Title

Improving estimation of kinetic parameters in dynamic force spectroscopy using cluster analysis

Journal

The Journal of Chemical Physics, 148(12)

ISSN

0021-9606

Authors

Yen, Chi-Fu
Sivasankar, Sanjeevi

Publication Date

2018-03-28

DOI

10.1063/1.5001325

Peer reviewed

Improving estimation of kinetic parameters in dynamic force spectroscopy using cluster analysis

Chi-Fu Yen[1] and Sanjeevi Sivasankar[1,2,a)]
[1]*Department of Electrical and Computer Engineering, Iowa State University, Ames, Iowa 50011, USA*
[2]*Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, USA*

(Received 12 May 2017; accepted 27 June 2017; published online 21 September 2017)

Dynamic Force Spectroscopy (DFS) is a widely used technique to characterize the dissociation kinetics and interaction energy landscape of receptor-ligand complexes with single-molecule resolution. In an Atomic Force Microscope (AFM)-based DFS experiment, receptor-ligand complexes, sandwiched between an AFM tip and substrate, are ruptured at different stress rates by varying the speed at which the AFM-tip and substrate are pulled away from each other. The rupture events are grouped according to their pulling speeds, and the mean force and loading rate of each group are calculated. These data are subsequently fit to established models, and energy landscape parameters such as the intrinsic off-rate (k_{off}) and the width of the potential energy barrier (x_β) are extracted. However, due to large uncertainties in determining mean forces and loading rates of the groups, errors in the estimated k_{off} and x_β can be substantial. Here, we demonstrate that the accuracy of fitted parameters in a DFS experiment can be dramatically improved by sorting rupture events into groups using cluster analysis instead of sorting them according to their pulling speeds. We test different clustering algorithms including Gaussian mixture, logistic regression, and K-means clustering, under conditions that closely mimic DFS experiments. Using Monte Carlo simulations, we benchmark the performance of these clustering algorithms over a wide range of k_{off} and x_β, under different levels of thermal noise, and as a function of both the number of unbinding events and the number of pulling speeds. Our results demonstrate that cluster analysis, particularly K-means clustering, is very effective in improving the accuracy of parameter estimation, particularly when the number of unbinding events are limited and not well separated into distinct groups. Cluster analysis is easy to implement, and our performance benchmarks serve as a guide in choosing an appropriate method for DFS data analysis. *Published by AIP Publishing.* https://doi.org/10.1063/1.5001325

INTRODUCTION

Dynamic force spectroscopy (DFS) experiments are widely used to characterize the dissociation kinetics and interaction energy landscape of protein-protein interactions,[1–3] DNA-protein binding,[4,5] and aggregation of misfolded proteins.[6,7] While these measurements can be performed using different micromanipulation tools including atomic force microscopy (AFM), micro-needle manipulation, optical tweezers, and magnetic tweezers,[8] AFM-based DFS experiments are widely used because of their sub-nanometer spatial resolution.[8]

In a typical DFS experiment, an AFM cantilever and substrate functionalized with flexible polymer linkers are decorated with the biomolecules of interest (Fig. 1).[9] The functionalized AFM tip and substrate are brought into contact, enabling opposing molecules to interact, and then pulled apart at a range of pulling speeds. The force applied to the protein complex is sensed by the deflection of the cantilever while the rate of applied force (the loading rate) is controlled by varying the separation-speed of the AFM tip and substrate. Histograms of rupture forces for each pulling speed are plotted to determine the most probable unbinding force; from the dependence of the rupture forces on loading rates, the energy landscape parameters of the system can be predicted.[10,11] In the widely used single barrier model, the intrinsic off-rate under zero force, k_{off}, and the width of energy barrier that inhibit protein dissociation, x_β, are determined by fitting the most probable force at different loading rates to the Bell-Evans model,

$$F^*(r) = F_\beta \ln(r/(k_{off} F_\beta)), \tag{1}$$

where $F^*(r)$ is the most probable unbinding force, r is the loading rate, $F_\beta = k_B T/x_\beta$, k_B is Boltzmann's constant, and T is the absolute temperature.[10,12] To increase the quality of the fit, several pulling speeds are used so that the loading rates cover a large dynamic range.[13,14]

In order to measure single molecule binding, DFS experiments are typically designed such that the chance of observing a specific unbinding event is less than 10%.[15] Under these conditions, collecting enough events to recover the unbinding force distribution and accurately estimate the most probable unbinding forces before the sample degrades is often impractical. Consequently, the mean or median rupture force is commonly used for data analysis instead of the most probable force.[16–18] Alternatively, the most probable force is

a)Author to whom correspondence should be addressed: sivasank@iastate.edu. Tel.: (515) 294-1220; Fax: (515) 294-6027.

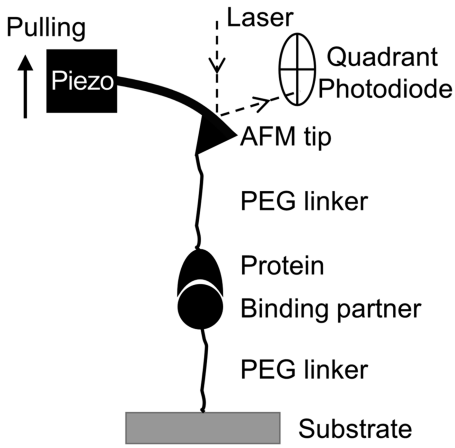

FIG. 1. Schematic of the experimental setup for an AFM based DFS measurement (*not to scale*). A receptor and its binding partner are tethered to the AFM tip and substrate via poly(ethylene glycol) (PEG) linkers. The receptor ligand complex is ruptured by translating the tip away from the substrate with a piezoelectric actuator. During this process, force and tip-surface distance are recorded.

determined by fitting the force histograms to a Gaussian distribution.[19] However, due to thermal fluctuations, heterogeneity of chemical bonds, and "contaminating" multiple unbinding and nonspecific adhesion events, the measured force distribution often varies from the theoretical model which decreases the accuracy of parameter estimation using the simplified mean or median force methods.[20,21] This distortion is most pronounced in the prediction of k_{off}, where errors are exponentiated with uncertainties of the same order of magnitude as the estimated value.[22] Different methods have been proposed to improve the accuracy of the estimated k_{off} and x_β, such as fitting the force and loading rate distribution with a probability density function,[23] introducing correction algorithms,[24] or by considering the force dependence of molecular oscillation frequencies.[25] However, a simple high-accuracy method to improve parameter estimation in DFS experiments, which retains the simplicity of using mean forces, is still lacking.

To overcome this bottleneck, we use cluster analysis to group unbinding events and improve the accuracy of fitted k_{off} and x_β in a typical DFS experiment. Cluster analysis is a widely used technique in identifying specific patterns from a large database, such as determining biologically relevant genes in microarray experiments,[26] identifying similar behaviors in marketing research,[27] and pattern recognition in computer vision.[28] Here, we use three clustering models: Gaussian mixture, logistic regression, and K-means to group single molecule unbinding events in DFS and to identify the most representative forces and loading rates for subsequent fitting. We simulated experimental data within a realistic range of k_{off}, x_β, thermal noise, number of pulling speeds, and number of events by performing Monte Carlo simulations. The simulated data were analyzed using both conventional analysis and cluster analysis, and the performances of different methods were compared. We show that clustering algorithms greatly improve the estimation of k_{off} and x_β, even when the amount of data is limited and where the unbinding events measured at multiple pulling speeds are not well separated from each other. Although our simulated data were analyzed using the classic Bell-Evans model, clustering analysis can be easily applied to other DFS models described in the literature.[29–31]

METHODS

Force-distance curve simulation

When a receptor-ligand complex is ruptured by withdrawing the cantilever away from the substrate using a piezoelectric actuator (Fig. 1), force-distance (FD) curves are the primary output of the measurement. We therefore simulated unbinding events as FD curves at a range of loading rates (Fig. 2). Our model parameters were chosen to relate the model in Fig. 1 to a realistic DFS experiment. We assumed that the receptor and ligand were immobilized on an AFM tip and substrate through polyethylene glycol (PEG; MW: 3400 Dalton) linkers. The spring constant of the cantilever was set to 40 pN/nm since soft probes with 10–100 pN/nm stiffness are usually used to measure weak biological interactions.[13] The measurements were simulated to occur at 25 °C with k_BT equal to 4.1 pN nm throughout the study. In order to mimic a realistic DFS experiment where loading rates usually span only two to three orders of magnitude,[13] we fixed the lower and upper bounds on loading rate to be 2000 and 10^6 pN/s.

We first calculated the probability distribution of rupture forces, $p(F)$, at a given loading rate, r, using the Bell-Evans model,[10]

$$p(F) = \frac{k_{off}}{r} \exp\left[\frac{F}{F_\beta} - \frac{k_{off} F_\beta}{r}(e^{F/F_\beta} - 1)\right]. \quad (2)$$

Since both the receptor and ligand were immobilized on the tip and substrate using flexible PEG linkers, a non-linear stretching of PEG tethers should be measured in each FD curve. We simulated the PEG stretching using the extended freely jointed chain model,[32]

$$D(F) = L_C \times \left(\coth\left(\frac{FL_K}{k_BT}\right) - \frac{k_BT}{FL_K}\right) + \frac{FL_C}{L_m K_s}, \quad (3)$$

where L_C and L_K are the contour length and Kuhn length of the PEG tethers and L_m and K_S are the average length and stiffness of a PEG monomer. Based on previous studies,[32] we used a value of 43.6898 nm for L_C, 0.7 nm for L_K, 0.2837 nm for L_m, and 150000 pN/nm for K_S. We simulated FD traces using Eq. (3) at 500 kHz data acquisition rate, with an unbinding force randomly sampled from the probability distribution of force [Eq. (2)] [Figs. 2(a) and 2(b)]. Since the FD curves are simulated by randomly sampling a rupture force from the probability distribution then attaching a linker spring to it, interactions between the PEG tethers and potential of the bond are assumed to be negligible.

Since the thermal fluctuations of the cantilever, which are detected by the AFM's Quadrant Photodiode (QPD) voltage, couple in as noise in both the measured force and the calculated tip-surface distance, we calculated the QPD voltage at each time point of the FD curve and calculated the noise in both force and tip-surface distance, using an optical lever sensitivity of 30 nm/V to correlate changes in the QPD

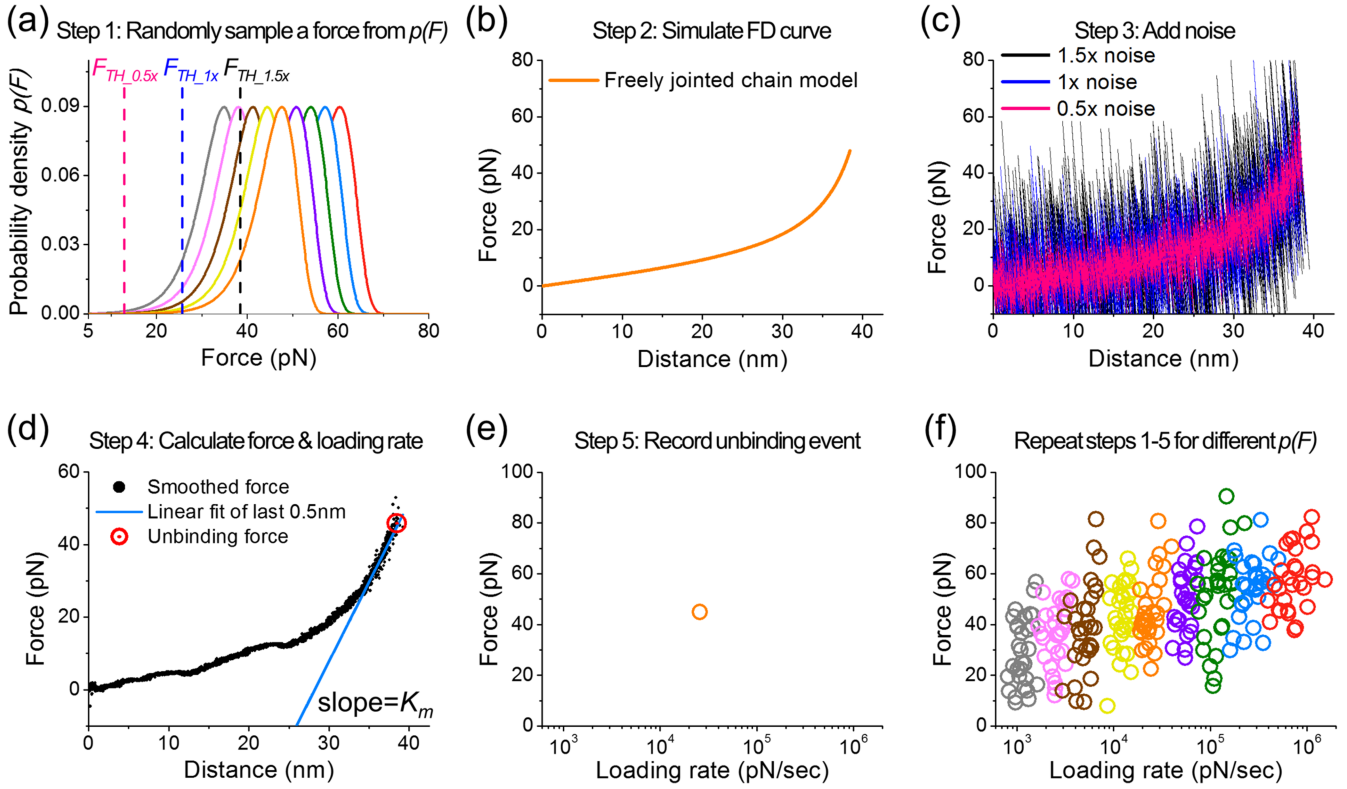

FIG. 2. Workflow for simulation of single molecule unbinding events. (a) An unbinding force greater than the force threshold (F_{TH}) was randomly sampled from the probability distribution of force. F_{TH} was set to 12.8 pN ($F_{TH\_0.5\times}$), 25.6 pN ($F_{TH\_1\times}$), and 38.4 pN ($F_{TH\_1.5\times}$) for conditions with 0.5×, 1×, and 1.5× thermal noise, respectively. (b) For the sampled unbinding force, a Force-Distance (FD) curve was simulated at 500 kHz using the freely jointed chain model to account for the stretching of PEG linkers. (c) The noise in force and distance due to thermal vibrations of the AFM cantilever was added to the FD curve at each time point. The calculated noise was normally distributed with standard deviations of 6.4 pN, 12.8 pN, or 19.2 pN in force and 0.16 nm, 0.32 nm, or 0.48 nm in distance, for conditions with 0.5× noise, 1× noise, and 1.5× noise, respectively. (d) To determine the unbinding force and loading rate, we first smoothed the noisy FD curve using a 4 nm moving average window and estimated the spring constant of the molecule, K_m, by fitting the last 0.5 nm data to a straight line. The loading rate was calculated by substituting K_m into Eq. (4). The last force reading was used as unbinding force. (e) The calculated force and loading rate for each FD curve were recorded. (f) By repeating the process described in panels (b)–(e), the unbinding events for a DFS measurement were simulated. Colors represent different pulling speeds.

voltage to cantilever fluctuations.[33] Since noise varies with factors such as AFM design, quality factor of cantilever, and environmental noise, we considered three levels of thermal noise in our simulations: 0.5×, 1×, or 1.5×. We accounted for these three thermal noise levels by adding normally distributed noise with standard deviations of 6.4 pN, 12.8 pN, or 19.2 pN in force and 0.16 nm, 0.32 nm, or 0.48 nm in distance to the FD curve [Fig. 2(c)].[34] In an actual DFS experiment, unbinding forces lower than a force threshold (F_{TH}), which depends on the level of noise, cannot be detected. We accounted for this in our simulations, by setting an F_{TH} value of 12.8 pN, 25.6 pN, or 38.4 pN for conditions with 0.5×, 1×, or 1.5× thermal noise and only sampled forces greater than F_{TH} [Fig. 2(a)].

Next, we estimated the loading rate for each FD curve, by modeling the cantilever and PEG linker as two springs that were pulled in series. While the spring constant of cantilever, K_C, was fixed, the spring constant of the PEG linker, K_m, was calculated as the slope of the tangent line to the FD curve at the unbinding force. Consequently, the loading rate, r, was calculated as

$$r = V_{pulling}(K_C K_m/(K_C + K_m)), \quad (4)$$

where $V_{pulling}$ is the pulling speed. We smoothed the noisy FD trace using a moving 4 nm window and estimated the spring constant of molecule, K_m, by fitting the last 0.5 nm data to a straight line [Fig. 2(d)]. The loading rate was determined by substituting K_m, K_C, and $V_{pulling}$ into Eq. (4). To simulate the dataset for a DFS experiment, we generated FD curves for different pulling speeds and recorded their rupture forces (last force reading in the FD curve) and loading rates [Figs. 2(e) and 2(f)].

Calculation of k_{off} and x_β

The simulated rupture events were sorted into groups using four methods (described below); the number of groups was limited to be the number of pulling speeds. To extract k_{off} and x_β, we determined the mean force and loading rate of each group and then fitted the mean force *vs.* loading rate to the Bell-Evans model [Eq. (1)] using a nonlinear least-squares fitting with bisquare weights. Simulations were repeated 100 times for each condition, and k_{off} and x_β were calculated for each simulation. Relative error in x_β was calculated as $[median(calculated\ x_\beta) - (preset\ x_\beta)]/[preset\ x_\beta]$. Relative error in k_{off} was calculated as $[e^{median(\ln(calculated\ k_{off}))} - (preset\ k_{off})]/[preset\ k_{off}]$. The algorithms for clustering have been derived in Ref. 28; Matlab code used for separating rupture events into groups was directly adopted from

Ref. 35 without modification. The methods we used to group data include the following:

Method 1: Pulling speed. This is the standard method in DFS data analysis where unbinding events with the same pulling speed are grouped together.

Method 2: Clustering using 2D Gaussian mixture model (GMM). Forces and loading rates were normalized for GMM since the ranges they span can be dramatically different. Each unbinding force F_i was normalized using $(F_i - F_{min})/(F_{max} - F_{min})$, where F_{max} and F_{min} are the maximum and minimum values for force. Loading rate r_i was normalized using $(\ln(r_i) - \ln(r_{min}))/(\ln(r_{max}) - \ln(r_{min}))$, where r_{max} and r_{min} are the maximum and minimum values for loading rate. As an initial guess for classification, events were grouped according to their loading rates; groups were assigned with equal number of events.

Method 3: Clustering using logistic regression model. Data were normalized as in *Method 2* and the initial guess was used as the training dataset for 2D logistic regression clustering.

Method 4: Clustering using 1D K-means. Events were separated into groups based on the normalized loading rates. The initial guess was the same as in GMM.

While k_{off} and x_β can be also extracted by directly fitting the entire data cloud to the Bell-Evans model, without sorting into groups, previous studies have shown that this fitting method results in large errors.[23] Consequently, we did not pursue cloud fitting in our study.

RESULTS AND DISCUSSION

Overview of cluster analysis

Cluster analysis is the process of sorting data into different groups such that events within the same category share similar characteristics. The main idea in applying this approach to a DFS experiment is that when a specific interaction is probed repeatedly using the same tip-sample pulling speed, the measured unbinding forces and loading rates are expected to be similar within a certain noise level. Therefore, unbinding events are expected to form clusters on a force versus loading rate plot. The mean forces and loading rates calculated from the clustered events share common characteristics such that the influence of outliers are reduced. To test this idea, we simulated unbinding events to closely mimic a realistic DFS experiment and grouped the events either using cluster analysis algorithms or according to pulling speed (the standard DFS analysis method where the unbinding events are grouped according to the tip-surface retraction speeds). We used three clustering analysis approaches in this work: Gaussian mixture model (GMM), logistic regression, and K-means (*methods*).[28,36]

Our rationale in using GMM[28] is that since the rupture force distribution at a constant pulling speed resembles a skewed Gaussian with a long tail at low forces,[10] the distribution of forces collected with many pulling speeds can be approximated as a mixture of Gaussian distributions. GMM was used to assign each unbinding event to a group by maximizing the posterior probability that the data point belongs to its assigned cluster such that the force distribution within each group is most likely to be a Gaussian.[28] In contrast to GMM, logistic regression identifies the boundaries between groups based on a training dataset, while K-means partitions data into clusters by minimizing the distance from the data point to the mean of its assigned cluster.[28,36] The theory and mathematical derivation of these methods are beyond the scope of this study; we merely adopt these clustering algorithms as data-analysis tools.

Overlap of data increases with increasing k_{off}, x_β, noise, number of unbinding events, and number of pulling speeds

Since the goal of cluster analysis is to partition data into groups by relocating ambiguous events at the boundaries into their proper categories, this approach is beneficial when unbinding events across multiple pulling speeds overlap. To generate overlapping datasets, we first examined how each parameter in our simulation affects data overlap (Fig. 3).

With a fixed range of loading rates, one would intuitively expect data overlap to increase with the number of pulling speeds, the number of unbinding events, and the level of

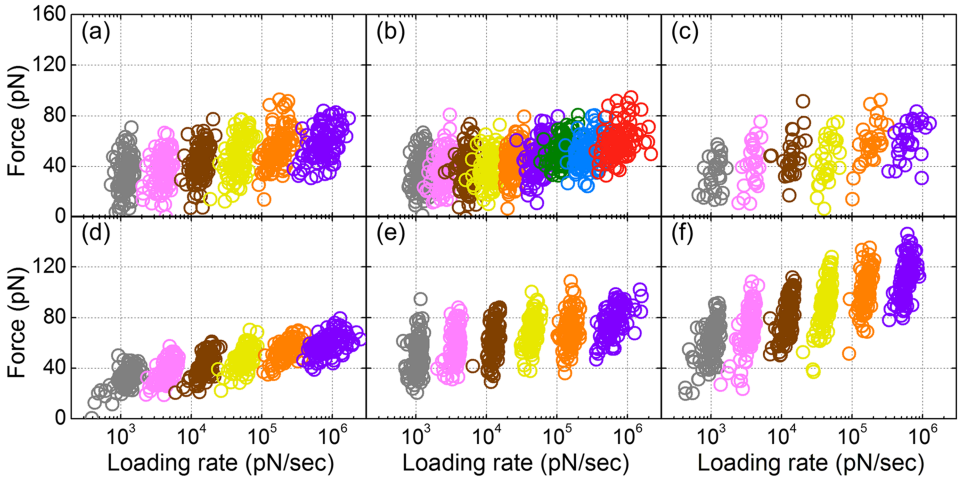

FIG. 3. Unbinding events across multiple pulling speeds overlap when the number of pulling speeds, number of data points, and thermal noise increases. (a) Data were simulated using k_{off} = 0.1 s^{-1}, x_β = 1 nm, 6 pulling speeds, 100 events per pulling speed with 1× thermal noise. Colors represent different speeds. (b)–(f) Similar to panel (a), unbinding events were simulated while varying only one parameter at a time. Simulations were carried out with (b) 9 pulling speeds; (c) 30 events per pulling speed; (d) 0.5× thermal noise; (e) k_{off} = 0.001 s^{-1}; (f) x_β = 0.5 nm. Each circle represents one unbinding event.

thermal noise. To confirm this, we first simulated a DFS experiment consisting of 600 unbinding events across 6 pulling speeds, using k_{off} = 0.1 s^{-1}, x_β = 1 nm, and 1× thermal noise [Fig. 3(a); *methods*]. Fixing all the other parameters in the simulation, we increased the number of pulling speeds [9 pulling speeds, Fig. 3(b)], decreased the number of events per pulling speed [30 events per pulling speed, Fig. 3(c)], decreased the noise level [0.5× thermal noise, Fig. 3(d)], reduced k_{off} [k_{off} = 0.001 s^{-1}, Fig. 3(e)], and reduced x_β [x_β = 0.5 nm, Fig. 3(f)]. As expected, the degree of data overlap increased with the number of pulling speeds, number of unbinding events, and a higher noise level [Figs. 3(a)–3(d)]. The simulated data also showed an increasing overlap as k_{off} and x_β increased [Figs. 3(a), 3(e), and 3(f)].

Cluster analysis improves the estimation of k_{off} and x_β

Next, we compared the accuracy of different cluster analysis methods and the standard pulling speed method on the estimation of k_{off} and x_β. An example dataset containing 270 simulated unbinding events evenly distributed over 9 tip-sample separation speeds with 1× thermal noise using k_{off} = 0.1 s^{-1} and x_β = 1 nm is shown in Fig. 4. The events were separated into 9 groups using either pulling speeds, GMM, logistic regression, or K-means (Fig. 4; *methods*). The most probable unbinding forces and loading rates were calculated by averaging data within each group (Fig. 4, red squares); k_{off} and x_β were then extracted by fitting those mean values to the Bell-Evans equation (Fig. 4, red lines). We performed simulations where only one parameter (either the thermal noise, number of unbinding events, number of pulling speeds, k_{off}, or x_β) was varied at a time, and the accuracy of different grouping methods on k_{off} and x_β estimation was compared. When the thermal noise, number of data points, or number of pulling speeds was varied, k_{off} and x_β were fixed at 0.1 s^{-1} and 1 nm, respectively. When k_{off} or x_β were varied, 9 pulling speeds and 30 unbinding events per speed with 1× thermal noise were used. Simulations for each condition were repeated 100 times, and the statistical distribution of estimated k_{off} and x_β was plotted (Figs. 5 and 6). In the following discussion, we focus on estimated errors in k_{off} since the estimation of x_β is accurate within 10% error in all analysis.

First, we tested k_{off} estimation using different grouping methods when the number of pulling speeds was varied (6, 9, and 12 pulling speeds). Our data showed that as the number of pulling speeds was increased, cluster analysis improved k_{off} estimation [Fig. 5(a)]. While relative errors in k_{off} using the standard pulling speed method were between −53% and −63% across all conditions, K-means analysis reduced the relative error from −43% at 6 pulling speeds to −11% and −8% at 9 and 12 pulling speeds, respectively. Logistic regression also decreased the error from −46% to −24% and −27% as the number of tip-surface retraction speeds was increased. In contrast, GMM showed a less significant improvement, with errors of −55%, −36%, and −42% at 6, 9, and 12 pulling speeds, respectively.

Next, we measured k_{off} estimation when the number of data points in each group equaled 10, 30, and 100 unbinding events. Our results showed that k_{off} estimation using the K-means method was superior, even when the amount of data was limited [Fig. 5(b)]. When individual groups had 10, 30, and 100 rupture events, the relative errors using K-means clustering were −9%, −11%, and −27% which was significantly lower than errors of −50%, −54%, and −58% measured with the standard pulling speed analysis. The logistic regression model also showed better accuracy, with errors of −24%, −24%, and −37% while GMM showed a less significant improvement, with errors of −51%, −36%, and −51% at 10 events, 30 events, and 100 events per group, respectively. Interestingly, while increasing the amount of data mainly increased precision, k_{off} accuracy did not increase.

The accuracy of k_{off} estimation decreased with an increase in thermal noise using all grouping methods [Fig. 5(c)]. When the force due to thermal fluctuations of the cantilever was 6.4 pN, 12.8 pN, and 19.2 pN (0.5×, 1×, and 1.5× thermal

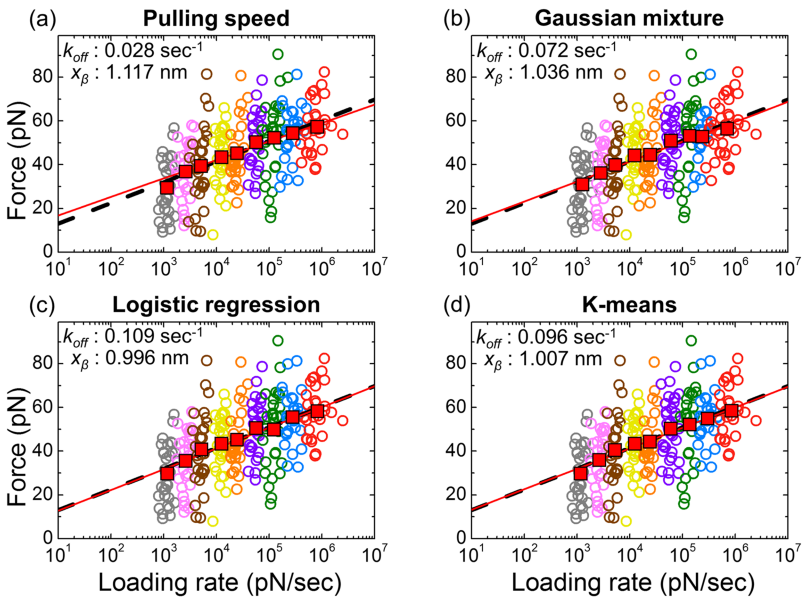

FIG. 4. Visual comparison of different clustering algorithms used. Unbinding events were simulated using k_{off} = 0.1 s^{-1}, x_β = 1 nm, 9 pulling speeds, 30 events per pulling speed with 1× thermal noise. Events were classified into different groups based on (a) pulling speeds; (b) Gaussian mixture cluster model; (c) logistic regression clustering; (d) K-means clustering. The results of grouping are indicated by colors. The average forces and loading rates (red squares) within each group were fit to the Bell-Evans model (red line) to estimate k_{off} and x_β. The estimated values are indicated. Black line represents the plot of the Bell-Evans equation using k_{off} = 0.1 s^{-1}, x_β = 1 nm. Differences between the different methods were observed at the boundaries between groups.

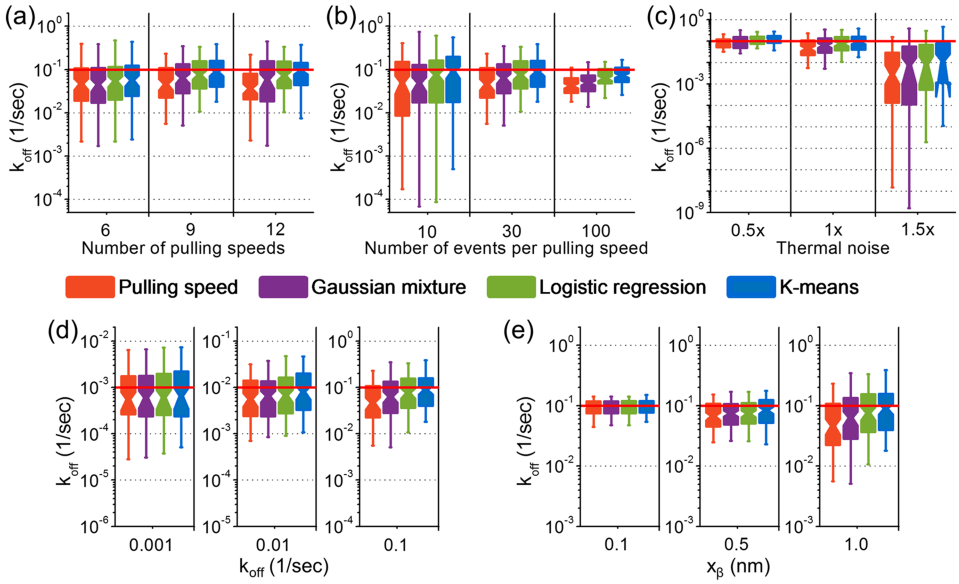

FIG. 5. Evaluation of cluster analysis for estimating k_{off} in DFS experiments. DFS data were generated using Monte Carlo simulations with k_{off} = 0.1 s^{-1}, x_β = 1 nm, 9 pulling speeds, 30 events per pulling speed, and 1× thermal noise. Four classification methods (pulling speeds, Gaussian mixture cluster model, logistic regression clustering, and K-means clustering) were used, and kinetic parameters were estimated using the Bell-Evans model. Simulations were repeated 100 times; the distribution of estimated k_{off} is plotted. Red line represents the preset values of k_{off}. The performance of methods was compared by varying (a) the number of pulling speeds, (b) the amount of data, (c) the thermal noise, (d) k_{off}, and (e) x_β. Cluster analysis significantly improved k_{off} estimation when the number of pulling speeds, number of unbinding events, and the thermal noise increased. Cluster analysis was especially accurate at high dissociation rates and wide energy barriers.

noise), the relative errors in estimated k_{off} using the standard pulling speed method (−15%, −54%, and −97%), k-means analysis (18%, −11%, and −88%), logistic regression (14%, −24%, and −92%), and GMM (−1%, −36%, and −92%) were comparable.

Importantly, our data showed that cluster analysis was especially useful for studying molecular interactions with high dissociation rates and wide energy barriers [Figs. 5(d) and 5(e)]. At dissociation rates of 10^{-3} s^{-1}, 0.01 s^{-1}, and 0.1 s^{-1}, the relative error in k_{off} estimated by K-means was −34%, −19%, and −11%, respectively. The relative error using logistic regression was comparable to K-means with values of −39% for 10^{-3} s^{-1}, −32% for 0.01 s^{-1}, and −24% for 0.1 s^{-1}, respectively. In contrast, k_{off} calculated using both GMM and the

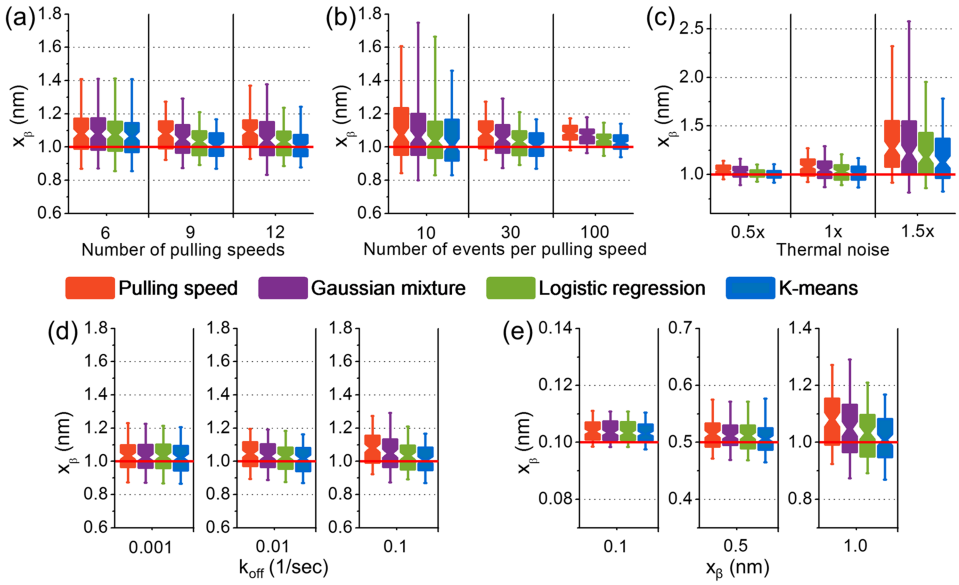

FIG. 6. Evaluation of cluster analysis for estimation of x_β. Unbinding events were generated using Monte Carlo simulations with k_{off} = 0.1 s^{-1}, x_β = 1 nm, 9 pulling speeds, 30 events per pulling speed, and 1× thermal noise. Pulling speeds, Gaussian mixture cluster model, logistic regression clustering, and K-means clustering were used to group data. Kinetic parameters were estimated using the Bell-Evans model. Simulations were repeated 100 times. Red line represents the preset values of x_β. Clustering methods were compared by varying (a) the number of pulling speeds, (b) the amount of data, (c) the thermal noise, (d) k_{off}, and (e) x_β. As in the case of k_{off} estimates shown in Fig. 5, cluster analysis significantly improved x_β estimation when the number of pulling speeds, the number of unbinding events, the thermal noise, k_{off}, and x_β increased.

standard pulling speed method showed larger relative errors (−40%, −35%, −36% for GMM and −44%, −44%, −54% for pulling speed) for off rates of 10^{-3} s^{-1}, 0.01 s^{-1}, and 0.1 s^{-1}, respectively.

Similarly, when the conventional pulling speed method was used, increasing the width of the energy barrier increased the error in k_{off} estimation (−2% error at 0.1 nm, −33% error at 0.5 nm, and −54% error at 1 nm). However, this increase in relative error was not observed when cluster analysis was used [Fig. 5(e)]. With barrier widths of 0.1 nm, 0.5 nm, and 1 nm, the errors in estimated k_{off} were 0%, −16%, and −11% for K-means; −1%, −26%, and −24% for logistic regression; −1%, −26%, and −36% for GMM.

Most importantly, the improvement in x_β estimation by cluster analysis followed a very similar trend as the k_{off} estimation (Fig. 6), indicating that the accuracy of both k_{off} and x_β estimates increased at the same time. This is particularly encouraging because it demonstrates that the increased accuracy of k_{off} by cluster analysis was not offset by a reduced accuracy in x_β estimation.

Finally, we validated the effect of clustering by simultaneously varying parameters across a range of values that have previously been measured in DFS experiments with biological systems including cell adhesion proteins and antigen-antibody complexes.[13] A total of 324 DFS experiments were simulated.

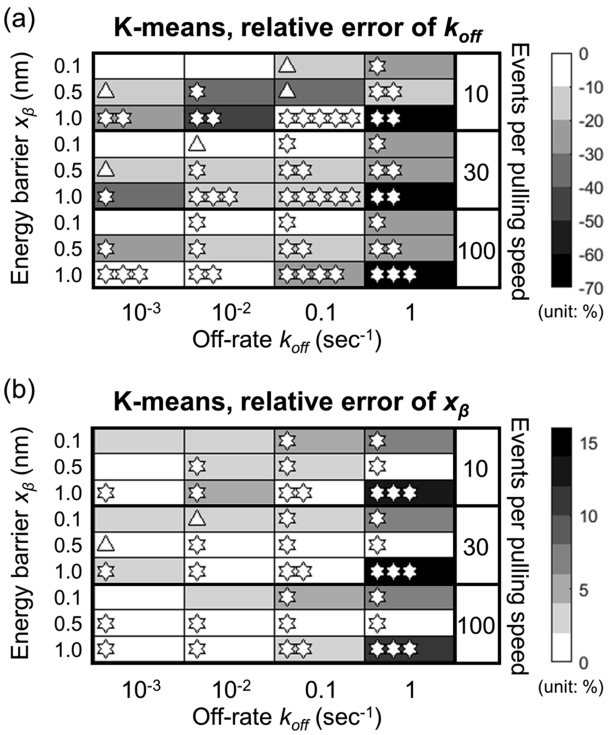

FIG. 7. Relative error in estimated (a) k_{off} and (b) x_β using K-means clustering. DFS data were simulated using 9 pulling speeds and 1× thermal noise. Other parameters include k_{off} = 10^{-3}, 0.01, 0.1, 1 s^{-1}; x_β = 0.1, 0.5, 1 nm; the number of events per pulling speed = 10, 30, 100. For each condition, the relative errors were compared with the pulling speed method. Stars indicate conditions where K-means is more accurate (0 stars: no improvement; 1 star: relative error is reduced ≤10% for k_{off} and ≤5% for x_β; 2 stars: relative error is reduced by 10%–20% for k_{off} and 5%–10% for x_β; etc.). Triangles indicate conditions where the K-means analysis is less accurate; relative error is increased ≤10% for k_{off} and ≤5% for x_β.

We used four values of k_{off} (10^{-3}, 0.01, 0.1, and 1 s^{-1}) in combination with three values of x_β (0.1, 0.5, and 1 nm). In order to account for different experimental conditions, we also varied the number of pulling speeds (6, 9, and 12), number of events per pulling speed (10, 30, and 100 events), and different levels of thermal noise (0.5×, 1×, and 1.5×). We applied cluster analysis to each condition and compared the estimated k_{off} and x_β to the results obtained using the pulling speed method. The complete results of our simulations are presented in the supplementary material (Figs. S1–S6; Tables S1–S3). In Fig. 7, we just present results using 9 pulling speeds and 1× thermal noise analyzed using K-means, to illustrate its power in estimating k_{off} and x_β (Fig. 7). We use stars to indicate conditions where K-means is more accurate and triangles to indicate conditions where the standard pulling speed analysis is more accurate; increase in the number of stars/triangles indicates a proportionally higher accuracy using K-means/standard-analysis.

As seen in Fig. 7, K-means clustering improved the accuracy of k_{off} and x_β even when the number of unbinding events was low. The improved accuracy of parameter estimation was more pronounced for wide energy barriers and high off-rates and when the unbinding events across different loading rates overlap (Fig. 7). For instance, at a dissociation rate of 0.1 s^{-1}, x_β of 1 nm, and 30 events per pulling speed, K-means reduced the relative error in k_{off} to −11% as compared to a −54% relative error using pulling speed analysis. In contrast, when the energy barrier was narrow or the off-rate was small, the unbinding events were already well-separated and cluster analysis did not significantly improve the estimation of k_{off} and x_β.

CONCLUSION

This manuscript presents a high accuracy method using clustering algorithms, to improve kinetic parameter estimation while retaining the simplicity of data collection and analysis of a conventional DFS experiment. We benchmarked the performance of different clustering algorithms, by testing them across an extensive range of conditions that mimic real-world experiments. The parameters we varied included the number of unbinding events, pulling speeds, and noise levels, across a range of k_{off} and x_β typical of receptor-ligand pairs. Under these conditions, the K-means method had the highest accuracy in estimating k_{off} and x_β. Although logistic regression and GMM were more accurate than the conventional pulling speed method, the improvement was not as significant as K-means.

The cluster analysis used in this study could be further improved, by grouping unbinding events using Bell-Evans force distributions[10] or more sophisticated distributions described in the literature.[29,37] The analysis method presented in our work can also be used to identify and eliminate artifacts due to the formation of multiple receptor ligand bonds and nonspecific binding events which are not tightly clustered on a force-loading rate plot.[38]

SUPPLEMENTARY MATERIAL

See supplementary material for comparison of different clustering methods using simulated data with different

values of k_{off}, x_β, pulling speeds, number of events, and thermal noise.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Science Foundation under award number PHY-1607550 and the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM121885. We thank Andrew Priest for help with the simulations.

[1] V. Montana, W. Liu, U. Mohideen, and V. Parpura, J. Physiol. **587**, 1943 (2009).
[2] A. Yersin, T. Osada, and A. Ikai, Biophys. J. **94**, 230 (2008).
[3] W. Baumgartner, P. Hinterdorfer, W. Ness, A. Raab, D. Vestweber, H. Schindler, and D. Drenckhahn, Proc. Natl. Acad. Sci. U. S. A. **97**, 4005 (2000).
[4] F. W. Bartels, B. Baumgarth, D. Anselmetti, R. Ros, and A. Becker, J. Struct. Biol. **143**, 145 (2003).
[5] J. Yu, Y. Jiang, X. Ma, Y. Lin, and X. Fang, Chem. - Asian J. **2**, 284 (2007).
[6] B. H. Kim, N. Y. Palermo, S. Lovas, T. Zaikova, J. F. Keana, and Y. L. Lyubchenko, Biochemistry **50**, 5154 (2011).
[7] C. F. Yen, D. S. Harischandra, A. Kanthasamy, and S. Sivasankar, Sci. Adv. **2**, e1600014 (2016).
[8] K. C. Neuman and A. Nagy, Nat. Methods **5**, 491 (2008).
[9] P. Hinterdorfer and Y. F. Dufrene, Nat. Methods **3**, 347 (2006).
[10] E. Evans and K. Ritchie, Biophys. J. **72**, 1541 (1997).
[11] J. T. Bullerjahn, S. Sturm, and K. Kroy, Nat. Commun. **5**, 4463 (2014).
[12] G. I. Bell, Science **200**, 618 (1978).
[13] A. R. Bizzarri and S. Cannistraro, *Dynamic Force Spectroscopy and Biomolecular Recognition* (CRC Press, Boca Raton, FL, 2012).
[14] A. Noy and R. W. Friddle, Methods **60**, 142 (2013).
[15] D. F. Tees, R. E. Waugh, and D. A. Hammer, Biophys. J. **80**, 668 (2001).
[16] D. A. Simson, M. Strigl, M. Hohenadl, and R. Merkel, Phys. Rev. Lett. **83**, 652 (1999).
[17] S. Loi, G. Sun, V. Franz, and H. J. Butt, Phys. Rev. E **66**, 031602 (2002).
[18] C. Gergely, J. Voegel, P. Schaaf, B. Senger, M. Maaloum, J. K. Horber, and J. Hemmerle, Proc. Natl. Acad. Sci. U. S. A. **97**, 10802 (2000).
[19] C. Ray, J. R. Brown, and B. B. Akhremitchev, J. Phys. Chem. B **110**, 17578 (2006).
[20] M. Raible, M. Evstigneev, F. W. Bartels, R. Eckel, M. Nguyen-Duong, R. Merkel, R. Ros, D. Anselmetti, and P. Reimann, Biophys. J. **90**, 3851 (2006).
[21] S. Getfert and P. Reimann, Biophys. J. **102**, 1184 (2012).
[22] C. Ray, J. R. Brown, and B. B. Akhremitchev, Langmuir **23**, 6076 (2007).
[23] C. Friedsam, A. K. Wehle, F. Kuhner, and H. E. Gaub, J. Phys.: Condens. Matter **15**, S1709 (2003).
[24] C. Ray, J. R. Brown, and B. B. Akhremitchev, J. Phys. Chem. B **111**, 1963 (2007).
[25] S. K. Sekatskii, F. Benedetti, and G. Dietler, J. Appl. Phys. **114**, 034701 (2013).
[26] P. D'haeseleer, Nat. Biotechnol. **23**, 1499 (2005).
[27] G. Punj and D. W. Stewart, J. Mark. Res. **20**, 134 (1983).
[28] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
[29] O. K. Dudko, G. Hummer, and A. Szabo, Phys. Rev. Lett. **96**, 108101 (2006).
[30] G. Hummer and A. Szabo, Biophys. J. **85**, 5 (2003).
[31] R. W. Friddle, A. Noy, and J. J. De Yoreo, Proc. Natl. Acad. Sci. U. S. A. **109**, 13573 (2012).
[32] F. Oesterhelt, M. Rief, and H. E. Gaub, New J. Phys. **1**, 6.1 (1999).
[33] R. Levy and M. Maaloum, Nanotechnology **13**, 33 (2002).
[34] H. J. Butt and M. Jaschke, Nanotechnology **6**, 1 (1995).
[35] M. Chen, Pattern Recognition and Machine Learning Toolbox, http://prml.github.io, 2016.
[36] J. C. Stoltzfus, Acad. Emerg. Med. **18**, 1099 (2011).
[37] O. K. Dudko, G. Hummer, and A. Szabo, Proc. Natl. Acad. Sci. U. S. A. **105**, 15755 (2008).
[38] T. Sulchek, R. W. Friddle, and A. Noy, Biophys. J. **90**, 4686 (2006).