Kalman Filter Estimates of the Contour Length of an Unfolding Protein in Single-Molecule Force Spectroscopy Experiments

Vicente I. Fernandez[1] , Pallav Kosuri[2], Vicente Parot[4] and Julio M. Fernández[3]

[1]Department of Mechanical Engineering, Massachusetts Institute of Technology, Boston 02139

[2]Dept of Biochemistry and [3]Department of Biological Sciences, Columbia University, New York 10027, [4]Pontificia Universidad Católica de Chile, Santiago

Abstract

Force spectroscopy measurements of single molecules using AFM have enabled the study of a range of molecular properties not accessible with bulk methods. These properties of interest must typically be inferred by manually fitting models to selected portions of measured data. As manual intervention in the fitting process easily introduces a bias in the analysis, there is a need for more sophisticated analysis methods capable of interpreting data in an unbiased and repeatable "hands-off" manner. Here we apply an extended Kalman filter to the estimation of protein contour length (Lc) during mechanical unfolding, based on force and extension data from an AFM experiment. This filter provides an online and fully automated estimate of Lc based on a system model, the experimental measurements, and noise statistics. The system model comprises a physical model of the cantilever and a nonlinear WLC approximation of the extended protein. When manually fitting the WLC model to force-extension data from ubiquitin proteins, the estimate of the change in contour length during unfolding is distributed normally with mean 23.3 nm and variance 10.2 nm². Testing the Kalman filter on the same protein yields ΔLc with a 24.8 nm mean and 2.0 nm² variance. As the variance limits resolution in estimating the number of amino acids released by unfolding, it is clear that the Kalman filter presents a substantial improvement over the conventional method. We thereby demonstrate that the Kalman filter provides a powerful unbiased approach to interpreting force spectroscopy data, capable of increasing resolution beyond the traditional experimental limit. Due to the flexibility of this approach, it can be extended to monitoring other state variables of molecular systems observed by various forms of force spectroscopy, including optical and magnetic tweezers.

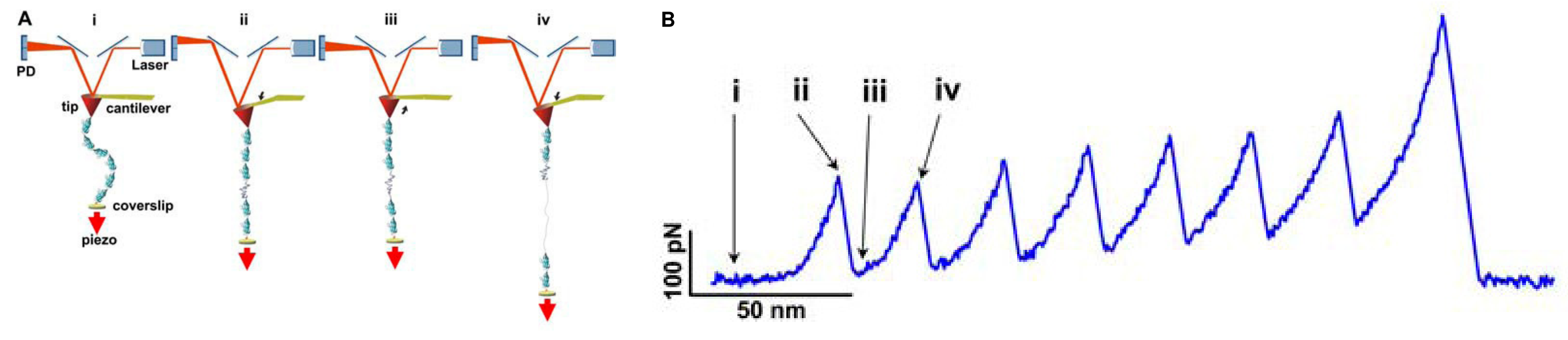


Figure 1 Mechanical stretching of a polyprotein using single-molecule atomic force microscopy. **(A)** *(i)* A single polyprotein molecule is held between the cantilever tip and the coverslip, whose position can be controlled with high precision using a piezoelectric positioner (piezo). *(ii)* Moving the coverslip away from the tip exerts a stretching force on the polyprotein, which in turn bends the cantilever. The bending of the cantilever changes the position of the laser beam on the split photo diode (PD), registering the pulling force. The applied force can be determined from the spring constant of the cantilever and the degree of cantilever bending. At this high pulling force, a protein domain unfolds. *(iii)* The unfolded domain can now readily extend, relaxing the cantilever. *(iv)* The piezo continues to move, stretching the polyprotein to a new high force peak, repeating the sequence until the whole polyprotein has unfolded. This process results in a force extension curve with a characteristic sawtooth pattern shape. **(B)** A typical sawtooth pattern curve obtained by stretching an I27 polyprotein. The labels *i–iv* represent the sequence of events shown in *A*.

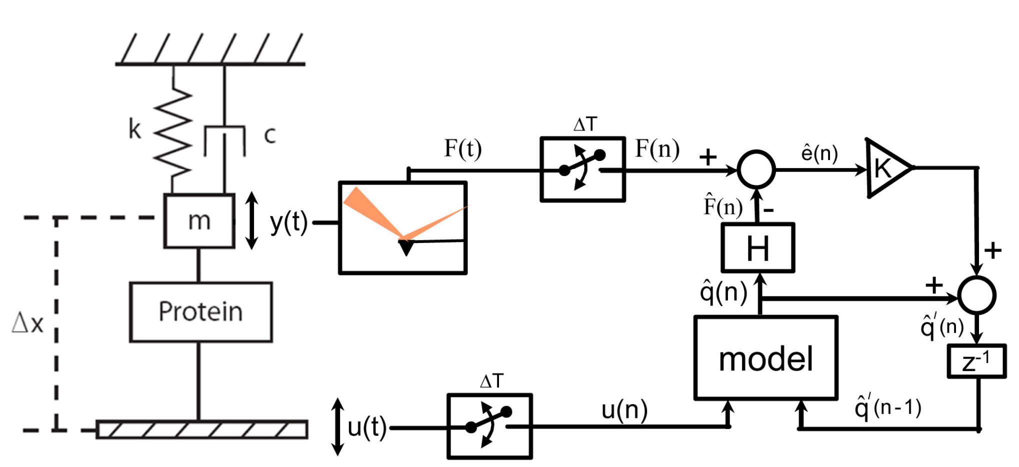


Figure 2 Schematic depiction of the Extended Kalman Filter (EKF) implementation for single protein force spectroscopy with an AFM. The EKF estimates the current contour length of the protein based upon the force measurements (F_t) up to the present time. It is an extension of the Kalman filter for systems with nonlinear dynamics. The Kalman filter is an optimal estimation algorithm given a known linear system with Gaussian noise.

In general, the extended Kalman filter (EKF) tracks an estimate of the state vector q_n The state vector is composed of the contour length of the protein and the position of the cantilever, which are the hidden variables that fully determine the system. The algorithm uses a model of the system to predict the measurement at timestep *n* based on the input at *n-1* and the estimate at *n-1*. The error between the predicted value and the true measured value at time *n* is then used to update the estimate with a gain *K* that is determined by the EKF algorithm. In order to optimally determine *K* for each timestep, an estimate of the covariance matrix is also tracked by the algorithm.

Discussed in Figure 3, the protein is modeled using the Worm-Like Chain (WLC) model of polymer elasticity, given by equation **(1)**. The WLC model provides the tension force in the protein (F_p) given the extension of the protein (*y-u*) and the contour length (*Lc*). The cantilever model in turn is specified by the transfer function between the input protein tension and the output cantilever force **(2)**. The corresponding coefficients are given in equations **(3)** and **(4)**. As mentioned in Figure 3, the result is a filter that depends on the previous two timesteps.

$$\text{(1)}\quad F_p = W\left(\frac{y-u}{Lc}\right) = \frac{k_B T}{p}\left(\frac{1}{4}\left(1-\frac{y-u}{Lc}\right)^{-2} - \frac{1}{4} + \frac{y-u}{Lc}\right)$$

$$\text{(2)}\quad \frac{F(z)}{F_p(z)} = \frac{b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}}$$

$$\text{(3)}\quad a = [-0.669\ \ 0.196]$$

$$\text{(4)}\quad b = [0.334\ \ 0.192]$$

$$\text{(5)}\quad q_n = [y_n \quad y_{n-1} \quad Lc_n \quad Lc_{n-1}]^T$$

$$\text{(6)}\quad q_{n+1} = \begin{bmatrix} y_{n+1} \\ y_n \\ Lc_{n+1} \\ Lc_n \end{bmatrix} = \begin{bmatrix} \sum_{j=0}^{1}\left(\frac{1}{k} b_{j+1} W\left(\frac{y_{n-j} - u_{n-j}}{Lc_{n-j}}\right) - a_{j+1} y_{n-j}\right) \\ y_n \\ Lc_n \\ Lc_n \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \bullet \begin{bmatrix} w_{1,n} \\ w_{2,n} \end{bmatrix}$$

$$\text{(7)}\quad F_n = [k \quad 0 \quad 0 \quad 0] \cdot q_n + v_n$$

Equation **(5)** shows the state vector used in this implementation. It is composed of the cantilever deflection (y) and the contour length of the protein being stretched (Lc). The current and previous values of y and Lc are both included in the state vector, as required by the cantilever model. The full system model used by the EKF algorithm is described by equations (6) and (7). Equation **(6)** describes the progression of the state vector. Both the cantilever deflection and the contour length are assumed to have independent white Gaussian process noise ($w_{i,n}$) sources. Apart from the noise, the contour length is modeled as constant. Although this is clearly incorrect globally, it models the period between unfolding events accurately. The cantilever deflection is updated by the combination of cantilever dynamics and WLC models. The nonlinearity in the WLC model is the reason that an extended Kalman filter must be used as opposed to a regular Kalman filter. Equation **(7)** is the measurement equation, linking the state variables to the experimentally measured quantity. In the measurement, there is an additional source of Gaussian noise (v_n). These equations fully define the problem for the application of the EKF algorithm. The algorithm itself is standard and can be found in texts such as [REF 1].

The EKF algorithm only utilizes the measurements that have already been made in order to create its estimate of the state variables. As a result, it can be run concurrently with the experiment itself. In the current analysis, the estimation was done after the experiment on a batch of traces. The causal estimation results are shown in Figure 5. These results show a substantial improvement over the commonly used hand-fitting methods.

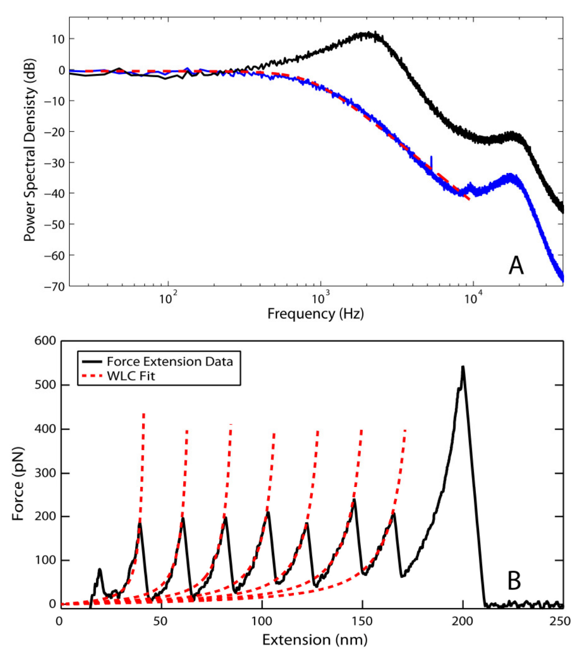


Figure 3 Models describing the experimental system for the extended Kalman filter implementation. The system model is divided into two sequential parts: A linear model of the cantilever dynamics driven by the output of a nonlinear protein model. **(A)** The cantilever model consists of a second order LTI system fit to the noise spectrum of a free cantilever. This assumes the noise is white and acts solely on the tip of the cantilever. It is important to distinguish the heavily damped cantilever spectrum near a surface (blue line) from the spectrum far from a surface (black line). **(B)** The protein forces are modeled by the Worm-Like Chain (WLC) model of protein elasticity. This model is known to fit the protein force-extension curve well for high forces only. Previously, data was fit with the WLC model as shown, where the unfolding step size and persistence length are chosen manually to obtain the best fit over the entire trace. For this example, ΔLc = 23.1 nm, and p = 0.37 nm.

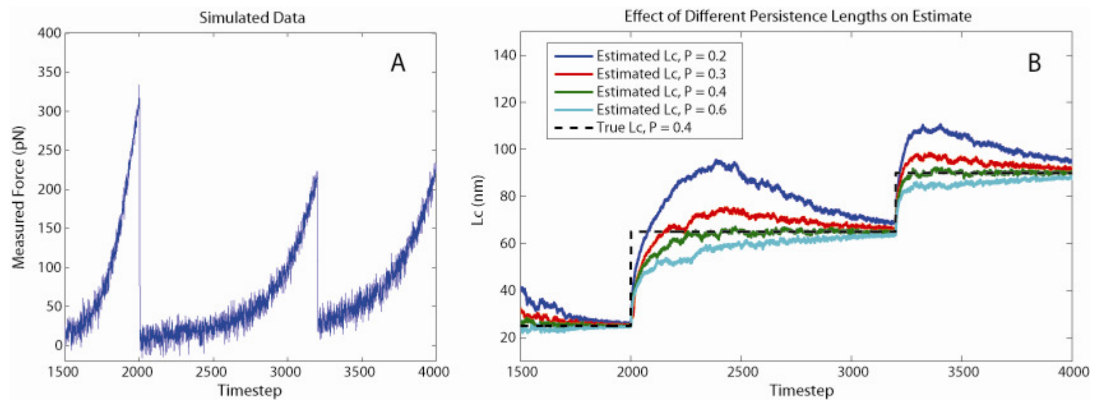


Figure 4 Kalman filter estimation applied to simulated data. (A) Simulated data generated directly from the cantilever and WLC models with a persistence length of 0.4 nm and a predetermined stepwise constant contour length (Lc). White Gaussian measurement noise with a standard deviation of 10 pN has been added to the simulation. **(B)** EKF estimates of the Lc based on various persistence lengths, against the true Lc (dotted line). The convergence behavior 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 Lc. If the persistence length is off, slow convergence results, with a slope that indicates the sign of the error.

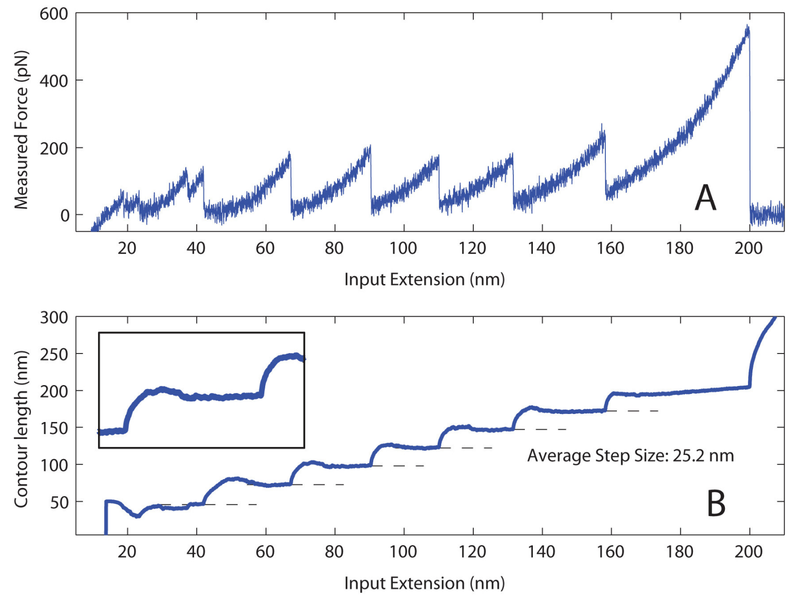
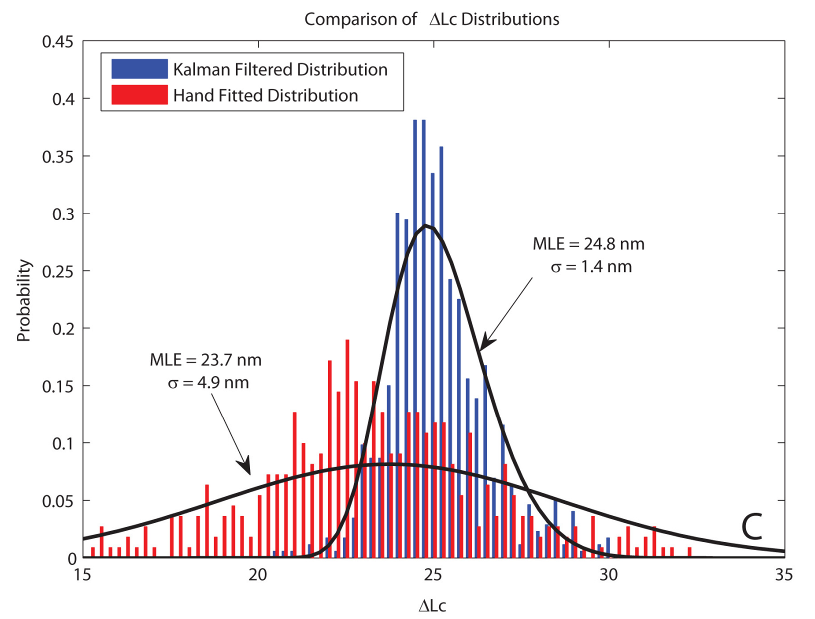


Figure 5. Results of EKF implementation on experimental data. Experiments with ubiquitin polyproteins at an extension rate of 400 nm/s produced 190 sawtooth traces for analysis. **(A)** A sample trace from the data set. The final peak is a result of the dissociation of the protein from the cantilever tip. These peaks were excluded from the analysis in Lc step sizes. **(B)** The resulting estimate of the contour length of the protein, corresponding to the data in part (A). A persistence length of 0.4 nm was chosen for the protein model, which is the commonly used value for ubiquitin. The inset enlarges one of the steps in Lc, showing that the convergence behavior does not match any of those from Figure 3. The overshoot that occurs immediately after the step 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 figure 2. The data for the hand-fitted steps is from [REF 2]. The spread of EKF estimates is substantially narrower, with a standard deviation less than half that of the hand-fitted distribution. A noticeable skew is observed in the EKF step size histogram. Therefore the fit plotted is not a Gaussian distribution, but a generalized extreme value distribution. The parameters of the distribution are: *ξ = -0.15 σ = 1.3* and *μ = 24.6*. *ξ*, *σ*, and *μ* are the shape, scale, and location parameters respectively. Though preliminary, this may be linked to the underlying protein mechanics in which large contour lengths are preferentially chosen among available configurations.



Conclusions

- **Implementing a Kalman filter enhances the estimate of the stepwise increases in contour length of unfolding proteins by providing more consistent and accurate results**
- **This approach provides an unbiased "hands off" way of extracting information in real time on molecular properties**
- **The choice of persistence length may be made before the experiment, or it can be automatically chosen in post-analysis, making the procedure fully independent of the experimenter**
- **The on-line operation of the Kalman filter opens the door to numerous applications such as improved feed-back systems**
- **The Kalman filter should be utilized more commonly in providing unbiased estimations of hidden state variables in single molecule experiments**

References:

[1] Mohinder S. Grewal and Angus P. Andrews, *Kalman Filtering – Theory and Practise using MATLAB, 2nd edition,* Wiley-Interscience, 2001

[2] Mariano Carrion-Vazquez et al., *The mechanical stability of ubiquitin is linkage dependent*, Nat Struct Biol, 2007