Learning and Estimation of Single Molecule Behavior

Sivaraman Rajaganapathy[1,a], James Melbourne[1,b], Tanuj Aggarwal[2,c], Rachit Shrivastava[1,d], and Murti V. Salapaka[1,e]

Abstract— **Data analysis in single molecule studies often involves estimation of parameters and the detection of abrupt changes in measured signals. For single molecule studies, tools for automated analysis that are crucial for rapid progress, need to be effective under large noise magnitudes, and often must assume little or no prior knowledge of parameters being studied. This article examines an iterated, dynamic programming based step detection algorithm (SDA). It is established that given a prior estimate, an iteration of the SDA necessarily improves the estimate. The analysis provides an explanation and a confirmation of the effectiveness of the learning and estimation capabilities of the algorithm observed empirically. Further, an alternative application of the SDA is demonstrated, wherein the parameters of a worm-like chain (WLC) model are estimated, for the automated analysis of data from single molecule protein pulling experiments.**

I. INTRODUCTION

The recently found ability of performing single-molecule experiments have provided crucial insights into the biological machinery of the cell [1], [2]. Single molecule experiments have revealed that many biological systems exhibit discrete behavior [3], [4], [5], [6]. For example, motor-proteins (also known as molecular motors) such as, kinesin and dynein, take discrete steps over microtubules while carrying cargo and form a fundamental mode of transport inside cells [5]. Instruments such as atomic force microscopes (AFMs) and optical tweezers have enabled resolution of measuring forces in the femto to multiple pico Newton range with spatial resolution at the nanometer scale. These instruments have made it possible to perform protein folding and unfolding experiments [7], [8], [9], [10], where domains fold and unfold with force differentials in the pico-Newton and femto-Newton range. The length of the domains are in the tens to hundreds of nanometers. The challenges of single molecule studies are many. For example, in force spectroscopy experiments using AFMs (see Fig. 1), a solution with protein molecules to be studied is deposited on a substrate, and a cantilever with a sharp tip is pressed against the substrate to enable a protein molecule to attach to the tip. Subsequent to pressing the cantilever against the substrate, the cantilever is retracted away from it. If there is a successful attachment the protein is extended and comes under tension. The domains of the protein unfold under the application of tension. The unfolding of a domain leads to a step change in the length of the protein. How much time is needed for the domains to unfold, the relationship to the force applied, the changes in the structure under folding and unfolding events are questions that are studied with AFM based force spectroscopy. Most of the studies involve numerous iterations of experiments. Often, data is collected from a large number of experiments, not just to ensure a high confidence on statistical information, but also because experiment success rates are sometimes intentionally kept low. For example, in single molecule force spectroscopy, the protein solution is diluted enough to ensure that the the success rate of an attachment forming between the tip and protein is between two to ten percent [4], [11]. The low probabilities of attachment ensure that the probability of attaching to multiple protein molecules is low [4], [11]. The task of obtaining accurate yet precise statistics of events under large uncertainties and the confounding dynamics of the instrumentation is a daunting one. For example, in AFM experiments, the protein domains can unfold at rapid rates where the dynamics of the cantilever cannot be ignored. The timing of domain unfolding or folding and changes in protein structure have to be ascertained from the measured cantilever deflection data which is corrupted by effects of thermal noise (process noise), measurement noise and the dynamics of the cantilever probe. It is evident that the needs of fundamental biophysics research necessitate the development of tools for automatic detection and estimation of abrupt changes in measurements and parameters of models used to describe the experimental data. In addition, these tools must be capable of operating with minimal prior information on the systems being studied. In this article, events characterized by steps appearing in data and parameters are emphasized; however, the methods discussed are applicable and extensible to many other characteristics of events (such as impulsive changes in the data).

A number of step detection tools for the study of single molecules are available [12], [13], [14], [15], [16], [17], [18], [19], [20], [21]. A comparative study of tools in [17], [18], [19], [20], [21] reported almost similar performance [3], with the χ^2 based technique in [21] observed to have improved temporal resolution. The step detection algorithm (SDA) [12], an iterative algorithm involving dynamic programming (Viterbi) based methodology, was demonstrated to successfully extract stepping statistics from signals with low signal to noise ratios (SNR), and with no prior information on the number of steps or their sizes. Further, the algorithm was capable of removing the undesirable effects of probe

[1] Department of Electrical & Computer Engineering, University of Minnesota, Minneapolis, MN 55455, USA

[2] Cymer LLC (An ASML Company), 17075 Thornmint Court, San Diego, CA 92127, USA

[a] <sivrmn@umn.edu>, [b] <melbo013@umn.edu>, [c] <aggarwaltanuj@gmail.com>, [d] <shriv058@umn.edu>, [e] <murtis@umn.edu>

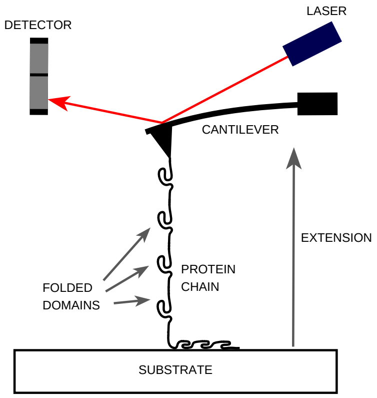

Fig. 1. Schematic of a protein pulling experiment. A polypeptide chain protein is stretched between a substrate and a cantilever tip of an AFM.

dynamics on the data. The SDA's effectiveness is evident from its use in a number of bio-physical studies that include, the detection of wear in molecular tracks [22], the detection of fluorophore events in the study of nanodot arrays [23], and HP1 chromatin binding [24]. Although SDA has found applications in multiple fundamental biophysics studies and its efficacy demonstrated in simulation studies (where the ground truth is known), the reasons for its effectiveness are not established. The algorithm is empirically observed to demonstrate learning, in which previous estimates of the step size distributions are used to provide a new, more accurate estimate of the same. Toward an analysis and substantial extension of SDA's applicability, this article makes two main contributions: (i) It provides the framework to evaluate and analyze the learning capability of the SDA. A key result of the article is that it rigorously establishes that with every learning iteration of the SDA, the estimation improves. (ii) It extends the SDA for learning parameters of models that explain single molecule behavior. Furthermore, it demonstrates the efficacy of the methods developed, on data obtained from simulations as well as experiments.

The article is organized as follows. Section II introduces the step detection problem and provides the details of the SDA. Section III establishes that the SDA's estimates improve with every iteration. Section IV extends the SDA for identifying worm like chain (WLC) models used in protein force spectroscopy experiments. The results of this reformulation are presented and discussed in Sections V & VI with simulation and experiment data.

II. THE STEP DETECTION PROBLEM

We begin with a description of the problem setting and the step detection algorithm proposed in [12].

A. The Step Detection Problem

We will denote stochastic variables by **bold** characters and their realizations by normal characters. Further, to denote a sequence $X = \{x_1, x_2, ..., x_T\}$, we will use capitalized symbols. A partial sequence $\{x_r, x_2, ..., x_s\}$ will be represented by X_r^s, where the subscript and superscript represent the start and end of sequence respectively.

Let $X^\dagger = \{x_1^\dagger, x_2^\dagger, ...x_T^\dagger\}$ be the true stepping signal generated by a system with its dynamics given by

$$\boldsymbol{x_t^\dagger} = \boldsymbol{x^\dagger_{t-1}} + \boldsymbol{u_t},$$

where, $\boldsymbol{u_t}$ is a stochastic variable with an unknown distribution $\boldsymbol{p(u_t)}$ independent of time t. Let $Y = \{y_1, y_2, ..., y_T\}$ be the observations of $X^\dagger$ corrupted by zero mean Gaussian noise with,

$$\boldsymbol{y_t} = \boldsymbol{x_t^\dagger} + \boldsymbol{\eta_t},$$

where, $\boldsymbol{\eta_t} \sim \mathcal{N}(0, \sigma^2)$ for all t. The task of the step detection algorithm is to find an accurate estimate $\hat{X} = \{\hat{x}_1, \hat{x}_2, ...\hat{x}_T\}$ of $X^\dagger$ given the observations Y, and, to obtain an estimate of the step-size distribution (that is, an estimate of $p(\boldsymbol{u})$). It is to be noted that even though we have not introduced probe dynamics that can confound measurements, the methodology is applicable in the presence of such dynamics. The simpler model is used to keep the presentation accessible and for space considerations.

B. SDA Formulation

As presented in [12], the SDA takes the measured data as input and a fits a staircase possibly with different step-sizes to the data (we assume that the staircase also admits negative steps). The SDA is iterative. Assuming the measurement noise is zero mean with variance σ^2, where, an estimate of the variance is known, the SDA has as its first iterate, the solution of the following optimization problem:

$$\hat{X}^* = \arg\min_X \left\{ \sum_{t=1}^{T} (y_t - x_t)^2 + W_t(x_t - x_{t-1}) \right\}, \quad (1)$$

where

$$W_t(x_t - x_{t-1}) = 9\sigma^2 \bar{\delta}(x_t - x_{t-1}), \quad (2)$$

with

$$\bar{\delta}(u) = \begin{cases} 0 & u = 0 \\ 1 & otherwise \end{cases}.$$

Equation 1 has a quadratic error term and a penalty term that penalizes steps in the fit to the data. The initial high penalty of nine times the variance of noise given by (2), ensures that any step included in the first staircase fit is a true step with high probability. The optimization problem is solved using dynamic programming. It is reasonable to expect that any steps in the fit with penalty (2) will have high probability of being a true step; however, many true steps will be missed in the fit. Subsequent to the first iteration, the staircase fit obtained is used to create a histogram of step sizes obtained from the fit. The histogram is normalized to obtain the penalty term for the subsequent iterate, where, step-sizes with higher frequency in the histogram are penalized with less severity compared to the ones with low frequency. This procedure of, fitting a staircase to the data using an estimate of step-size distribution followed by obtaining an estimate of step-size distribution from the fit, is iterated until no further change is observed in the histogram of step-sizes. Remarkably, the algorithm is empirically shown to perform well and is used extensively in many biophysics studies [24], [23], [22]. It is to be remarked that the SDA assumes no prior knowledge of step-sizes, where, multiple step sizes are also admissible.

In this section, the SDA's operational steps are cast in a formal setting, to enable rigorous analysis. First the connection of each iterate with maximum a posteriori probability estimate is presented. The maximum a posteriori probability (MAP) fit is obtained by

$$\hat{X}^{MAP} := \arg\max_X \left\{ p(X|Y) \right\},$$

where, $p(X|Y)$, is the probability density function of $\boldsymbol{X}$ given $\boldsymbol{Y} = Y$ evaluated at X . If $\gamma(\cdot)$ is the distribution over X, then

$$\hat{X}^{MAP} = \arg\max_X \left\{ p(Y|X)\gamma(X) \right\}.$$

Consider the case where the step-size distribution from a previous fit is used to obtain a probability distribution $\gamma(\cdot)$ over $\boldsymbol{u}$. Then, Theorem 1 demonstrates the link between the SDA estimate and a MAP estimate.

Theorem 1. *Given a prior probability distribution $\gamma(\cdot)$ over step size $\boldsymbol{u} = \boldsymbol{x_t} - \boldsymbol{x_{t-1}}$, $\hat{X}^{MAP} = \hat{X}^{SDA}$, when $x_0 = 0$ with*

$$\hat{X}^{SDA} = \arg\min_X \left\{ \sum_{t=1}^{T} (y_t - x_t)^2 + W_t(x_t - x_{t-1}) \right\} \quad (3)$$

and

$$W_t(x_t - x_{t-1}) = -2\sigma^2 \log\left[\gamma(\boldsymbol{u_t} = x_t - x_{t-1}|X_1^{t-1})\right].$$

Proof. See [12] for a proof. □

Theorem 1 shows that if the true step size distribution were known, then optimizing the objective function $\sum_{t=1}^{T}(y_t - x_t)^2 + W_t(x_t - x_{t-1})$ provides the MAP estimate of $X^{\dagger}$.

The SDA thus assumes that the system generating the step signals has a generative model as shown in Fig. 2. Here the prior probability $\gamma(\cdot)$ over $\boldsymbol{u}$ is assumed to be approximated by a histogram with fixed bin ranges $R = \{r_1, r_2, ..., r_K\}$ of equal widths Δr and free bin heights $\Theta = \{\theta_1, \theta_2, ..., \theta_k\}$. The choice of the number of bins K and the ranges R is chosen to approximate $\gamma(\cdot)$ accurately without overfitting. Thus, the bin heights Θ parameterize the distribution $\gamma(\cdot)$. We first cast the SDA formally into two principal stages, namely the minimization stage (M-Stage) followed by the evaluation stage (E-Stage), which are iterated N times to give an estimate $\left[\hat{X}^{(N)}, \Theta^{(N)}\right]$. Let $n \in \{1, 2, ..., N\}$ represent iteration count. Then the two stages are as follows:

M-Stage: Minimize the objective function $J(X, \Theta^{(n-1)})$ with respect to X to give $\hat{X}^{(n)}$, where

$$\hat{X}^{(n)} := \arg\min_X J(X, \Theta^{(n-1)}), \quad (4)$$

with,

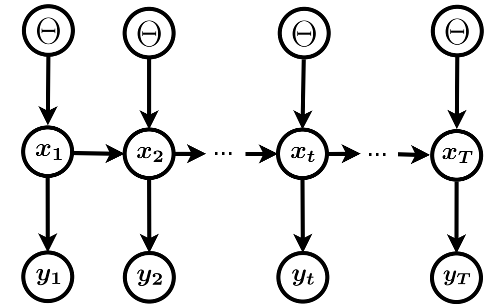

Fig. 2. Generative model assumed by the SDA to reconstruct step signals. Here $\Theta = \{\theta_1, ..., \theta_K\}$ are the time invariant parameters of the unknown step size distribution. x_t are the true steps that are hidden, and y_t the noisy observations of x_t.

$$J(X, \Theta^{(n-1)}) := \sum_{t=1}^{T}(y_t - x_t)^2 + W_t(X, \Theta^{(n-1)}).$$

Here, $W_t(X, \Theta^{(n-1)})$ is a regularization term used to prevent overfitting and is defined as:

$$W_t(X, \Theta^{(n-1)}) = \begin{cases} 9\sigma^2\bar{\delta}(x_t - x_{t-1}) & n = 1 \\ -2\sigma^2 \log\left[p(u_t|X_1^{t-1}, \Theta^{(n-1)})\right] & n > 1 \end{cases}$$

with

$$\bar{\delta}(u) = \begin{cases} 0 & u = 0 \\ 1 & otherwise \end{cases}.$$

Here, $p(u_t|X_1^{t-1}, \Theta^{(n-1)})$ is the probability density function of taking a step $\boldsymbol{u_t}$ evaluated for the step size u_t, with $u_t = x_t - x_{t-1}$ in iteration n, given $\Theta^{(n-1)}$, the previous estimate of parameters of the step size distribution. In this step, $\Theta^{(n-1)}$ is assumed as known.

E-Stage: Evaluate $\Theta^{(n)} = \{\theta_1^{(n)}, \theta_2^{(n)}, ..., \theta_K^{(n)}\}$ using

$$\theta_k^{(n)} = \frac{s_k^{(n)}}{T} \text{ for all } k, \quad (5)$$

where $s_k^{(n)}$ is the number of steps, $u_t^{(n)}$, in the fit $\hat{X}^{(n)}$ contained within the bin r_k. Thus $s_k^{(n)} = |\{t : u_t^{(n)} \in r_k\}|_1$ where $|\Gamma|_1$ is the cardinality of the set Γ. Here T is the total number of observations of steps $u_t^{(n)}$ in the estimate $\hat{X}^{(n)}$, (thus equal to $\sum_{k=1}^{K} s_k^{(n)}$). The SDA is summarized in Algorithm 1.

Algorithm 1 Step Detection Algorithm

```
Input: Time series measurements $Y = \{y_1, y_2, ..., y_T\}$.
Output: Step fit $\hat{X}^{(N)} = \{\hat{x}_1^{(N)}, \hat{x}_2^{(N)}, ..., \hat{x}_T^{(N)}\}$ and Histogram bin parameters $\Theta^{(N)} = \{\theta_1^{(N)}, \theta_2^{(N)}, ..., \theta_K^{(N)}\}$.
Set initial penalty using (2)
Compute step fit $\hat{X}^{(1)}$ using (1)
for all $n \in \{2, 3, ..., N\}$ do
    E-Stage: Compute $\Theta^{(n)}$ using (5)
    M-Stage: Compute step fit $\hat{X}^{(n)}$ using (4)
end for
```

In the implementation, dynamic programming is used in the M-Stage for the minimization of $J(X, \Theta^{(n-1)})$ with respect to X. The total number of iterations N is chosen such that changes in $\hat{X}$ and Θ are within desired tolerances. With the formalization of the framework complete, the main result is derived in the next section.

III. ITERATIVE LEARNING IN THE STEP DETECTION ALGORITHM

In [12], it is empirically observed that the fits $\hat{X}^{(n)}$ improve in accuracy with increasing n. The SDA, in each iteration, consisting of the M-Stage and E-Stage improves upon the previous estimates in the sense of increasing the joint conditional probability density $p(\hat{X}, \Theta|Y)$ evaluated at $\hat{X}^{(n)}, \Theta^{(n)}$. The SDA uses the knowledge gained from the previous estimates in the M-Stage, in the form of a previously estimated distribution ($\Theta^{(n-1)}$) of the step sizes. However, in the E-Stage, the re-estimation of the step size distribution $\Theta^{(n)}$ from the fit $\hat{X}^{(n)}$ is kept unbiased. For example, consider the case where the distribution Θ is represented by a histogram with K bins $\{\theta_1, \theta_2, ..., \theta_K\}$. Then, we assume that any two histogram bin heights $\Theta^{'}$ and $\Theta^{''}$ satisfy $p(\Theta^{'}) = p(\Theta^{''})$. Such an assumption is feasible, since for any integer K, $\Omega = \{\Theta : \sum_{i=1}^{K} \theta_i = 1,\ \theta_i \geq 0\}$, is a compact set of an affine subspace $\mathbb{R}^K$, and thus can be endowed with a uniform probability measure. Using the natural correspondence between Ω and the space of $K-$bin histograms, we can proceed.

Theorem 2. *For Θ uniformly distributed, let the observations Y be corrupted by zero mean Gaussian noise with known variance σ^2. That is, $\boldsymbol{y_t} = \hat{\boldsymbol{x}}_{\boldsymbol{t}} + \boldsymbol{\eta_t}$, with $\boldsymbol{\eta_t} \sim \mathcal{N}(0, \sigma^2)$ for all t. Then, given an estimate $[\hat{X}^{(n-1)}, \Theta^{(n-1)}]$, the step detection algorithm's iteration (see Algorithm 1) produces an improved estimate $[\hat{X}^{(n)}, \Theta^{(n)}]$ where*

$$p(\hat{X}^{(n)}, \Theta^{(n)}|Y) \geq p(\hat{X}^{(n-1)}, \Theta^{(n-1)}|Y).$$

Proof. The proof is uses Lemma 1 to show an improvement in both the M-Stage and the E-Stage and omitted due to space constraints. □

Lemma 1. *Let $U = \{u_1, u_2, ..., u_T\}$ be the observations of a stationary stochastic variable. Let $\mathcal{H}$ be a histogram approximating the distribution of U, where $\mathcal{H}$ has K bins and bin ranges $R = \{r_1, r_2, ..., r_K\}$ of equal widths Δr such that all the observations are contained within the union of the ranges of the bins. Let the bin heights be $\Theta = \{\theta_1, \theta_2, ..., \theta_K\}$. Let s_k be the number of observations u_t contained within the bin r_k, that is $s_k = |\{t : u_t \in r_k\}|$. Then, setting the bin heights $\theta_k = \frac{s_k}{T}$ maximizes the likelihood $p(U|\Theta)$ with respect to Θ.*

Proof. The proof uses the method of Lagrange multipliers to solve the constrained optimization problem and is left to the reader. □

IV. WLC MODEL FITTING

Here we apply the step detection algorithm for the analysis of data from atomic force microscope (AFM) based protein pulling experiments.

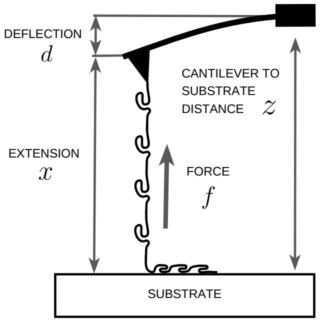

Fig. 3. Measurements in a pulling experiment. The deflection in the cantilever as well as the distance between the tip and substrate are measured, allowing the force versus extension characteristics to be determined.

A. Protein Pulling Experiments

In a protein pulling experiment, a single molecule of a polypeptide chain protein is attached between a substrate and a cantilever tip of an AFM as shown in Fig. 3. The distance z between the cantilever tip and the substrate is measured and is controlled by a piezoelectric transducer. The deflection d in the cantilever is measured via a laser photo-diode sensor arrangement. Thus, the extension $(x = z - d)$ can be calculated. The deflection in the cantilever is multiplied by its spring constant to convert it to the force f in the protein. During an experiment, the distance z is increased at a constant speed, so as to apply an increasing tensile force on the protein. It is known (see [25]) that force f is well characterized by the force-extension relation of the worm like chain model f^{WLC}, described by (6).

$$f^{WLC}(x, l, p) = \frac{k_B T}{p} \left[\frac{1}{4} \left(1 - \frac{x}{l} \right)^{-2} - \frac{1}{4} + \frac{x}{l} \right], \quad (6)$$

where k_B is the Boltzmann constant, and T is the temperature. Here, l, which represents the contour length and p, which represents the persistence of the protein are the free parameters whose estimates are of interest in the experimental study. Further, during a pulling experiment, folded domains in the protein unfold rapidly, thereby abruptly changing both the contour and persistence lengths [4]. The data analysis then involves the detection of such abrupt changes in the contour and persistence lengths.

B. SDA Reformulation Strategy

Our task now is to find the estimates $\hat{L}$ and $\hat{P}$ of $L = \{l_1, l_2, ..., l_T\}$ and $P = \{p_1, p_2, ..., p_T\}$ given the measurements of force $F = \{f_1, f_2, ..., f_T\}$ and extension $X = \{x_1, x_2, ..., x_T\}$. While experimental data shows that the noise in the measured force is stationary, the corresponding noise in the contour and persistence lengths when inverting the WLC function is non-stationary and dependent on the extension. Further, the percentage change in persistence length is typically smaller than the percentage change in the contour length, leading to a poorer SNR in the estimates of p compared to those of l. Due to these challenges, the SDA is reformulated as follows:

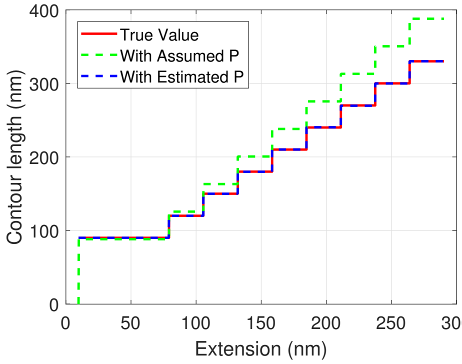

Fig. 4. Estimates of the contour length with the reformulated SDA terminated at 20 iterations. The estimates in green were obtained with the persistence length assumed constant, which leads to the expected error in the contour length magnitudes. The estimates recomputed locally to improve accuracy are shown in blue. This is done after estimating the corresponding persistence lengths.

M-Stage: Minimize the objective function $J^{WLC}(L,\Theta^{(n-1)})$ with respect to L to give $\hat{L}^{(n)}$, that is

$$\hat{L}^{(n)} = \arg\min_{X} J^{WLC}(L,\Theta^{(n-1)}),$$

where, $J^{WLC}(L,\Theta^{(n-1)})$ is defined as

$$\sum_{t=1}^{T}\left[f_t - f_t^{WLC}(x_t, l_t, p_o)\right]^2 + W_t(L,\Theta^{(n-1)}),$$

with

$$W_t(L,\Theta^{(n-1)}) = \begin{cases} 9\sigma^2\bar{\delta}(l_t - l_{t-1}) & n=1 \\ -2\sigma^2\log\left[p(u_t|L_1^{t-1},\Theta^{(n-1)})\right] & n>1 \end{cases}.$$

E-Stage: Evaluate $\Theta^{(n)}$ using $\Theta^{(n)} = \arg\max_{\Theta} p(U^{(n)}|\Theta)$. Here σ^2 is the variance of noise in the force measurements F, n the iteration count, and p_o is an initial guess of the mean persistence length. The steps u_t are redefined as $u_t = l_t - l_{t-1}$. This reformulation allows us to determine the steps in L. Since the step locations in L are caused by unfolding events, the locations of the steps in P must be the same. At this stage, the regions of the force versus extension curves corresponding to the step changes in L and P are extracted. Local estimates for both are then recomputed using a least squares fit to improve the estimation accuracy.

V. Results

A. Simulations

The reformulated SDA was applied to simulated data containing 8 unfolding events with additive Gaussian noise ($\mathcal{N}(0,\sigma = 10\,\text{pN})$) in the measured forces. The contour lengths were incremented from 90 nm in steps of 30 nm, and the persistence lengths incremented from 115 pm in steps of 35 pm. These simulation parameters were chosen to represent a typical scenario encountered in our experiments.

In Fig. 4, the estimates $\hat{L}$ from the SDA are compared with the ground truth. The algorithm was initialized with $p_o = 130$ pm, and thus the errors in the estimate for $\hat{L}$ are lesser in the region where P was close to this value. Although, the errors appear to increase in other regions, we note that the location of steps in the estimate match those of the truth. These regions are then used to find $\hat{P}$ and improve the estimates of $\hat{L}$, as shown in Fig. 5 and Fig. 4 respectively. Fig. 6 compares the true and measured forces with those computed from the estimates $\hat{L}$ and $\hat{P}$.

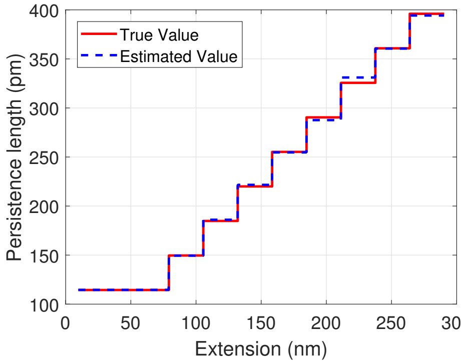

Fig. 5. Persistence length estimates compared with the actual values when applying the SDA on simulated data.

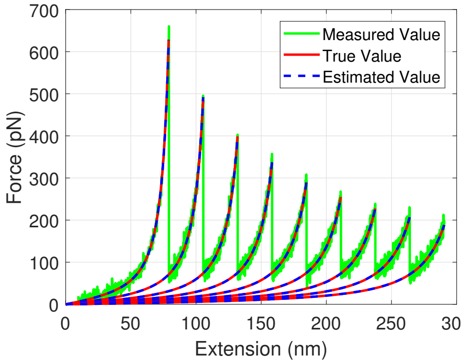

Fig. 6. Estimates of the force versus extension curve compared with the true and measured curves from the simulated system.

B. Experimental Application

The reformulated SDA was used to analyze the data from single molecule force spectroscopy on the I27O AFM reference protein from AthenaES, which is a combination of 8 repeats of the domain Ig 27 from human titin. The MFP-3D AFM from Oxford Instruments was used with Bio-cantilevers (BL-RC-150VB) from Asylum Research with a mean spring constant of 6 pN nm^{-1}. Further, the actual cantilever spring constant is measured by analyzing their thermal fluctuations. All the experiments were conducted at a temperature of 298 K, with the protein solution applied to a fresh gold substrate. The AFM cantilever's tip is then pressed against the substrate with a force in the range of 0.5nN to 2.0nN for a duration of 3s. The tip is retracted away at a constant speed. Whenever different ends of a segment of the protein are adsorbed to the tip and to the substrate, a tensile force is applied on the segment. To ensure that only a single molecule is present between the tip and the substrate, the concentration of the protein solution is kept low (typically

TABLE I

TITIN PARAMETERS: REPORTED VS SDA ESTIMATES

	Reported	Estimated with SDA	% Error
ΔL (nm)	28.4 (see [26])	28.04	-1.3
P (pm)	300 ± 70 (see [27])	292	-2.7

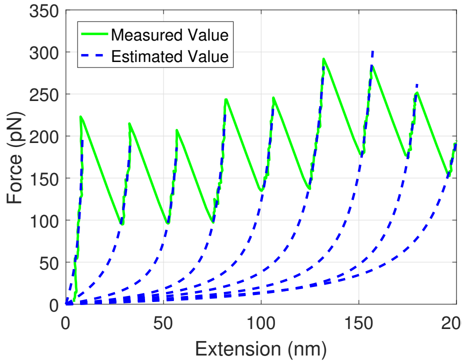

Fig. 7. Sample force versus extension curve from pulling experiments on titin. The force estimates from the WLC models fit after each unfolding event are overlaid on the measurements.

below 200 nM l^{-1}).

The estimates of the most probable persistence length (P) and the most probable contour length per domain (ΔL) from analyzing over 3000 experiments with a 6% success rate are compared in Table I with the reported values in [26], [27]. A sample force curve and its estimation is shown in Fig. 7. The automated analysis of the data using the SDA yielded estimates within acceptable tolerances.

VI. CONCLUSIONS

This article builds a mathematical framework for the analysis of a step detection algorithm and establishes that every iterate of the learning process improves the estimates over the previous estimate. The article extends the use of the step detection algorithm for estimating parameters of a model for polymer chains. Furthermore, it shows the efficacy of the methods developed on force spectroscopy data obtained using atomic force microscopy.

ACKNOWLEDGMENT

We thank Sayan Ghosal for developing the protocols and providing training for conducting the AFM based force spectroscopy experiments. The work in the article was made possible by NSF, CMMI grant and NSF, CPS grant to MS.

REFERENCES

[1] S. Weiss, “Fluorescence spectroscopy of single biomolecules,” *Science*, vol. 283, no. 5408, pp. 1676–1683, 1999.

[2] F. Ritort, “Single-molecule experiments in biological physics: methods and applications,” *Journal of Physics: Condensed Matter*, vol. 18, no. 32, p. R531, 2006.

[3] B. C. Carter, M. Vershinin, and S. P. Gross, “A comparison of step-detection methods: how well can you do?” *Biophysical journal*, vol. 94, no. 1, pp. 306–319, 2008.

[4] A. F. Oberhauser, P. K. Hansma, M. Carrion-Vazquez, and J. M. Fernandez, “Stepwise unfolding of titin under force-clamp atomic force microscopy,” *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 468–472, 2001.

[5] K. Svoboda, C. F. Schmidt, B. J. Schnapp, and S. M. Block, “Direct observation of kinesin stepping by optical trapping interferometry,” *Nature*, vol. 365, no. 6448, pp. 721–727, 1993.

[6] T. Ha, “Single-molecule methods leap ahead,” *Nature methods*, vol. 11, no. 10, pp. 1015–1018, 2014.

[7] E. M. Puchner and H. E. Gaub, “Force and function: probing proteins with afm-based force spectroscopy,” *Current opinion in structural biology*, vol. 19, no. 5, pp. 605–614, 2009.

[8] F. M. Fazal and S. M. Block, “Optical tweezers study life under tension,” *Nature photonics*, vol. 5, no. 6, pp. 318–321, 2011.

[9] M. Ludwig, M. Rief, L. Schmidt, H. Li, F. Oesterhelt, M. Gautel, and H. Gaub, “Afm, a tool for single-molecule experiments,” *Applied Physics A: Materials Science & Processing*, vol. 68, no. 2, pp. 173–176, 1999.

[10] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, “Reversible unfolding of individual titin immunoglobulin domains by afm,” *science*, vol. 276, no. 5315, pp. 1109–1112, 1997.

[11] R. Law, P. Carl, S. Harper, P. Dalhaimer, D. W. Speicher, and D. E. Discher, “Cooperativity in forced unfolding of tandem spectrin repeats,” *Biophysical journal*, vol. 84, no. 1, pp. 533–544, 2003.

[12] T. Aggarwal, D. Materassi, R. Davison, T. Hays, and M. Salapaka, “Detection of steps in single molecule data,” *Cellular and molecular bioengineering*, vol. 5, no. 1, pp. 14–31, 2012.

[13] T. Aggarwal, *Novel Tools for Biophysics Research*. University of Minnesota, 2012.

[14] L. S. Milescu, A. Yildiz, P. R. Selvin, and F. Sachs, “Extracting dwell time sequences from processive molecular motor data,” *Biophysical journal*, vol. 91, no. 9, pp. 3135–3150, 2006.

[15] F. E. Müllner, S. Syed, P. R. Selvin, and F. J. Sigworth, “Improved hidden markov models for molecular motors, part 1: basic theory,” *Biophysical journal*, vol. 99, no. 11, pp. 3684–3695, 2010.

[16] S. Syed, F. E. Müllner, P. R. Selvin, and F. J. Sigworth, “Improved hidden markov models for molecular motors, part 2: extensions and application to experimental data,” *Biophysical journal*, vol. 99, no. 11, pp. 3696–3703, 2010.

[17] W. Hua, E. C. Young, M. L. Fleming, and J. Gelles, “Coupling of kinesin steps to atp hydrolysis,” *Nature*, vol. 388, no. 6640, p. 390, 1997.

[18] N. J. Carter and R. Cross, “Mechanics of the kinesin step,” *Nature*, vol. 435, no. 7040, p. 308, 2005.

[19] B. M. Sadler and A. Swami, “Analysis of wavelet transform multiscale products for step detection and estimation,” ARMY RESEARCH LAB ADELPHI MD, Tech. Rep., 1998.

[20] ——, “Analysis of multiscale products for step detection and estimation,” *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 1043–1051, 1999.

[21] J. W. Kerssemakers, E. L. Munteanu, L. Laan, T. L. Noetzel, M. E. Janson, and M. Dogterom, “Assembly dynamics of microtubules at molecular resolution,” *Nature*, vol. 442, no. 7103, p. 709, 2006.

[22] E. L. Dumont, C. Do, and H. Hess, “Molecular wear of microtubules propelled by surface-adhered kinesins,” *Nature nanotechnology*, vol. 10, no. 2, pp. 166–169, 2015.

[23] H. Cai, H. Wolfenson, D. Depoil, M. L. Dustin, M. P. Sheetz, and S. J. Wind, “Molecular occupancy of nanodot arrays,” *ACS nano*, vol. 10, no. 4, pp. 4173–4183, 2016.

[24] L. C. Bryan, D. R. Weilandt, A. L. Bachmann, S. Kilic, C. C. Lechner, P. D. Odermatt, G. E. Fantner, S. Georgeon, O. Hantschel, V. Hatzimanikatis *et al.*, “Single-molecule kinetic analysis of hp1-chromatin binding reveals a dynamic network of histone modification and dna interactions,” *Nucleic Acids Research*, 2017.

[25] M. Rief, J. Pascual, M. Saraste, and H. E. Gaub, “Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles,” *Journal of molecular biology*, vol. 286, no. 2, pp. 553–561, 1999.

[26] M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez, “Mechanical and chemical unfolding of a single protein: a comparison,” *Proceedings of the National Academy of Sciences*, vol. 96, no. 7, pp. 3694–3699, 1999.

[27] A. E. Systems, “I27o afm reference protein,” AthenaES, 1450 South Rolling Road Baltimore, MD 21227 USA, Tech. Rep. Catalog Number: 0304, 2011.