{x2} = 5.8 nm for $(WW)_2$ and μ_{x1} = 35 nm, μ_{x2} = 30 nm, and μ_{x3} = 20 nm for Fg. Initial values of σ_{fj} were set to be equal to $\sigma_{f,(WW)_2}/\sqrt{2}$ and $\sigma_{f,Fg}/\sqrt{2}$, where $\sigma_{f,(WW)_2}$ = 8.2 pN and $\sigma_{f,Fg}$ = 31.2 pN are the standard deviations for data from $D_{(WW)_2}$ and D_{Fg}. In all cases, the convergence of EM algorithm was reached in less than 20 steps. The histograms and nonparametric densities for combined data sets $D_{(WW)_2}$ and D_{Fg} are compared with the theoretical pdfs of unfolding forces for weighted superposition, $\sum_j \pi_j \mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$, and for each transition type, $\mathcal{N}(f_i|\mu_{fj}, \sigma_{fj})$, in Figure 7, which shows very good agreement both for $(WW)_2$ and Fg. The predicted values of π_j, μ_{fj}, and σ_{fj} are compared with actual values of these quantities in Table 4. The errors for fixed (variable) prior probabilities are L_1^1 = 0.13 (0.14) and L_2^1 = 0.13 (0.12) for $(WW)_2$ and L_1^1 = 0.13 (0.15), L_2^1 = 0.07 (0.10), and L_3^1 = 0.08 (0.10) for Fg (Table 4). The agreement is very good both for $(WW)_2$ and for Fg notwithstanding the small sample size ($M_{(WW)_2}$ = 100 and M_{Fg} = 82) and overlapping force ranges (Figure S2). Hence, the EM-based approach described above can be applied directly to the experimental data to resolve the distributions of unfolding forces.

CONCLUSIONS

To conclude, we developed and tested SVM- and EM-based approaches to statistical learning from single-molecule forced unfolding experiments. We showed that results from molecular modeling *in silico* can be used as a training set to construct the maximal margin classifier (with SVM) or maximum likelihood classifier (with EM). These can then be used to classify and model the experimental unfolding forces. An input from the computational molecular modeling is desirable for meaningful interpretation of complex forced unfolding data characterized by low signal-to-noise ratio. We also proposed a simple EM-based approach (Case Study 6), which can be applied directly to the experimental unfolding data. This approach uses information about the number of unfolding transitions of different types (available from the shape of experimental force histograms), but it does not involve data classification and data division into training and test observations. This approach performs well when the sample size is small and unfolding transitions of different types have overlapping force ranges. The developed SVM- and EM-based methods can be implemented in a single-molecule experimental setting to understand and model the protein unfolding data.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c02334.

SVM calculations, tables summarizing descriptions of unfolding transitions, and figures of forced unfolding transitions, nonparametric density estimates of the distributions of unfolding forces, and experimental unfolding data, (PDF)

AUTHOR INFORMATION

Corresponding Author

Valeri Barsegov − *Department of Chemistry, University of Massachusetts, Lowell, Massachusetts 01854, United States;* orcid.org/0000-0003-1994-3917; Phone: +1 978-934-3661; Email: Valeri_Barsegov@uml.edu

Authors

Farkhad Maksudov − *Department of Chemistry, University of Massachusetts, Lowell, Massachusetts 01854, United States*

Lee K. Jones − *Department of Mathematical Sciences, University of Massachusetts, Lowell, Massachusetts 01854, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c02334

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the NIH grant R01HL148227 and NSF grant MCB-2027530 to V.B.

REFERENCES

(1) Simpson, L. J.; Tzima, E.; Reader, J. S. Mechanical forces and their effect on the ribosome and protein translation machinery. *Cells* **2020**, *9*, 650.

(2) Maillard, R. A.; Chistol, G.; Sen, M.; Righini, M.; Tan, J.; Kaiser, C. M.; Hodges, C.; Martin, A.; Bustamante, C. ClpX (P) generates mechanical force to unfold and translocate its protein substrates. *Cell* **2011**, *145*, 459−469.

(3) Olivares, A. O.; Baker, T. A.; Sauer, R. T. Mechanical protein unfolding and degradation. *Annu. Rev. Physiol.* **2018**, *80*, 413−429.
(4) Schönfelder, J.; Alonso-Caballero, A.; De Sancho, D.; Perez-Jimenez, R. The life of proteins under mechanical force. *Chem. Soc. Rev.* **2018**, *47*, 3558−3573.
(5) Pollard, T. D.; Cooper, J. A. Actin, a central player in cell shape and movement. *Science* **2009**, *326*, 1208−1212.
(6) Lieleg, O.; Schmoller, K. M.; Cyron, C. J.; Luan, Y.; Wall, W. A.; Bausch, A. R. Structural polymorphism in heterogeneous cytoskeletal networks. *Soft Matter* **2009**, *5*, 1796−1803.
(7) Shah, E. A.; Keren, K. Mechanical forces and feedbacks in cell motility. *Curr. Opin. Cell Biol.* **2013**, *25*, 550−557.
(8) Jansen, K. A.; Atherton, P.; Ballestrem, C. Mechanotransduction at the cell-matrix interface. *Semin. Cell Dev. Biol.* **2017**, *71*, 75−83.
(9) Humphrey, J. D.; Dufresne, E. R.; Schwartz, M. A. Mechano-transduction and extracellular matrix homeostasis. *Nat. Rev. Mol. Cell Biol.* **2014**, *15*, 802−812.
(10) Somara, S.; Gilmont, R.; Bitar, K. N. Role of thin-filament regulatory proteins in relaxation of colonic smooth muscle contraction. *Am. J. Physiol. Liver Physiol.* **2009**, *297*, G958−G966.
(11) Rivas-Pardo, J. A.; Eckels, E. C.; Popa, I.; Kosuri, P.; Linke, W. A.; Fernández, J. M. Work done by titin protein folding assists muscle contraction. *Cell Rep.* **2016**, *14*, 1339−1347.
(12) Hessel, A. L.; Lindstedt, S. L.; Nishikawa, K. C. Physiological mechanisms of eccentric contraction and its applications: a role for the giant titin protein. *Front. Physiol.* **2017**, *8*, 70.
(13) Niederländer, N.; Raynaud, F.; Astier, C.; Chaussepied, P. Regulation of the actin–myosin interaction by titin. *Eur. J. Biochem.* **2004**, *271*, 4572−4581.
(14) Raynaud, F.; Astier, C.; Benyamin, Y. Evidence for a direct but sequential binding of titin to tropomyosin and actin filaments. *Biochim. Biophys. Acta, Proteins Proteomics* **2004**, *1700*, 171−178.
(15) Weisel, J. W. The mechanical properties of fibrin for basic scientists and clinicians. *Biophys. Chem.* **2004**, *112*, 267−276.
(16) Litvinov, R. I.; Weisel, J. W. Fibrin mechanical properties and their structural origins. *Matrix Biol.* **2017**, *60-61*, 110−123.
(17) Brown, A. E.; Litvinov, R. I.; Discher, D. E.; Purohit, P. K.; Weisel, J. W. Multiscale mechanics of fibrin polymer: gel stretching with protein unfolding and loss of water. *Science* **2009**, *325*, 741−744.
(18) Campbell, R. A.; Aleman, M. M.; Gray, L. D.; Falvo, M. R.; Wolberg, A. S. Flow profoundly influences fibrin network structure: Implications for fibrin formation and clot stability in haemostasis. *Thromb. Haemostasis* **2010**, *104*, 1281−1284.
(19) Flamm, M. H.; Diamond, S. L. Multiscale systems biology and physics of thrombosis under flow. *Ann. Biomed. Eng.* **2012**, *40*, 2355−2364.
(20) Purohit, P. K.; Litvinov, R. I.; Brown, A. E. X.; Discher, D. E.; Weisel, J. W. Protein unfolding accounts for the unusual mechanical behavior of fibrin networks. *Acta Biomater.* **2011**, *7*, 2374−2383.
(21) Mitsui, K.; Hara, M.; Ikai, A. Mechanical unfolding of a2-macroglobulin molecules with atomic force microscope. *FEBS Lett.* **1996**, *385*, 29−33.
(22) Zlatanova, J.; Lindsay, S. M.; Leuba, S. H. Single molecule force spectroscopy in biology using the atomic force microscope. *Prog. Biophys. Mol. Biol.* **2000**, *74*, 37−61.
(23) Clausen-Schaumann, M.; Seitz, H.; Krautbauer, R.; Gaub, H. E. Force spectroscopy with single bio-molecules. *Curr. Opin. Chem. Biol.* **2000**, *4*, 524−530.
(24) Hoffmann, T.; Dougan, L. Single molecule force spectroscopy using polyproteins. *Chem. Soc. Rev.* **2012**, *41*, 4781−4796.
(25) Barsegov, V.; Klimov, D. K.; Thirumalai, D. Mapping the energy landscape of biomolecules using single molecule force correlation spectroscopy: theory and applications. *Biophys. J.* **2006**, *90*, 3827−3841.
(26) Ashkin, A.; Dziedzic, J. M.; Yamane, T. Optical trapping and manipulation of single cells using infrared laser beams. *Nature* **1987**, *330*, 769−771.
(27) Zhang, X.; Ma, L.; Zhang, Y. High-resolution optical tweezers for single-molecule manipulation. *Yale J. Biol. Med.* **2013**, *86*, 367−383.
(28) Moffitt, J. R.; Chemla, Y. R.; Smith, S. B. Bustamante, Recent advances in optical tweezers. *Annu. Rev. Biochem.* **2008**, *77*, 205−228.
(29) Capitanio, M.; Pavone, F. S. Interrogating biology with force: single molecule high-resolution measurements with optical tweezers. *Biophys. J.* **2013**, *105*, 1293−1303.
(30) Smith, S. B.; Finzi, L.; Bustamante, C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **1992**, *258*, 1122−1126.
(31) Chen, L.; Offenhäusser, A.; Krause, H. J. Magnetic tweezers with high permeability electromagnets for fast actuation of magnetic beads. *Rev. Sci. Instrum.* **2015**, *86*, 044701.
(32) De Vlaminck, I.; Dekker, C. Recent advances in magnetic tweezers. *Annu. Rev. Biophys.* **2012**, *41*, 453−472.
(33) Le, S.; Liu, R.; Lim, C. T.; Yan, J. Uncovering mechanosensing mechanisms at the single protein level using magnetic tweezers. *Methods* **2016**, *94*, 13−18.
(34) Kilinc, D.; Lee, G. U. Advances in magnetic tweezers for single molecule and cell biophysics. *Integr. Biol.* **2014**, *6*, 27−34.
(35) Neuman, K. C.; Nagy, A. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat. Methods* **2008**, *5*, 491−505.
(36) Svoboda, K.; Schmidt, C. F.; Schnapp, B. J.; Block, S. M. Direct observation of kinesin stepping by optical trapping interferometry. *Nature* **1993**, *365*, 721−727.
(37) Finer, J. T.; Simmons, R. M.; Spudich, J. A. Single myosin molecule mechanics: piconewton forces and nanometre steps. *Nature* **1994**, *368*, 113−119.
(38) Puchner, E. M.; Alexandrovich, A.; Kho, A. L.; Hensen, U.; Schafer, L. V.; Brandmeier, B.; Grater, F.; Grubmuller, H.; Gaub, H. E.; Gautel, M. Mechanoenzymatics of titin kinase. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 13385−13390.
(39) Yusko, E. C.; Asbury, C. L. Force is a signal that cells cannot ignore. *Mol. Biol. Cell* **2014**, *25*, 3717−3725.
(40) Vogel, V. Mechanotransduction involving multimodular proteins: converting force into biochemical signals. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 459−488.
(41) Chakrabarti, S.; Hinczewski, M.; Thirumalai, D. Plasticity of hydrogen bond networks regulates mechanochemistry of cell adhesion complexes. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 9048−9053.
(42) Olivares, A. O.; Nager, A. R.; Iosefson, O.; Sauer, R. T.; Baker, T. A. Mechanochemical basis of protein degradation by a double-ring AAA + machine. *Nat. Struct. Mol. Biol.* **2014**, *21*, 871.
(43) Aubin-Tam, M. E.; Olivares, A. O.; Sauer, R. T.; Baker, T. A.; Lang, M. J. Single-molecule protein unfolding and translocation by an ATP-fueled proteolytic machine. *Cell* **2011**, *145*, 257−267.
(44) Zhmurov, A.; Dima, R. I.; Barsegov, V. Order statistics theory of unfolding of multimeric proteins. *Biophys. J.* **2010**, *99*, 1959−1968.
(45) Kononova, O.; Jones, L.; Barsegov, V. Order statistics inference for describing topological coupling and mechanical symmetry breaking in multidomain proteins. *J. Chem. Phys.* **2013**, *139*, 121913.
(46) Bura, E.; Klimov, D. K.; Barsegov, V. Analyzing forced unfolding of protein tandems by ordered variates, 1: Independent unfolding time. *Biophys. J.* **2007**, *93*, 1100−1115.
(47) Bura, E.; Klimov, D. K.; Barsegov, V. Analyzing forced unfolding of protein tandems by ordered variates, 2: dependent unfolding times. *Biophys. J.* **2008**, *94*, 2516−2528.
(48) Brockwell, D. J.; Beddard, G. S.; Clarkson, J.; Zinober, R. C.; Blake, A. W.; Trinick, J.; Olmsted, P. D.; Smith, D. A.; Radford, S. E. The effect of core destabilization on the mechanical resistance of I27. *Biophys. J.* **2002**, *83*, 458−472.
(49) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*; Springer: New York, 2013.
(50) Baldi, P.; Brunak, S.; Bach, F. *Bioinformatics: the machine learning approach*; MIT Press: Cambridge, MA, 2001.
(51) Libbrecht, M. W.; Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321−332.
(52) Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of compounds for machine-learning prediction of

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808

physical properties. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 144110.
(53) Lookman, T., Eidenbenz, S., Alexander C, F., Eds. *Barnes, Materials discovery and design: By means of data science and optimal learning*; Springer, 2018.
(54) Ryan, K.; Lengyel, J.; Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **2018**, *140*, 10158−10168.
(55) Furmanchuk, A. O.; Agrawal, A.; Choudhary, A. Predictive analytics for crystalline materials: bulk modulus. *RSC Adv.* **2016**, *6*, 95246−95251.
(56) Noordik, J. H. *Cheminformatics Developments: History, Reviews and Current Research*; IOS Press: Amsterdam, 2004.
(57) Ziatdinov, M.; Dyck, O.; Maksov, A.; Li, X.; Sang, X.; Xiao, K.; Unocic, R. R.; Vasudevan, R.; Jesse, S.; V Kalinin, S. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. *ACS Nano* **2017**, *11*, 12742−12752.
(58) Jäger, M.; Nguyen, H.; Crane, J. C.; Kelly, J. W.; Gruebele, M. The folding mechanism of a β-sheet: the WW domain. *J. Mol. Biol.* **2001**, *311*, 373−393.
(59) Ferguson, N.; Berriman, J.; Petrovich, M.; Sharpe, T. D.; Finch, J. T.; Fersht, A. R. Rapid amyloid fiber formation from the fast-folding WW domain FBP28. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 9814−9819.
(60) Karanicolas, J.; Brooks, C. L. The structural basis for biphasic kinetics in the folding of the WW domain from a formin-binding protein: Lessons for protein design? *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 3954−3959.
(61) Cheung, M. S.; Klimov, D.; Thirumalai, D. Molecular crowding enhances native state stability and refolding rates of globular proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 4753−4758.
(62) Kollman, J. M.; Pandi, L.; Sawaya, M. R.; Riley, M.; Doolittle, R. F. Crystal structure of human fibrinogen. *Biochemistry* **2009**, *48*, 3877−3886.
(63) Zhmurov, A.; Brown, A. E. X.; Litvinov, R. I.; Dima, R. I.; Weisel, J. W.; Barsegov, V. Mechanism of fibrin(ogen) forced unfolding. *Structure* **2011**, *19*, 1615−1624.
(64) Hyeon, C.; Dima, R. I.; Thirumalai, D. Pathways and Kinetic Barriers in Mechanical Unfolding and Refolding of RNA and Proteins. *Structure* **2006**, *14*, 1633−1645.
(65) Zhmurov, A.; Dima, R. I.; Barsegov, V. Order statistics theory of unfolding of multimeric proteins. *Biophys. J.* **2010**, *99*, 1959.
(66) Weisel, J. W. Fibrinogen and fibrin. *Adv. Protein Chem.* **2005**, *70*, 247.
(67) Mickler, M.; Dima, R. I.; Dietz, H.; Hyeon, C.; Thirumalai, D.; Rief, M. Revealing the bifurcation in the unfolding pathways of GFP by using single-molecule experiments and simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 20268−20273.
(68) Dima, R. I.; Joshi, H. Probing the origin of tubulin rigidity with molecular simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 15743−15748.
(69) Koepf, E. K.; Petrassi, H. M.; Sudol, M.; Kelly, J. W. WW: An isolated three-stranded antiparallel β-sheet domain that unfolds and refolds reversibly; evidence for a structured hydrophobic cluster in urea and GdnHCl and a disordered thermal unfolded state. *Protein Sci.* **1999**, *8*, 841−853.
(70) Bura, E.; Zhmurov, A.; Barsegov, V. Nonparametric density estimation and optimal bandwidth selection for protein unfolding and unbinding data. *J. Chem. Phys.* **2009**, *130*, 015102.
(71) Silverman, B. W. *Density estimation for statistics and data analysis*; CRC Press, 1986.
(72) Scott, D. W. *Multivariate density estimation: theory, practice, and visualization*; John Wiley & Sons, 2015.
(73) Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*; Springer Science & Business Media, 2009.
(74) McLachlan, G. J.; Krishnan, T. *The EM algorithm and extensions*, second ed.; John Wiley & Sons, 2007.

https://doi.org/10.1021/acs.jpcb.1c02334
J. Phys. Chem. B 2021, 125, 5794−5808