Force-clamp analysis techniques reveal stretched exponential unfolding kinetics in ubiquitin

Herbert Lannon*, Eric Vanden-Eijnden+, and J. Brujic*
*Department of Physics and Center for Soft Matter Research, New York University, 4 Washington Place, New York, NY, 10003, USA
+Courant Institute of Mathematical Sciences, New York University, New York, NY, 10012

Supporting Materials

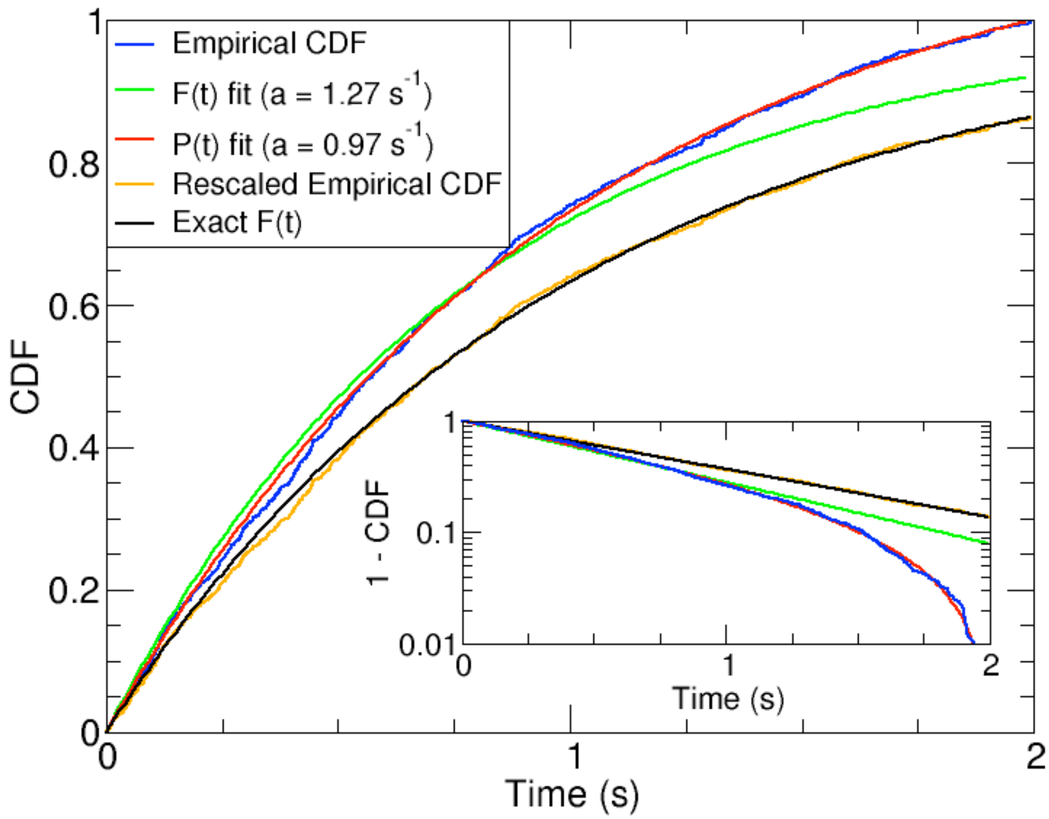

Figure S1. Comparison of fits of P(t) and F(t). A synthetic data set is generated by drawing single exponential dwell times with a rate constant of *a* = 1 sec^{-1} and only keeping those times that are less than t_c =2 sec. The empirical CDF of the generated data, hat P(t) [Eq. (1) in the main text], is shown in blue. Fitting using P(t) (red) [Eq. (2) in the main text] recovers the original constant *a* within 3%, while fitting with F(t) gives rise to a 27% discrepancy. By properly rescaling the empirical CDF (orange) using Eq. (2), the exact F(t) is recovered (black). Log scale of 1 - CDF in the inset highlights the changes in shape between F(t) and P(t).

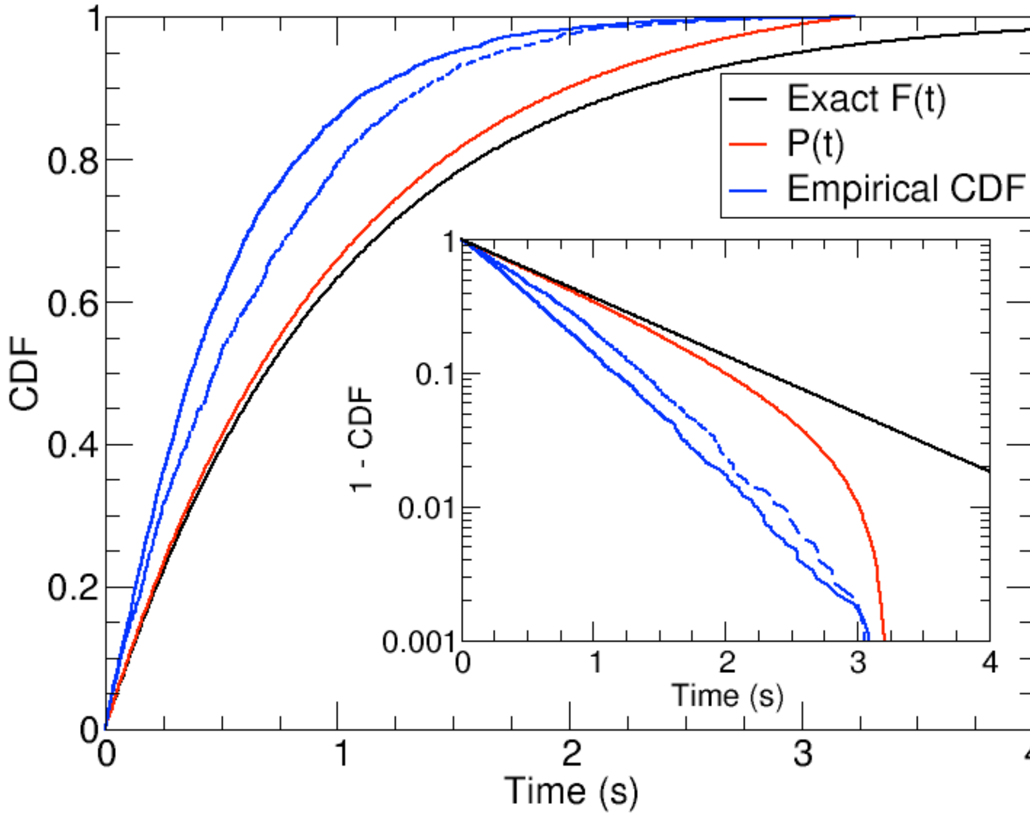

Figure S2. Mimicking the effects of random detachment time and polyprotein length. A synthetic dataset is generated by: (i) picking at random a chain of length N and a detachment time t_d, (ii) generating N dwell times from a single exponential distribution with a rate constant *a* = 1 sec^{-1}, (iii) keeping only those dwell times that are less than t_d, and (iv) repeating these operations. Operations (i)-(iii) mimic the pulling of a single polyprotein, and by repeating them we generate 1000 dwell times (like in the experiment reported in the main text). The exact exponential F(t) from which the times were picked is shown in black. The full and dashed blue lines are the two empirical CDFs obtained by the procedure above with two different distributions p(N) for N. These CDFs are different from one another, and different from the empirical CDF hat P(t), shown in red, that was generated as in Fig. S1, with a t_c =3.25 sec comparable to the largest dwell time kept in the blue data sets. The inset shows the CDFs plotted on a log scale. We can see large deviations between these CDFs due to the experimental biases from N and t_d. How to remove these biases by filtering the datasets is explained in the main text and in Fig. S4.

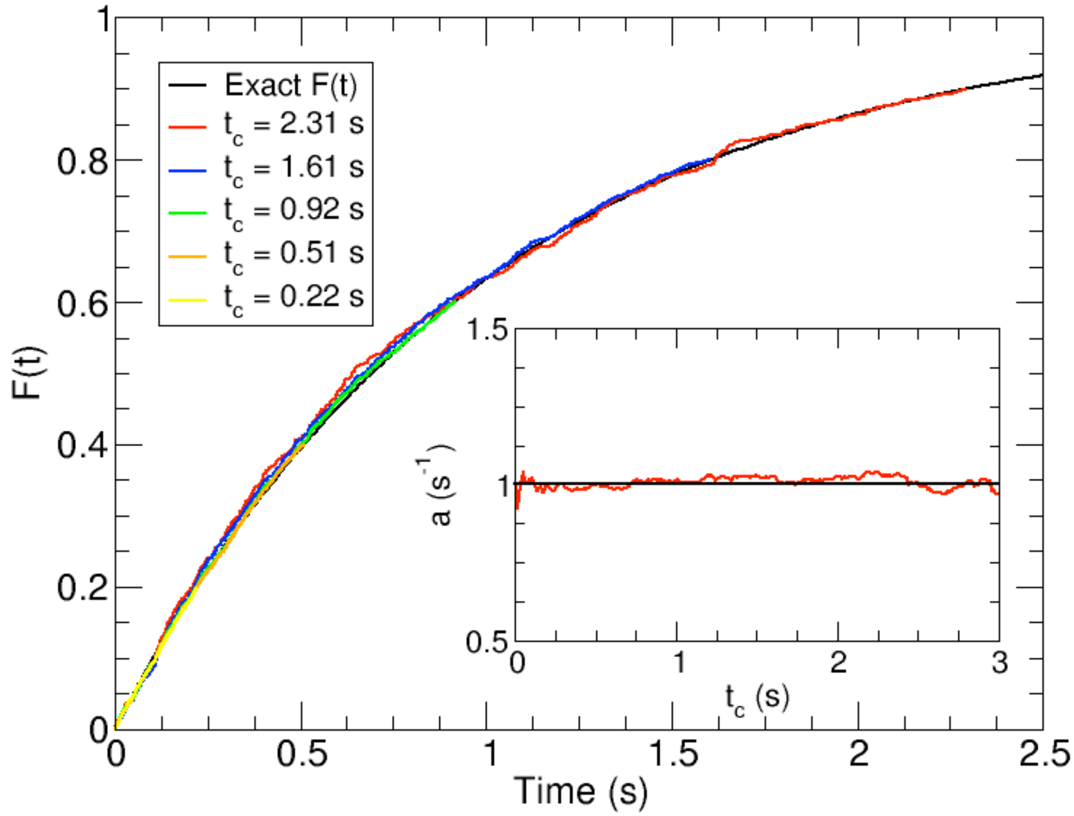

Figure S3. Correct filtering procedure. One of the blue datasets shown in Fig. S2 is used to construct empirical CDFs by the filtering procedure described in the main text: keep only the dwell times that (i) are less than a given cutting time t_c and (ii) come from traces with $t_d > t_c$. The exact F(t) is shown in black. The solid color curves are the empirical CDFs obtained with different cutting times t_c, and rescaled by $F(t_c)$. The inset is a plot of the rate constant a found by fitting the empirical CDFs, showing that the filtering procedure gives robust values for a at all t_c.

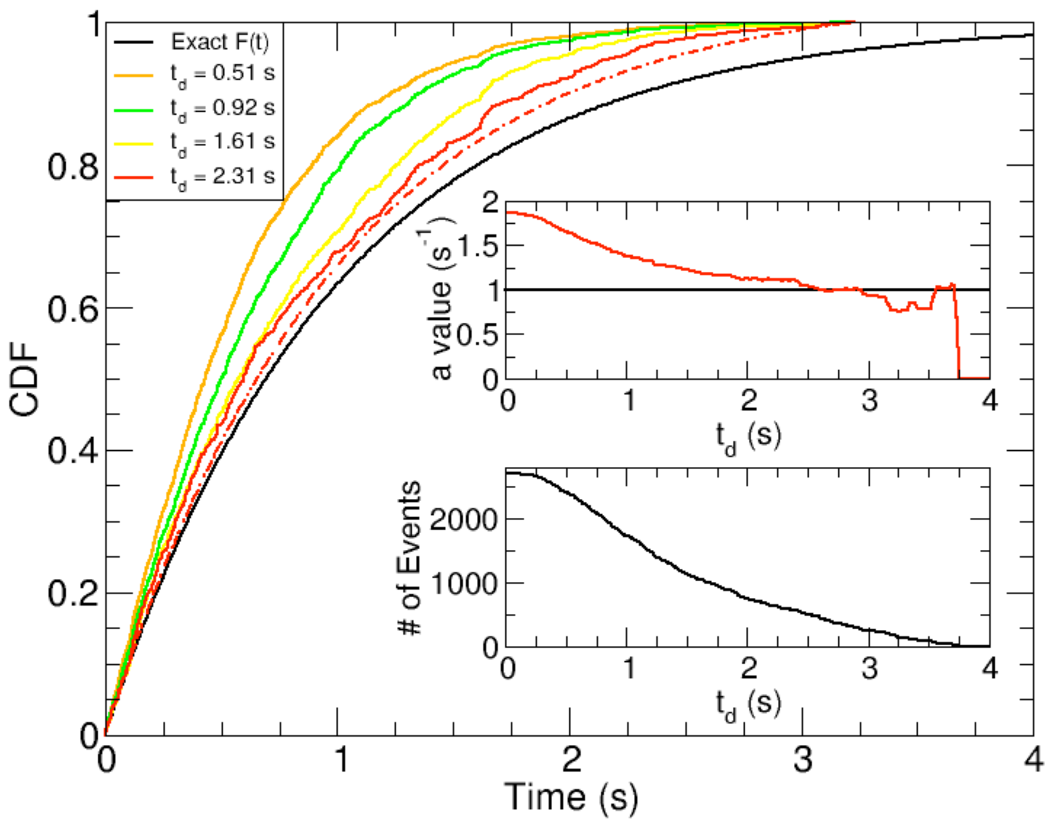

Figure S4. Naive filtering procedure. One of the blue datasets shown in Fig. S2 is used to construct empirical CDFs by a naive filtering in which all the dwell times less than t_d are kept (solid colored curves). The exact F(t) is shown in black. Due to the experimental biases from N and $t_{d,}$ each empirical CDF is different from the corresponding P(t) obtained by setting $t_{max} = t_d$ in Eq. (2) in main text, and they only begin to match as t_d is increased: the dashed line shows P(t) for $t_{max} = t_d = 2.31$s. The top inset is a plot of the rate constant *a* found by fitting the empirical CDFs, showing that this naive filtering procedure cannot recover the correct value for *a*. The bottom inset is a plot of the number of points remaining in the dataset at different t_d.

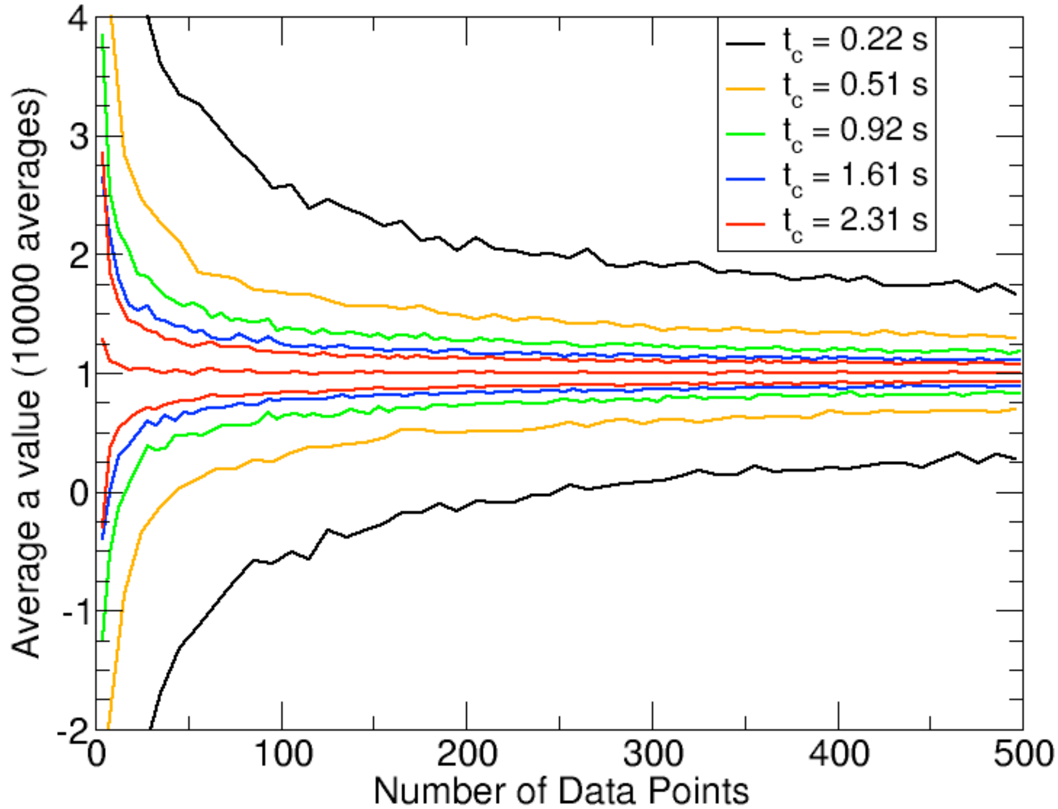

Figure S5. Error analysis. Rate constant value found by MLE fitting to empirical CDFs constructed using Eq. (2) in the main text with different $t_{max} = t_c$, effectively mimicking an experiment with a finite ending time. Dwell time datasets of varying size were generated using an exponential distribution with a = 1 sec^{-1}. At each dataset size, 1000 datasets where generated and fit to give an average value and a standard deviation for the rate constant found. The fitting at each t_c recovers the rate constant used in generating the data, seen as a flat red line at a = 1 sec^{-1}. The confidence intervals of one standard deviation in the fits at each t_c are seen as solid colored lines. The fits become more accurate as either the number of points in the data set or t_c grow. In particular, for a fixed number of dwell times in the dataset, a larger time window in the experiment gives more accurate results.

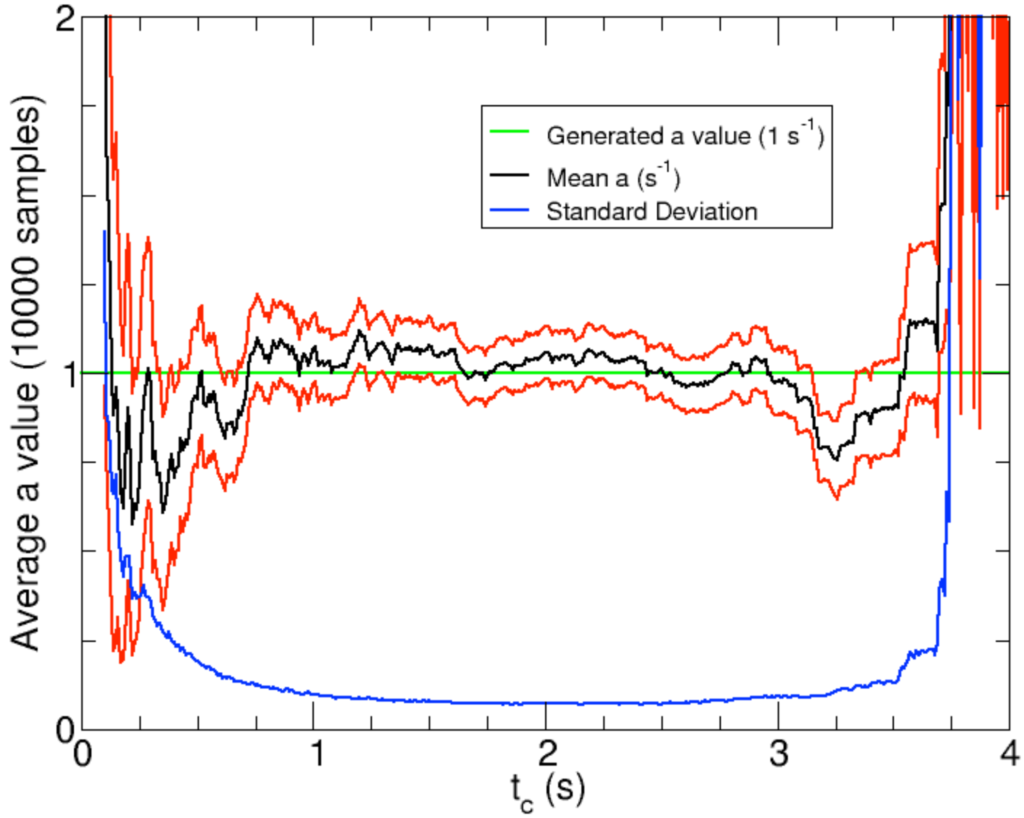

Figure S6. Bayesian sampling. Rate constant found by fitting the empirical CDFs constructed from an exponential dataset with a = 1 sec^{-1} (green line) using different $t_c = t_{max}$ values. The fits were obtained by Bayesian sampling to find the mean a (black curve) and its interval of confidence at one standard deviation (red curves). The absolute value of the standard deviation is shown in blue. At small and large t_c values, the standard deviation is large and the fitting is inaccurate. For t_c between 1 and 3 sec, the standard deviation is small and the fits are accurate. For t_c < 1 sec the range of the empirical CDFs is too short, resulting in poor accuracy; for t_c > 3 sec the number of points kept in the dataset becomes too small. There were 500 dwell times in the original dataset before filtering by t_c in this case.