Extended Kalman filter estimates the contour length of a protein in single molecule atomic force microscopy experiments

Vicente I. Fernandez,[1,a)] Pallav Kosuri,[2] Vicente Parot,[3] and Julio M. Fernandez[4]
[1]*Department of Mechanical Engineering, MIT, Massachusetts 02139, USA*
[2]*Department of Biochemistry and Molecular Biophysics, Columbia University, New York 10027, USA*
[3]*Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile*
[4]*Department of Biological Sciences, Columbia University, New York 10027, USA*

(Received 13 September 2009; accepted 30 September 2009; published online 6 November 2009)

Atomic force microscopy force spectroscopy has become a powerful biophysical technique for probing the dynamics of proteins at the single molecule level. Extending a polyprotein at constant velocity produces the now familiar sawtooth pattern force-length relationship. Customarily, manual fits of the wormlike chain (WLC) model of polymer elasticity to sawtooth pattern data have been used to measure the contour length L_c of the protein as it unfolds one module at a time. The change in the value of L_c measures the number of amino acids released by an unfolding protein and can be used as a precise locator of the unfolding transition state. However, manual WLC fits are slow and introduce inevitable operator-driven errors which reduce the accuracy of the L_c estimates. Here we demonstrate an extended Kalman filter that provides operator-free real time estimates of L_c from sawtooth pattern data. The filter design is based on a cantilever-protein arrangement modeled by a simple linear time-invariant cantilever model and by a nonlinear force-length relationship function for the protein. The resulting Kalman filter applied to sawtooth pattern data demonstrates its real time, operator-free ability to accurately measure L_c. These results are a marked improvement over the earlier techniques and the procedure is easily extended or modified to accommodate further quantities of interest in force spectroscopy. © *2009 American Institute of Physics.*

[doi:10.1063/1.3252982]

I. INTRODUCTION

The extension of a single protein using atomic force microscopy (AFM) techniques has opened new avenues for the exploration of the physical mechanisms underlying protein dynamics. The original use for the AFM was surface microscopy by tracing a solid object in order to image it.[1] However, it was found that molecules can be induced to stick to a cantilever tip by an unknown mechanism simply by applying positive pressure onto a layer of molecules deposited on a gold surface.[2,3] Taking advantage of this, a single protein can be attached to the cantilever tip and to a movable base. Moving the base away from the cantilever pulls on the protein, thus deflecting cantilever and inducing tension on the protein [Fig. 1(a)]. The nanometer scale deflections of the cantilever are measured through a laser beam which is reflected off of the tip of the cantilever and onto a split photodiode [Fig. 1(a)].

By ligating multiple copies of the cDNA (complementary DNA) coding for a single protein and expressing the resultant gene in bacteria, it is now possible to produce "polyproteins" consisting of multiple copies of a single protein.[3] The mechanical properties of these engineered polyproteins can be examined in detail. When a polyprotein is picked up and stretched, the resulting force-extension curve has the characteristic appearance of a sawtooth pattern [Fig. 2(b)]. The sawtooth pattern results from the sequential unfolding of all the protein modules as the protein is elongated. Several characteristic features of a sawtooth pattern are revealing of the mechanical architecture of the protein being studied. For example, the peak force reached before an unfolding event measures mechanical stability.[4] The unprecedented resolution afforded by these single molecule experiments contains further detailed information about the dynamically unfolding protein which is not directly evident in the sawtooth patterns. One important state variable that cannot be measured directly from the data is the contour length of the protein, which counts the number of amino acids that contribute to the entropic recoil of a protein. The contour length increase triggered by unfolding is characteristic of each protein type, providing a unique fingerprint that permits protein identification and the resolution of fine details in the mechanical architecture of a protein.[5,6]

Proteins are composed of covalently linked amino acids (polypeptide) that fluctuate in their position with respect to each other, driven by thermal energy. The properties of a fluctuating polypeptide have been described using the wormlike chain (WLC) model of polymer elasticity.[2,3,7,8] The WLC model was primarily developed to explain the extensibility of DNA molecules.[9] Although this model is based on assumptions that very clearly do not apply to proteins,[10] its use became widespread as a convenient empirical description of protein elasticity. The WLC model fits the nonlinear relationship between force and extension of unfolded proteins,

[a)] Author to whom correspondence should be addressed. Electronic mail: vicentef@mit.edu.

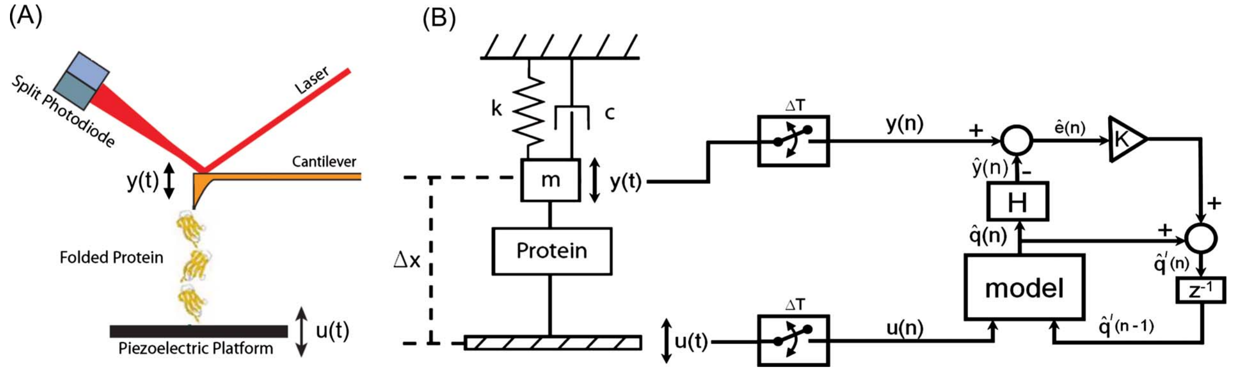

FIG. 1. (Color online) Single molecule AFM implementation of a KF, for the measurement of the contour length of a protein. (a) Diagram of the experimental arrangement. The deflection of a laser beam by a cantilever measures the force generated by a single protein being stretched by the motion $u(t)$ of a piezoelectric actuator. (b) Diagram of the KF implementation. A physical model of the cantilever-protein arrangement (left) is used to generate an *a priori* estimate of the state variables of the measuring system. The *a priori* state variables are used to generate an estimate of the measured cantilever deflection $\hat{y}(n)$, which is compared with the measured value. The resulting error $\hat{e}(n)$ is multiplied by the optimized Kalman gain to produce a refined *a posteriori* estimate of the state variable L_c; the contour length of the protein. The KF optimizes the state variable estimates in order to minimize the variance of the error.

albeit only at high forces. The parameters of the WLC model are the contour length of the molecule, L_c, and the persistence length p which describe the random-coil behavior of a polymer.[9] Manual fits of the WLC model to force-extension traces has become the method of choice for the measurement of contour length in a wide variety of proteins.[5,6] Here we demonstrate the implementation of state variable estimation techniques that are of widespread use in systems control engineering. We demonstrate the use of an extended Kalman filter (EKF) to track the changes in contour length of a polyprotein unfolding under a stretching force. KFs excel at extracting state variables from noisy data, and do so in real time.[11] The application of a KF to single molecule force spectroscopy data exhibits several substantial improvements over the current practice.

II. THEORY AND EXPERIMENTS

A. Single-molecule force spectroscopy

In our experiments we used polyproteins made of nine identical repetitions of ubiquitin, a small protein consisting of 76 amino acids. The polyproteins were cloned into pT7Blue vector-XL1Blue *E.coli* system and expressed in pQE80L vector-BLR(DE3) *E.coli* cells. The proteins were purified using Ni^{2+} affinity chromatography followed by size-exclusion chromatography. The details of the custom-made atomic force microscope and its mode of operation have been described elsewhere.[12] In our force-extension experiments we used silicon nitride cantilevers (Veeco, Santa Barbara, CA) with an average spring constant of ~30 pN/nm which was measured using equipartition.[13] The typical experiment consisted of adding 5–10 μl of the ubiquitin polyprotein solution (0.1–0.2 mg/ml) to a drop of phosphate buffered saline (PBS, pH 7.0) solution placed on a gold-coated cover slide. After ~10 min, a 10–30-nm-thick layer of protein adsorbed onto the gold surface. Contact between an AFM cantilever and the protein covered surface occasionally results in a single polyprotein adhering to both the gold surface and the cantilever. In such cases, extending the polyprotein by means of the piezoelectric actuator yielded highly reproducible and recognizable sawtooth patternlike force-extension traces. Our experiments were done at room temperature, at a pulling rate of 400 nm/s. The analog to digital converter was set to a sampling rate of 14.3 kHz and the resulting data were low pass filtered with an eight pole Bessel filter set to a 5 kHz cutoff frequency.

B. Mechanical model of the cantilever-protein system

The KF is a well-known algorithm for estimating the state variables of a linear system in the presence of white Gaussian noise.[14] Under these conditions, it is optimal in minimizing the expected variance of the error.[14] However, the quality of the estimate depends on the accuracy of the models used to represent the systems that is to be monitored. For the application to experiments in protein extension, the system was divided into a model of the cantilever deflection and a model of the protein tension. In series, the two models compose the entire system for study (Fig. 1).

In order to have an estimator for an unobserved state of the system, it is necessary to characterize the dynamic behavior of the tip of the AFM. As seen from Fig. 1(a), the cantilever deflection is the sole link between the protein contour length state and the experimental measurements. We start by describing a simple physical model for an AFM cantilever: a second order dynamic system in one dimension, which is characterized by the equation of motion

$$\frac{T}{m} = \ddot{y} + 4\zeta\pi f_n\dot{y} + 4\pi^2 f_n^2 y. \qquad (1)$$

This equation represents the continuous motion of a damped harmonic oscillator, where y is the deflection of the cantilever, f_n is the undamped natural frequency of the system expressed in hertz, ζ is the damping ratio, m is the point mass of the system, and T is the external force applied to the system. It is generally not possible to stimulate an AFM cantilever directly at its tip in order to generate the correct frequency response for evaluating the parameters in Eq. (1). However, the spectrum of the free thermal vibrations of the

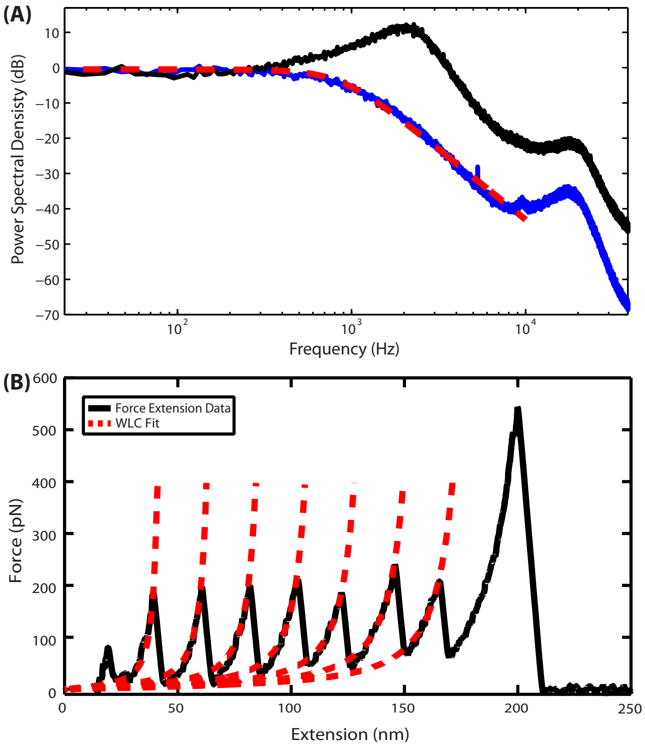

FIG. 2. (Color online) Mechanical properties of the AFM cantilever and of an extending polyprotein. (a) Under thermal equilibrium and free from any molecules attached to it, the typical AFM cantilever used in our experiments shows a principal resonant peak at ~1 kHz. It is much more pronounced with higher modes when the cantilever is free (black line) than when it is near the surface (blue line). A fit of the frequency response function of the simple cantilever model described in Eq. (3) (dotted red line), produces a reasonable description of the data derived from f_n=1.207 kHz and ζ=0.25. (b) Force-length relationship of a single ubiquitin polyprotein as it extends in a single molecule AFM experiment, as depicted in Fig. 1(a). The characteristic sawtooth pattern shape of these traces results from a series of sequential protein unfolding events which increase the contour length of the polyprotein at each peak, by an amount that is protein specific. The contour length increases are obtained by fitting the force-length relationship up to each force peak, using the WLC model of polymer elasticity [WLC, Eq. (4), dashed red lines]. Consecutive least-squares fits to the data show that the contour length increases from L_c=29.9 nm and up to L_c=179 nm in discrete increases of ~23–26 nm.

cantilever can be used as a substitute. This assumes that the cantilever was being stimulated by white thermal noise which was distributed throughout the cantilever. Although this noise stimulates all modes of the cantilever, the model being used only relies on the primary mode and therefore the spectrum was adequate for the system identification. In order to utilize this approach, a slightly different form of Eq. (1) was used

$$T \propto \ddot{F} + 4\zeta\pi f_n \dot{F} + 4\pi^2 f_n^2 F, \tag{2}$$

where F represents the force of the deflected cantilever and is related to the deflection through the cantilever spring constant k. Due to the unknown strength of the white noise process that is driving the system, there is already ambiguity in the constant of proportionality. However, with this formulation, it is clear that in the steady state the tension in the protein should balance the force in the cantilever. Therefore the noise spectrum can be normalized as in Fig. 2(a) in order to match the transfer function. Note that there is a substantial difference in the damping between the vibration of the cantilever several microns away from the surface [top trace in Fig. 2(a)], as compared with that observed ~200 nm from the surface [bottom trace in Fig. 2(a)]. Since the experiments occurred close to the surface, the latter spectrum was used to generate the cantilever model. Also, although a simple model was used that ignores many of the vibrational modes of the cantilever, it can be seen from Fig. 2(a) that the resulting model fits the power spectrum well in the frequency range that corresponds to the sampling rate being used in the experiments.

Equations (1) and (2) represent a continuous model; however, the experimental data is obtained as a discrete time series by sampling with an analog to digital converter. In order to make the model match the discrete nature of the measurements, Eq. (2) was converted to a second order discrete filter with transfer function

$$\frac{F(z)}{T(z)} = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}, \tag{3}$$

where z^{-1} refers to a unit delay at the sampling frequency. The particular values that represent the fit shown in Fig. 2(a) with a sampling frequency of 625 kHz, were b_1=0.334, b_2=0.192, a_1=−.0669, and a_2=0.196. This is the final form of the model utilized in the EKF.

In addition to simulating the cantilever, a model of the protein tension's dependence on contour length is needed. This is the second component of the system and is represented by a nonlinear function relating the end-to-end extension of the protein (Δx, see Fig. 1) and the contour length of the protein, L_c, to the tension T in the protein. In order to obtain a representative system model that can be used in our KF scheme, a reasonable approximation to the force-extension relationship of a protein is needed. The typical physical approach to modeling the tension of a protein is the WLC model of polymer elasticity.[5,6,9] This steady state model assumes that the protein behaves as a thin, flexible, inextensible string and that the sole cause of tension can be described by the entropy of the system. The WLC model of polymer elasticity is represented in Eq. (4).

$$T\left(\frac{\Delta x}{L_c}\right) = \frac{k_B t}{p}\left[\frac{1}{4}\left(1 - \frac{\Delta x}{L_c}\right)^{-2} - \frac{1}{4} + \frac{\Delta x}{L_c}\right], \tag{4}$$

where T is the protein tension, k_B is Boltzmann's constant, t is the temperature in Kelvin, p is the persistence length, and L_c is the contour length of the protein. The persistence length p is a parameter that describes the flexibility of the protein and is typically not well known. It can be thought of as the length over which a disturbance in the protein shape will propagate. The contour length of the protein measures the number of amino acids of the protein that are extended during the force-extension measurements.[5,6] The WLC model has been found to accurately describe the protein's force-extension relationship for values of $\Delta x/L_c > 0.7$. For smaller values of this extension ratio, not only entropy, but other phenomena such as hydrophobic interactions contribute to the force-length relationship in the protein.[10] This is apparent in Fig. 2(b) where the WLC model of polymer elasticity is shown to fit the force-extension relationship well near each unfolding peak, albeit with an increased value of the contour length after each unfolding event.

In order to implement the WLC model for the EKF, we estimated the most typical value of persistence length p for the particular protein being used, by normalizing and plotting a set of force-extension traces leading up to 130 different unfolding peaks. Each trace was normalized by its contour length L_c (assumed constant between unfolding events) and scaled by the ratio $\Delta x/L_c$. For values of the persistence length close to the correct one, these normalized traces collapsed onto a single curve. We found that a WLC function with $p=0.2$ nm provided a good fit through all the data sets, and we used it in the implementation of the KF described here. However, this approach is not required, and in many cases *a priori* knowledge of the persistence length is sufficient. Section II D considers the question of the persistence length further.

C. EKF estimates of the contour length of an unfolding protein

Due to the nonlinear tension model for the protein [Eq. (4)], we implemented an EKF (Ref. 15) to estimate the length of a protein in real time during an experiment. This implementation uses the previously described models of both the cantilever and the WLC tension function of the protein.

A brief outline of the EKF implementation for single molecule AFM experiments follows. For a more general treatment of EKFs, see for example Kwakernaak and Sivan.[11] In our experiments the position of the piezoelectric stage is considered a known input but the true cantilever deflection and contour length of the protein are not. Therefore the state vector $\underline{q}$ tracked by the EKF consists of the deflection and contour length states. The exact form of the state vector is dependent on the model used for the cantilever dynamics. Here, the second order cantilever model results in dependencies in the previous two time steps of the deflections and contour lengths. Thus the state vector $(\underline{q})$ consists of the current and previous values of the cantilever deflection and contour length

$$\underline{q}_t = [X_t \;\; X_{t-1} \;\; L_t \;\; L_{t-1}]^T.$$

The subscripts refer to the time step of the variable. Thus X_t and L_t refer to the cantilever deflection and contour length at time step t, respectively.

Based on the models defined in the previous section, the state transition equation can then be written as follows:

$$\underline{q}_{t+1} = \begin{bmatrix} X_{t+1} \\ X_t \\ L_{t+1} \\ L_t \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{1} \left[\frac{1}{k} b_{j+1} W\left(\frac{X_{t-j} - u_{t-j}}{L_{t-j}} \right) - a_{j+1} X_{t-j} \right] \\ X_t \\ L_t \\ L_t \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w_{1,t} \\ w_{2,t} \end{bmatrix}, \tag{5}$$

where the function $W(\cdot)$ refers to the WLC model described earlier, k is the cantilever spring constant, u_t is the piezoelectric stage position at time t, and the $w_{i,t}$ are Gaussian noise variables at time t. The indexed coefficients a_i and b_i are those from the cantilever model [Eq. (3)]. This equation describes how the state vector at one time step is related to the state vector at the next. Besides using a combination of the cantilever and WLC models to express the next cantilever deflection variable, Eq. (5) also predicts that the contour length does not change (except by noise). Clearly this is incorrect in the scale of an entire experiment but on the local scale it includes the knowledge that the contour length does not change when the protein is not unfolding. The addition of white noise is necessary in order to keep the estimate of the contour length from converging to a fixed point. The noise vector $\underline{w}_t$ has a diagonal covariance matrix Q.

The result of an experiment is a measurement of the force on the cantilever, therefore a simple linear measurement equation can be used

$$F_t = H\underline{q}_t + v_t = [k \;\; 0 \;\; 0 \;\; 0]\underline{q}_t + v_t. \tag{6}$$

Here k is again the cantilever spring constant and v_t is the Gaussian measurement noise variable at time t, independent of the other noise and of itself at earlier times. The noise v_t has variance R.

The EKF algorithm estimates the state recursively in successive update and prediction steps. There are four quantities tracked by the algorithm: $\hat{\underline{q}}_{t|t}$, $\hat{\underline{q}}_{t|t-1}$, $P_{t|t}$, and $P_{t|t-1}$. Here, $\hat{\underline{q}}_{t|t}$ and $\hat{\underline{q}}_{t|t-1}$ are the estimates of the state vectors at time t based on all the measured values of the cantilever force up to time t and $t-1$, respectively. Similarly, $P_{t|t}$ and $P_{t|t-1}$ are estimates of the error covariance matrices based on the measurements of the cantilever force up to time t and $t-1$, respectively.

The update equations compute $\hat{\underline{q}}_{t|t}$ and $P_{t|t}$ from $\hat{\underline{q}}_{t|t-1}$ and $P_{t|t-1}$

$$\hat{\underline{q}}_{t|t} = \hat{\underline{q}}_{t|t-1} + K_t(F_t - H\hat{\underline{q}}_{t|t-1}), \tag{7}$$

$$P_{t|t} = P_{t|t-1} - K_t H P_{t|t-1}, \tag{8}$$

where $K_t = (P_{t|t-1}H^T)/(HP_{t|t-1}H^T + R)$ is the Kalman gain, H is the output vector from the measurement Eq. (6), and F_t is the measurement at time t.

The prediction equations compute $\hat{\underline{q}}_{t+1|t}$ and $P_{t+1|t}$ from $\hat{\underline{q}}_{t|t}$ and $P_{t|t}$. The prediction of $\hat{\underline{q}}_{t+1|t}$ comes directly from the nonlinear state transition Eq. (5) above. In order to predict the *a priori* error covariance matrix for the next time step, the state transition equation must be linearized into the form $\underline{q}_{t+1} = A_t\underline{q}_t + B_t\underline{w}_t$. B_t is already linear from Eq. (5). The linearized state transition matrix A_t is the Jacobian evaluated at $\hat{\underline{q}}_{t|t}$, given explicitly by

$$A_t = \begin{bmatrix} J_{11} & J_{12} & \frac{b_1}{k}\left.\frac{\partial W}{\partial L}\right|_{\hat{X}_{t|t},\hat{L}_{t|t},u_t} & \frac{b_2}{k}\left.\frac{\partial W}{\partial L}\right|_{\hat{X}_{t-1|t},\hat{L}_{t-1|t},u_{t-1}} \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$J_{11} = \frac{b_1}{k}\left.\frac{\partial W}{\partial X}\right|_{\hat{X}_{t|t},\hat{L}_{t|t},u_t} + a_1 \quad J_{12} = \frac{b_2}{k}\left.\frac{\partial W}{\partial X}\right|_{\hat{X}_{t-1|t},\hat{L}_{t-1|t},u_{t-1}} + a_2. \tag{9}$$

The prediction equation for $P_{t+1|t}$ is then given by

$$P_{t+1|t} = A_tP_{t|t}A_t^T + B_tQB_t^T. \tag{10}$$

This algorithm can be run in real time with the experiment, and provides estimates of the contour length taking into account all the available data in $\hat{\underline{q}}_{t|t}$. In order to begin the algorithm, all that is needed is an initial estimate of the state (such as the expected value of the deflection and contour length) and the error covariance matrix (such as the identity matrix). The only caveat is that the ratio of the extension to the contour length must always be less than one according to the protein model. In general, a good estimate for the noise variance in the measurement may be obtainable but it is much more difficult to obtain reasonable values for the variances in Q. In practice, the ratio of these two values determines the tradeoff between convergence speed and noise rejection and must be tuned at the start to obtain the peak performance.

D. EKF recovers ΔL_c from synthetic data.

In order to investigate the behavior of the EKF implementation with the described models, the algorithm was applied to simulated data in which the true value of the contour length was known exactly (Fig. 3). The simulated data was generated using the polymer and cantilever models described in Sec. II B. To simulate protein unfolding, two stepwise increases in the value of $L_c = 40$ and 30 nm were introduced at different times. We added Gaussian noise with a variance of 15 pN to simulate the experimental conditions. This represents an ideal case in which the model used in the EKF implementation is exactly the same as that which generated the synthetic data. Not surprisingly, the algorithm described in Eqs. (5)–(10) accurately recovers the step increases in contour length introduced during the simulation (Fig. 3). The simulations also allowed us to investigate the effect of a mismatch in the persistence length, which is a parameter that is not known *a priori* and is not estimated by our EKF implementation. As is clear from Fig. 3(b), the persistence length determines the convergence characteristics of the state estimate but not whether it converges. A mismatch in the persistence length between the system and the KF results in an initial over or underestimate of the contour length, depending on the sign of the mismatch. In all the cases, however, the estimate of the contour length improves as the protein is stretched further, until the next unfolding event. This implies that small errors in matching the persistence length will not significantly affect the contour length estimation, especially if using the most reliable estimate, just before the unfolding event. In addition, this presents a manner for obtaining an estimate for the persistence length from a few experimental traces. Minimizing the absolute value of the slope of the contour length state estimate convergence of the KF between unfoldings will lead to an approximation for the true persistence length. This approach was implemented and found to work adequately (not shown). In addition, it may be possible to augment the state vector of the KF to include the persistence length.

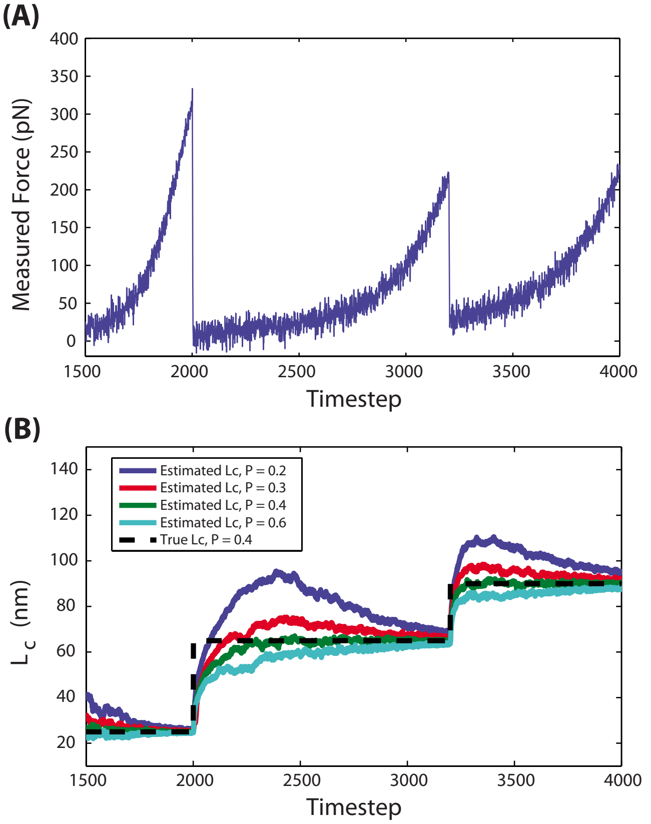

FIG. 3. (Color online) Analysis of the extended KF implementation. (a) Simulated data created for the purpose of examining the behavior of the observer, generated by stepwise changes in the contour length and utilizing the same system model described in the text. Independent Gaussian noise with a variance of 15 pN was used to simulate the measurement process. (b) A comparison of the contour length estimates from extended Kalman filters based upon different values of persistence length, compared against the true contour length (dotted line). The convergence behavior of the estimate is strongly dependent on the value of the persistence length. For the true value of the persistence length, the estimate quickly converges to the true L_c. If the persistence length is off, slower convergence results and the manner of convergence identifies whether the persistence length was too large or small.

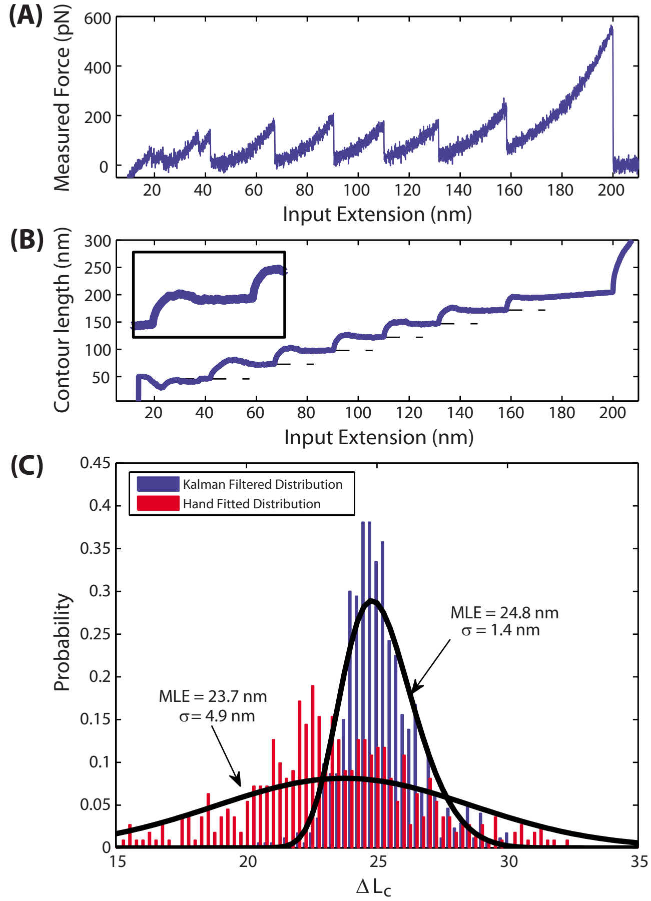

FIG. 4. (Color online) Experiments with ubiquitin polyproteins at an extension rate of 400 nm/s produced 190 sawtooth traces for analysis. (a) A sample trace from the data set. The final peak is a result of the dissociation of the protein from the cantilever tip. These peaks were excluded from the analysis in L_c step sizes. (b) The resulting estimate of the contour length of the protein, corresponding to the data in part (a). The inset enlarges one of the steps in L_c, showing that the convergence behavior does not match any of those from Fig. 3(b). The overshoot that occurs immediately after the step most likely implies an error in the protein model and confirms the expected failure of the WLC model at low forces. (c) The distribution and statistics of the estimated changes in contour lengths during unfolding compared with earlier results fit by hand with the WLC model, as in Fig. 2(b). The data for the hand-fitted steps is from Carrion-Vazquez *et al.* (Ref. 17). A noticeable skew is observed in the EKF step size histogram. Therefore the fit plotted is not a Gaussian distribution as in the case of the hand-fitted data but a generalized extreme value distribution. The parameters of the distribution are: $\xi=-0.15$ $\sigma=1.3$ and $\mu=24.6$. ξ, σ, and μ are the shape, scale, and location parameters, respectively. The maximum likelihood estimates and variances are listed in the figure for each approach. In the case of the Gaussian fit, the maximum likelihood estimate is equivalent to the mean.

E. EKF measurements of contour length increments in an unfolding polyprotein

The application of the EKF to real data yields an extraordinary set of results (Fig. 4). We applied the EKF to a set of 190 force-extension traces obtained from single polyubiquitin proteins as described previously.[16] All the analyzed force-extension traces have the unmistakable fingerprint of a sawtooth pattern.[2,3] Applying the EKF algorithm to a force-extension trace with a sawtooth pattern [Fig. 4(a)] produces a very distinct staircase of contour length increases estimated by the EKF [Fig. 4(b)]. This procedure was repeated for each sawtooth pattern trace. The actual increase in contour length was referenced to just prior to the next unfolding event, as the EKF estimate converged to a well-defined value. Interestingly, the contour length estimates overshoot significantly at low forces, converging rapidly at higher forces [Fig. 4(b), inset]. This effect was not seen when analyzing the simulated data and does not resemble a mismatch in the persistence length. It was observed even when the persistence length was varied over the range (0.1–2 nm) and in addition it is different in character than the types of deviations observed with mismatched persistence lengths [see Fig. 3(b)]. It is likely that the observed deviations are a reflection of the fact that the WLC model does not correctly describe the physics of a protein at low forces. This is not a surprise given that the WLC model assumes that the behavior of the molecule is driven solely by entropy, ignoring the large enthalpic contributions that dominate when a protein collapses at low force.[10] However, the rapid convergence of the contour length estimates in the high force range is in agreement with the observation that the physics of a highly extended protein is dominated by entropic elasticity.

A histogram of the EKF estimates of the contour length increments measured from these recordings ($n=693$) was fit with a generalized extreme value distribution, in order to take into account the substantial skew observed in the distribution [blue bars; Fig. 4(c)]. This fit made it possible to more accurately identify the maximum likelihood estimate, which was found to be 24.8 nm. The standard deviation of the data was 1.4 nm. This result can be compared with an earlier evaluation of ubiquitin unfolding data, where the traces were low-pass filtered at ~175 Hz before they could be manually fitted with the WLC. Despite the extensive filtering and careful visual validation of each fit, this procedure yielded an estimate of ΔL_c with mean 23.7 nm and a standard deviation of 4.9 nm [Fig. 4(c); red bars from Carrion-Vazquez *et al.*[16]]. Clearly, the fully automated EKF approach not only saves time and effort but also markedly decreases the variance of the data.

III. DISCUSSION

Kalman filters have been extensively used in control systems and instrumentation[11,17] but only rarely in the analysis of single molecule data.[18] Here we demonstrate the application of an EKF in the measurement of the contour length of a single polyprotein during a single molecule force spectroscopy experiment. Measurements of the contour length of an unfolding protein can precisely locate the mechanical transition state and count the number of amino acids involved in the ensuing extension. Estimates of contour length are obtained by fits of the WLC model of polymer elasticity. Indeed, fits of the WLC model to protein unfolding traces have become a standard practice in force spectroscopy by AFM[2,6] and also by optical tweezers.[7,19] These fits can be done manually, by adjusting the persistence length and contour length of the WLC model until a visually agreeable trace is produced. Alternatively, a setpoint or two boundaries can be placed in a trace and a fit can be generated using algorithms such as Levenberg–Marquardt that minimizes the squares of the deviations. However, in all these cases, the operator is free to choose the location of the setpoints as well as the evaluation of what constitutes a good fit. Ideally, an independent model based automated observer can measure state variables such the contour length of a protein, without the intervention of an operator. As we have shown here, this can be realized using an extended Kalman filter algorithm that uses a model of the cantilever-protein system to estimate of the state variables describing the extension of a single protein. Such an EKF implementation can be very effective in a hands-off measure of contour length increments from sawtooth pattern data. Furthermore, given that the EKF requires only the current and immediately past values of a force-extension relationship, estimates can be obtained in real time during an experiment. This feature can be of value in automated measurements.

The biggest limitation of our EKF algorithm is its reliance on models that accurately represent the physics of the cantilever and the protein. While the mechanical behavior of a cantilever can be complex, its physics is essentially understood. By contrast, the use of the WLC model of polymer elasticity to describe the behavior of a protein can only be considered as a first approximation. Indeed, the physics of a collapsed protein is not well understood.[10] One important alternative in implementing the EKF is that its design need not be grounded in models such as the WLC. That is, absent a thorough understanding of the physics underlying the systems, a nonparametric description [such as Eq. (3)] fit to experimental data suffices. However, such an approach would be useful solely from the systems control point of view in an automated instrument. Nevertheless, as the models of force-length relationship for extended proteins improve, new implementations of the EKF may be possible where several relevant state variables of the proteins being studied, are monitored in real time during an experiment. Such an advance would open up extraordinary opportunities for monitoring the conformational dynamics of a single protein, in real time.

[1] G. Binnig, C. F. Quate, and C. Gerber, Phys. Rev. Lett. **56**, 930 (1986).

[2] M. Rief, M. Gautel, F. Oesterhelt, J. M. Fernandez, and H. E. Gaub, Science **276**, 1109 (1997).

[3] M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **96**, 3694 (1999).

[4] H. B. Li, W. A. Linke, A. F. Oberhauser, M. Carrion-Vazquez, J. G. Kerkviliet, H. Lu, P. E. Marszalek, and J. M. Fernandez, Nature (London) **418**, 998 (2002).

[5] M. Carrion-Vazquez, P. E. Marszalek, A. F. Oberhauser, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **96**, 11288 (1999).

[6] S. R. Ainavarapu, J. Brujic, H. H. Huang, A. P. Wiita, H. Lu, L. Li, K. A. Walther, M. Carrion-Vazquez, H. Li, and J. M. Fernandez, Biophys. J. **92**, 225 (2007).

[7] M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier, and C. Bustamante, Science **276**, 1112 (1997).

[8] M. Rief, J. M. Fernandez, and H. E. Gaub, Phys. Rev. Lett. **81**, 4764 (1998).

[9] C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, Science **265**, 1599 (1994).

[10] K. A. Walther, F. Grater, L. Dougan, C. L. Badilla, B. J. Berne, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **104**, 7916 (2007).

[11] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems* (Wiley, New York, 1972).

[12] M. Schlierf, H. Li, and J. M. Fernandez, Proc. Natl. Acad. Sci. U.S.A. **101**, 7299 (2004).

[13] E. L. Florin, M. Rief, H. Lehmann, M. Ludwig, C. Dornmair, V. T. Moy, and H. E. Gaub, Biosens. Bioelectron. **10**, 895 (1995).

[14] R. E. Kalman, Trans ASME J. Basic Eng. **82**, 35 (1960).

[15] H. Cox, IEEE Trans. Autom. Control **9**, 5 (1964)

[16] M. Carrion-Vazquez, H. B. Li, H. Lu, P. E. Marszalek, A. F. Oberhauser, and J. M. Fernandez, Nat. Struct. Biol. **10**, 738 (2003).

[17] B. Mokaberi and A. A. G. Requicha, IEEE Trans. Autom. Sci. Eng. **3**, 199 (2006).

[18] L. S. Milescu, A. Yildiz, P. R. Selvin, and F. Sachs, Biophys. J. **91**, 3135 (2006).

[19] C. Cecconi, E. A. Shank, C. Bustamante, and S. Marqusee, Science **309**, 2057 (2005).