Manuscript submitted to **Biophysical***Journal*
Article

Autoregressive HMM resolves biomolecular transitions from passive optical tweezer force measurements

Brian A Dawes[1] and Maria Kamenetska[1,2,3,*]

[1]Department of Physics, Boston University, Boston, Massachusetts 02215, USA
[2]Department of Chemistry, Boston University, Boston, Massachusetts 02215, USA
[3]Division of Materials Science and Engineering, Boston University, Boston, Massachusetts 02215, USA
[*]Correspondence: mkamenet@bu.edu

ABSTRACT Optical tweezer (OT) single molecule force spectroscopy is a powerful method to map out the energy landscape of biological complexes and has found increasing applications in academic and pharmaceutical research. The dominant method to extract molecular conformation transitions from the thermal diffusion-broadened trajectories of the microscopic OT probes attached to the single molecule of interest is through hidden Markov models (HMM). In standard applications, the HMMs assume a white noise spectrum of the probes superimposed onto the molecular signal.Here, we demonstrate, through theoretical derivation, computer modeling and experimental measurements that this standard white noise HMM (wnHMM) misses key features of real OT data. The deviation is most pronounced at higher frequencies because the white noise model does not account for the over-damped nature of particle diffusion in an OT harmonic potential in aqueous environments. To address this, we derive how to incorporate autoregression (ar) between consecutive data points into an HMM, and demonstrate through modeling and experiment that such an arHMM captures real OT data behavior across all frequency ranges. Through analysis of real OT data we recorded on a single DNA hairpin undergoing folding and unfolding transitions, we show that the wnHMM extracts lifetimes that are at least a factor of 2 faster and less consistent than the arHMM results which match expectations and prior measurements. Overall, our work suggests that arHMM should be the default model choice for analysis OT single molecule transitions and that its use will improve the fidelity and accuracy of single molecule force spectroscopy measurements.

SIGNIFICANCE This work derives and experimentally validates an improved Bayesian inference method for fitting and interpreting single molecule force spectroscopy data. The impact of this work will be to substantially improve accuracy and reliability of single molecule measurements using optical tweezers, which are becoming wide-spread in biophysical studies of living systems. The method we derive works on data collected using the simplest passive mode of the OT, requiring no feedback, and can serve to make OT measurements and analysis accessible to new users.

INTRODUCTION

Single molecule force spectroscopy using optical tweezers (OT) is a powerful method for uncovering molecular mechanisms at the root of biological phenomena by tracking dynamics of molecular processes in real time. Recognized by the Nobel prize in physics in 2018, the technique allows real-time monitoring of a single biological complex (1, 2). Unlike ensemble measurements which average over a macroscopic number of unsynchronized molecules stochastically evolving in a thermal bath, single molecule measurements with OT track a single complex to yield insights on the molecule's structural conformation and dynamics. In recent years, state-of-the-art OT instruments have been made available commercially and are increasingly relied on by users in academics, industry and pharmaceuticals with no specialized prior training.

A typical OT experiment takes place at the focus of an optical microscope. Two probe particles are tethered together by the molecule of interest using biochemical linkages and suspended in solution. Two trapping lasers focused into the microscope create harmonic potentials that localize each probe away from each other to extend the tether, generating tension (2, 3). Translation of the trapping lasers moves the probes and

extends (contracts) the tether, increasing (decreasing) tension. The force applied by the trap, and thus the tension in the tether, can be measured by observing the deflection of the trapping laser, typically at ≈ kHz rates. Any conformational transition of the biological complex causes a measurable change in the probe positions. By correctly identifying these molecular transitions and the time at which they occur in the probe movement OT signal, researchers can measure the rates at which biologically-relevant transformations take place (4–6). Additionally, the energy landscape of the transition can be extracted from the force dependence of the transition rates (5–7). Using OTs, kinetics and energetics of biological transitions can be mapped out, including in the presence of biologically important perturbations such as binding partners or ligands (7, 8).

However, the task of mapping OT measurements of probe positions to the internal states of the molecule is complicated by the thermal motion of the probe due to Brownian motion. The confining potential of the trap is never sufficiently strong to completely cancel this nanoscale fluctuation of microsphere probe particles in aqueous ambient conditions. Any molecular signal is recorded on top of this noisy background (4). Real-time feedback mechanisms can help filter this noise by maintaining the molecular construct at a constant force, but the bandwidth of the experiment will be fundamentally limited by the feedback bandwidth (9). Implementing a high bandwidth feedback system is non-trivial and sometimes impossible on commercial OTs. Measurements in passive OT mode, where the trapping laser beams are held at a fixed location with no feedback, and monitoring the motion of the probe beads while the molecule undergoes transitions, is a much simpler approaches that simplifies and extends to reach of OT measurements to a greater number of users.

A common approach to analyze passive data and filter out the fluctuations of the bead particles due to Brownian motion is to use unsupervised learning algorithms to help infer the hidden molecular states from the measured probe positions (10–12). If the molecular transitions follow simple chemical kinetics with a single transition state, a well-founded assumption for many biological systems, the time series can be described by a hidden Markov model (HMM). The HMM framework can be used to estimate the properties of the underlying molecular states and assign the state to the bead position measured at each time point (13, 14). HMMs have become a go-to method for analyzing single molecule force measurements and are often assumed to be the gold standard in interpreting OT data (2, 15). However, additional assumptions must be made in an HMM as shown in fig. 1A (left): one needs to specify the emission model which quantifies the probabilistic distribution of x_i given the hidden molecular state s_i (14). The emission model must be chosen to properly represent the statistical properties of the thermal background. Most prior OT HMM analyses have assumed this background is uncorrelated white noise, meaning that subsequent bead positions depend only on the hidden state as indicated (10, 12, 14, 15). We denote this model here as the white noise HMM (wnHMM) model.

The assumptions of the wnHMM, however, are known to deviate significantly from the physical system of micron-sized probes in an OT liquid environment. For example, it is well known that OT data has a Lorentzian power spectral density (PSD), where it is white at low frequencies but transitions to $1/f^2$ Brown noise at higher frequencies. In fact, the frequency at which this transition occurs is a signature of the damped harmonic potential experienced by the probes due to the OT confinement in water and is often used for OT calibration (16, 17). The implications of this non-white noise for analysis of time series data of bead forces or positions has not been thoroughly explored. Significantly, these considerations imply that wnHMM is insufficient to accurately extract molecular transitions occurring at time scales comparable to bead diffusion dynamics.

Here, we derive and experimentally validate an improved HMM model and fitting procedure for extracting molecular transitions from OT single molecule measurements. First, we demonstrate analytically and with experimental and computational data that position time-series as measured in OT experiments deviate significantly from a white noise signal due to bead diffusion in the trap. Specifically, we show that OT data is autocorrelated, contrary to the assumptions of the wnHMM. Next, we integrate autocorrelation analytically into the emission model of an HMM to generate an improved model for OT data interpretation which we term auto regressive HMM (arHMM). Finally, we demonstrate an improved fitting protocol to generate initial parameters for seeding both HMMs and use them to analyze real OT measurements of DNA hairpin transitions we record. By comparing the distribution of molecular state life-times identified with both models with expected Markovian results, we determine that the wnHMM incorrectly classifies a fraction of bead fluctuation events as molecular state transitions. This results in non-exponentially distributed lifetimes, which violates the predictions for a two state system like our hairpin. Overall, the wnHMM-extracted molecular state transition rates are typically a factor of at least 2 faster than the parameters estimated via arHMM and contain significant variability. Analysis of the force dependence of the arHMM transition rates agrees with previous literature which the wnHMM rates fail to replicate without filtering data to remove outliers. Our results suggest that arHMM models should be the default choice of model for assigning OT force transitions in single molecule measurements.

MATERIALS AND METHODS

OT measurement overview

A schematic of an OT experiment in a Lumicks C-Trap, used to examine here the folding dynamics of a DNA hairpin, is shown in fig. 1B. This dual-beam setup containing two optical traps is a common geometry in the field (2). The molecule of

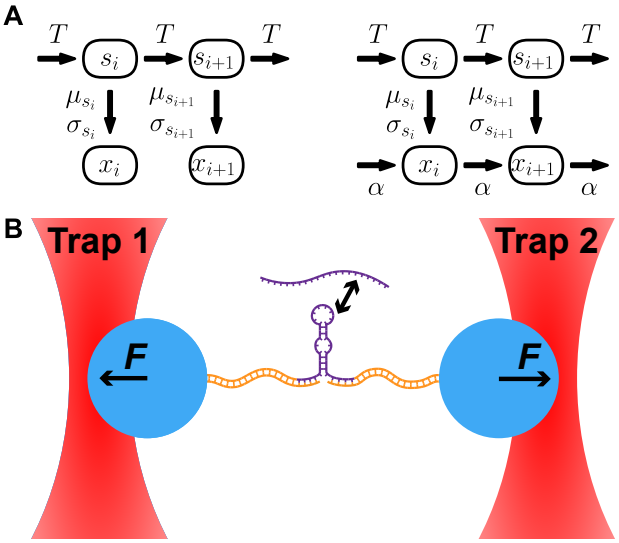

Figure 1: Diagrams of the HMM schema and a typical dual-trap OT experiment. *(A)* A standard wnHMM (left) consists of a sequence of hidden states s_i and a sequence of observed emissions x_i. The likelihood of observing x_i is a Gaussian with mean and standard deviation μ_{s_i}, σ_{s_i} which depend on the state. An arHMM (right) has a similar structure but the emission likelihood also depends on the previous emission via an autoregressive coefficient α. *(B)* Two optically trapped beads are connected via a single molecule tether and the trapping force F is recorded. The tether depicted has two dsDNA handles (orange) and a region of interest (purple) with a folded structure. Unfolding of the region of interest increases the tether contour length and decreases F.

interest (purple) is tethered between two microscopic glass or plastic beads often using double-stranded DNA (dsDNA) handles (orange). The beads are trapped in aqueous solution using a focused laser beam which creates a harmonic potential which localizes the bead at the focus of an optical microcsope. The trapping force on the bead F is linearly related to the displacement of the bead from the trap center x by a Hookean stiffness κ that depends on experimental conditions. The deflection of the trapping laser by the bead is measured using back-focal plane interferometry at kHz rates and is experimentally calibrated against Brownian fluctuations into either force F, or equivalently position x (16). By manipulating the position of the lasers, a force is applied to the beads and they can be tweezed–moved apart–to exert tension on the molecule.

A common and easily accessible approach to measuring single molecule dynamics is to simply watch the real-time fluctuations of the molecular conformation while keeping the trap fixed in passive mode with no feedback applied (18, 19). In this technique, the two beads are positioned a certain distance apart such that some tension is applied to the tethered molecule. Under this force, the molecule undergoes transitions which change the overall contour length of the construct and appear as jumps in bead displacement or force. This technique has the advantage of not requiring feedback to maintain a constant force and is well-suited for studying molecular dynamics which occur faster than the ≈10 ms time scale of most feedback systems. It is also simple to implement by non-experts and requires the least technical expertise and no extra instrumentation beyond a standard OT. However, HMM methods or other analysis tools to locate molecular transitions in thermally broadened OT data are required.

For details about sample preparation, experimental passive mode expeirmental protocols and data analysis procedures, please see the SI.

RESULTS AND DISCUSSION

OT data is described by an autoregressive model

We first focus on only one trap with a single probe particle, for example trap 1 in fig. 1B. In this situation, no molecular tether is attached to the bead and no net tension is exerted that can pull it out of the trap. A trapped probe particle experiences stochastic Brownian fluctuations, viscous drag, and a linear restoring force from the trap. The particle's position x(t) obeys a harmonic Langevin equation (16):

$$m\dot{x}(t) = \sqrt{2D\gamma}\eta(t) - \gamma\dot{x} - \kappa[x(t) - x_{\mathrm{eq}}] \qquad (1)$$

where γ is the drag coefficient of the solvent, η is an uncorrelated white noise term with 0 mean and unit variance, D is the diffusion coefficient, κ is the trap stiffness, and x_{eq} is the equilibrium position.

As the probe is overdamped in typical OT experiments,

this can be simplified to the Ornstein-Uhlenbeck form by dropping the inertial term (16, 20):

$$\dot{x}(t) = \sqrt{2D}\eta(t) - \frac{\kappa}{\gamma}[x(t) - x_{\text{eq}}] \tag{2}$$

If the continuous signal $x(t)$ is sampled at rate f_s, the discrete signal is described by the discretized version of eq. (2):

$$x_i = x_{i-1} + (1-\alpha)(x_{\text{eq}} - x_{i-1}) + (1-\alpha^2)\sigma\eta_i \tag{3}$$

$$\alpha = \exp\left(\frac{-\kappa}{\gamma f_s}\right); \quad \sigma^2 = \frac{\gamma D}{\kappa}$$

where η_i are independent normally distributed values with 0 mean and unit variance. The second term in eq. (3) is the motion towards the trap center which is controlled by α. The final term represents the stochastic fluctuations and is scaled so that $\text{Var}(x_i) = \sigma^2$. In statistics, this model is known as AR(1), an autoregressive model of order 1, as each value depends explicitly on 1 previous value. Each data point has a correlation of α with the previous data point, which gives rise to an exponentially decaying autocorrelation function (ACF):

$$R_{xx}(\tau) = \sigma^2\alpha^{-\tau f_s} = \sigma^2\exp(-\tau\kappa/\gamma) \tag{4}$$

Sample position measurements of a single bead localized at the focus of one optical trap potential with no tethered molecules attached, as described above, is shown in fig. 2A (top). The ACF of this time series is plotted in black in fig. 2B. We observe an exponential decay of the ACF in agreement with eq. (4).

We fit the time series to the AR(1) model via analytical MLE to extract $x_{\text{eq}}, \sigma^2, \alpha$. Feeding these into eq. (3) above, we generate artificial AR(1) data to compare to our experimental result. The resulting time-series is shown in blue in fig. 2A (middle) and is qualitatively similar to the real data plotted in black. Importantly, the calculated ACF of the AR(1) model is in good agreement with experiment.

Interestingly, over a time span much greater than the autocorrelation time, the histogram of OT data is Gaussian as seen in the histogram in fig. 2A (top). These Gaussian histograms can be generated as well by a simpler white noise process:

$$x_i = x_{\text{eq}} + \sigma\eta_i \tag{5}$$

However, this model does not accurately replicate the original data, which we show by generating a white noise time-series using eq. (5), plotted in green in fig. 2A (bottom). As expected, the histograms of forces for the real, AR(1) and white noise data are indistinguishable, characterized by the same mean and width σ. However, by inspection of the time-series, we observe that the dynamics of bead motion predicted by the white noise at short time-scales are distinct from the real and AR(1) time-series, indicating that autocorrelations are important for accurately capturing the dynamics of the diffusive motion of the bead in an OT experiment.

The failure of the white noise model at high frequencies is especially clear when the three time-series are plotted in the frequency domain in fig. 2C. While the power spectral density (PSD) of white noise is flat (orange), it is well known that OT data has a Lorentzian PSD (16):

$$S_{xx}(f) = \frac{D}{2\pi\left(f^2 + f_c^2\right)}; \quad f_c = \frac{2\pi\kappa}{\gamma} \tag{6}$$

The PSD of our real OT data and of the AR(1), shown in black and blue respectively, are well described by eq. (6). Below the corner frequency f_c we observe that the PSD is approximately white but above f_c the PSD falls off as $1/f^2$. In fact, we can derive that f_c and α are interrelated:

$$\alpha = \exp\left(-\frac{2\pi f_s}{f_c}\right) \tag{7}$$

where f_s is the data sampling rate. Overall, the white noise underestimates (overestimates) the power at low (high) frequencies, resulting in the faster fluctuations seen in the time series.

A critical difference between models is that the real and AR(1) data contain many extended excursions to regions far from the mean, as observed in the time series in fig. 2A (top and middle). These fluctuations are significant, as they may appear as transitions to a different hidden state of the molecule in a dual-beam OT measurement shown in fig. 1B. To quantify this difference between the white noise and the real or AR(1) data, we plot in fig. 2D the frequency of bead excursions beyond 2 standard deviations away from the mean for a given number of data points. While the chance of seeing a run of several outliers is vanishingly small for the white noise data, both real and AR(1) data contain many runs of several outliers, reflecting the finite speed of bead diffusion back towards the mean. These outlier runs can be particularly troublesome for an HMM which may falsely infer that the molecule is transitioning rapidly between two states.

arHMM accurately extracts molecular transitions in dual-beam OT data

We now turn to analysis of a typical dual-beam OT experiment, shown in fig. 1B, which uses two trapped probes connected by a tether containing a central molecule of interest which has a folded and unfolded state. The average force on the probes is sampled at a rate f_s, yielding a sequence of experimental data x_i as in the single probe case. The state of the molecule at each timepoint s_i exists but is not known by the experimentalist and must be inferred from the experimental signal, i.e., s_i is a hidden parameter. If there is only a single transition path between the folded and unfolded states, which is common in biomolecules studied by OT, the transitions obey first-order chemical kinetics or equivalently Markovian dynamics and can be described by a transition matrix T (2). This framework containing a hidden state s_i (the conformation of

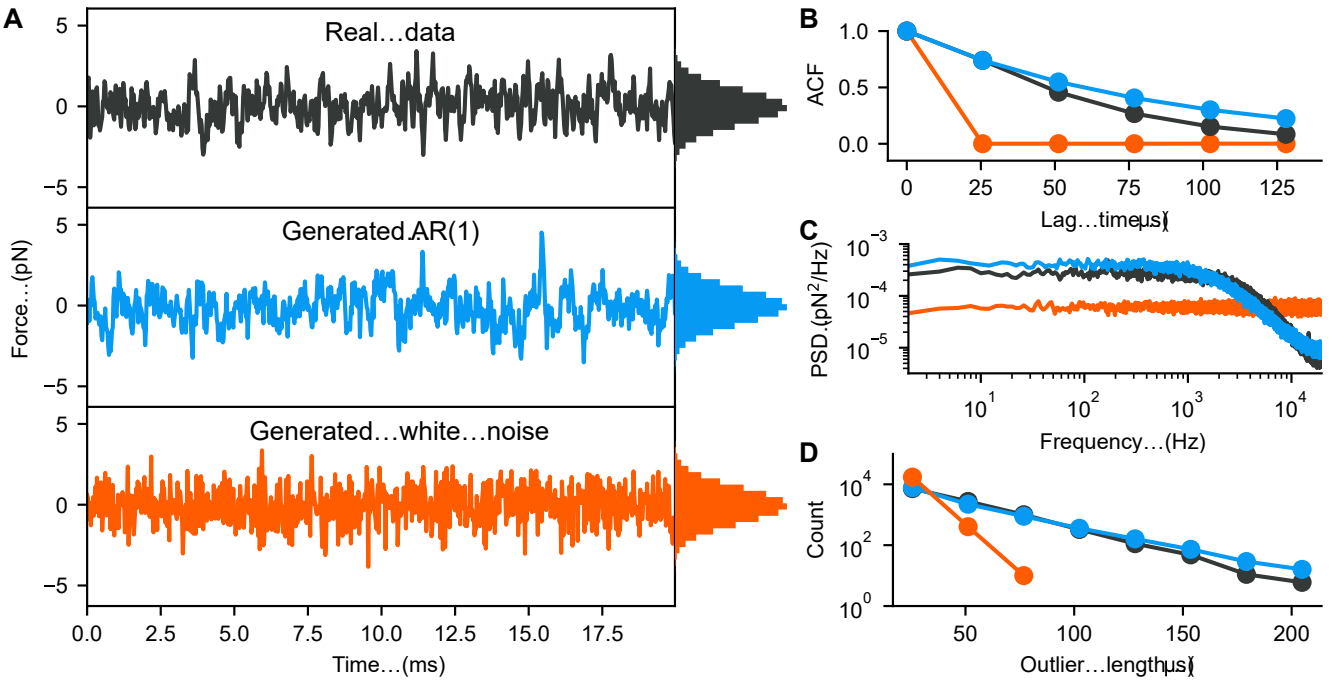

Figure 2: Comparison of optical tweezer force data for a single bead (black) with the AR(1) (blue) and white noise (orange) models. AR(1) and white noise data were generated using parameters obtained from fitting the real data with MLE. Each dataset is 20 s long and sampled at 39.0625 kHz. *(A)* Representative 15 ms time traces from the three datasets. The real and AR(1) data are much smoother than the white noise data. *(B)* Normalized autocorrelation functions of the three datasets. The white noise model displays no autocorrelation while the real and AR(1) data show an exponential decay. *(C)* The power spectral densities of the three datasets. The real and AR(1) data display similar Lorentzian spectra unlike the flat white noise spectrum. *(D)* Histogram of observed runs of data points above 2 standard deviations from the mean by run length. Only runs of the exact length are counted, i.e., subsequences of longer runs are not counted. Both the real and AR(1) data display extended runs in the outlier region that are not present in the white noise data.

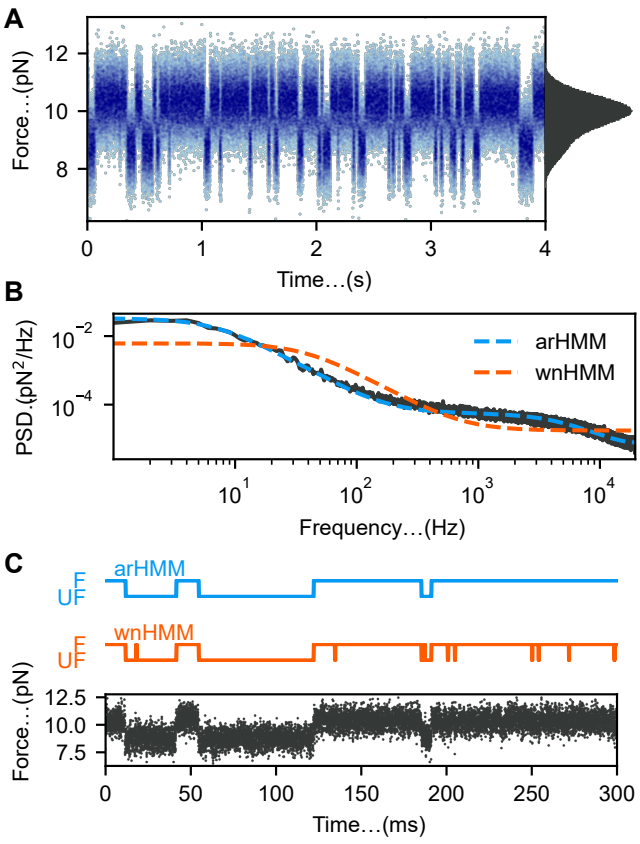

Figure 3: Results of fitting 21 s of molecular transitions with arHMM and wnHMM. *(A)* A 4 s subset of the measured force vs time. Data is sampled at 39.0625 kHz and colored by density to resolve overplotting. The histogram on the right is generated from the full 21 s. *(B)* 300 ms of data from the above time series. Both arHMM and wnHMM were used to predict the most likely state at each timepoint, folded (F) or unfolded (UF). The wnHMM erroneously predicts two short unfolding and refolding events. *(C)* In black, the PSD of the time series from A. Parameters from the HMM fitting were used to generate PSDs for the arHMM (blue) and wnHMM (orange) models. The arHMM properly models the high frequency falloff due to the OT.

the molecule) evolving based on a transition matrix T (the chemical rate constants) and a signal dependent on the state x_i (the experimentally determined probe displacement/force signal) is known as a hidden Markov model (HMM). An HMM can be fit to learn the transition matrix T, providing thermodynamic information about the molecular transitions (11, 12). Additionally, an HMM can also be smoothed to estimate the molecular state at each timepoint to pinpoint when molecular transitions occur.

The evolution of the observed data based off the molecular state is encoded by the emission model of the HMM which typically assumes that the observed values are independent of each other and normally distributed based on the molecular state:

$$x_i = x_{\mathrm{eq},s_i} + \sigma_{s_i}\eta_i \tag{8}$$

where x_{eq,s_i} and σ_{s_i} are the equilibrium position and standard deviation of the probe when the molecule is in a state s_i and η_i is an independent random value drawn from the standard normal distribution. This is analogous to the single-probe white noise model in eq. (5) but includes the dependence of bead positions x_i on the hidden state as shown in Figure 1A (left). Thus we term it the white noise HMM (wnHMM). Figure 1

However, as demonstrated above, the white noise model does not accurately capture background fluctuations in OT data. Instead, we propose using an emission model based on the AR(1) model of eq. (3) which properly incorporates the autocorrelation of the measurements:

$$x_i = x_{i-1} + (1-\alpha)(x_{\mathrm{eq},s_i} - x_{i-1}) + (1-\alpha^2)\sigma_{s_i}\eta_i \tag{9}$$

We call this the auto-regressive HMM (arHMM) (21, 22). Figure 1A (right) shows a schematic for the arHMM. To summarize, the arHMM has a modified emission probability which takes into account the correlation of displacement between time points in addition to the hidden state s_i. As a result, the likelihood of assigning a molecular state transition at any time point depends not only on the present molecular state, but also on the previous bead displacement. /stThis parameter allows the model to account for excursions of the beads far from equilibrium due to stochastic diffusion which commonly occur in OT data, as we have previously shown.

We test both models on real force data collected on a hairpin construct shown in fig. S1 in a dual-beam OT setup, as shown in fig. 3A. In this case, a DNA hairpin is tethered between two beads and the trap positions are fixed as described in the methods. The higher and lower force states centered around 12 and 10 pN correspond to the hairpin in a folded or unfolded state respectively. The probability density of forces, plotted in the right inset, is a mixture of two Gaussians. A fit of these Gaussians is shown in fig. S3A.

The PSD of this signal, shown in fig. 3B, displays a double-Lorentzian distribution, in contrast to the spectrum of a single diffusing bead in the trap in fig. 2C (black). Fitting the double Lorentzian to the data (shown in fig. S3B), we

extract two characteristic time scales. At higher frequency, we obtain the f_c of the optical trap as discussed previously and in eq. (6). As shown in eq. (7), this parameter is equivalent to the autoregressive parameter α which is only accounted for by the arHMM.

Turning to the lower frequency component, we observe that it is absent from the single-probe data in fig. 2 as it originates from the Markov switching of the molecule as derived in the supplemental. The corner frequency of this molecular process f_m can be written as:

$$f_m = \frac{k_1 + k_2}{2\pi} \tag{10}$$

where k_1, k_2 are the forward and backward transitions from hidden states s_1 and s_2. In our case, these rates correspond to the folding and unfolding rates of the hairpin k_F, k_{UF}. This signal is present in both the wnHMM and arHMM; however, only the arHMM generates the second Lorentzian peak.

To fit both models to the data we use standard iterative methods to find the MLE parameters that maximize the likelihood.(12, 21, 22) Since a a poor choice of initial parameter guesses can cause the fitting procedure to converge to an inaccurate estimate, we developed a novel methodology to help automate this procedure and generate more robust initializations for both the wnHMM and arHMM. The details are provided in the supplement. In brief, we use the parameters extracted from the double Gaussian fit to the force histogram in fig. S3A and the estimate of f_m obtained from fitting the PSD in fig S3B to provide initial guess to both models. The α parameter for the arHMM is also provided from the higher frequency Lorentzian PSD fit.

Tables S2 and S3 show the resulting transition rates and Bayesian information criteria (BIC) for all 8 datasets for the arHMM and wnHMM. Table S4 shows a comparison between the arHMM and wnHMM fits. In each dataset, the BIC for the arHMM fit is lower, indicating that the arHMM fits the data better and that this is not a result of overfitting by introducing extra parameters. In fig. 3B, dashed lines, we also observe that the parameters output by the arHMM fit replicate the measured PSD while the wnHMM parameters cannot replicate the higher frequency falloff. To compensate, the predicted f_m of the wnHMM is shifted to higher frequencies, indicating that the wnHMM is overestimating the transition rates. This is significant as estimating the transition rates is often the end goal of single molecule biophysics measurements.

We find that this overestimation of transition rates also results in misclassification of molecular states.The resulting assignment of a small subsection of the data is plotted for arHMM and wnHMM in blue and orange respectively in fig. 3C. Both models identified the distinct probe displacements in the real data, plotted in black, as corresponding to two molecular states at ≈10 and ≈12 pN. However, the wnHMM was significantly more likely to assign fast transitions when probes strayed far from either equilibrium. In contrast, the arHMM assigned these to rare bead diffusion events rather than to a change of the hidden state.

The distribution of folded and unfolded state lifetimes as extracted by the wnHMM and arHMM models are plotted in fig. 4A in orange and blue, respectively. As expected, the wnHMM shows much shorter lifetimes. For a two-state system, the survival plots should show an exponential decay (plotted in grey and black dashes) with rate constants set by the molecular unfolding and folding rates, k_F and k_{UF}. The lifetimes measured by the arHMM model follow this prediction. In contrast, the assignment by the wnHMM model results in very clear and drastic deviations from single exponential behavior, with a large excess of very short lifetimes. This deviation is particularly evident for the less stable state, which corresponds to the unfolded state in this particular example. We conclude that these fast events are due to erroneous assignment of long-lived bead fluctuations to state transitions, such as the one plotted in fig. 3C.

We perform similar measurements at a range of forces and repeat the analysis to measure transition rates as a function of force. The resulting rates for folding and unfolding as estimated by the wnHMM and arHMM are plotted in fig. 4B in orange and blue respectively. According to Arrhenius kinetics, the transition rates should follow (23):

$$k(F) = k_0 \exp\left(\pm \frac{F\Delta x^{\ddagger}}{k_B T}\right) \tag{11}$$

where k_0 is the zero force rate and $\Delta x^{\ddagger}$ is the distance to the transition state. Dashed lines in fig. 4B show the exponential fits to the rates produced by the arHMM. In contrast, the wnHMM-derived rates have a wider distribution, with several outliers from the predicted exponential relationship. The extracted wnHMM rates are consistently faster than the arHMM as already discussed.

Fitting eq. (11) to the arHMM rates, the transition state is estimated to be 3.9 nm from the unfolded state which can be identified with the GC basepair just before the hairpin loop in fig. S1. This is consistent with previous measurements of hairpin unfolding dynamics which show the transition state often occurs near the hairpin loop (24). Removing the outliers from the wnHMM-determined rates in fig. 4B, we extract a transition state to be 6.6 nm from the unfolded state, much further from the loop.

Additionally, we can calculate the Gibb's free energy difference between the two states at 0 force as (23):

$$\Delta G = F_{1/2}\Delta x_{\text{tot}} \tag{12}$$

where $F_{1/2}$ is the force at which $k_F = k_{UF}$ and Δx_{tot} is the total length change between the states. Using 0.44 nm/nt and accounting for the 2 nm width of the helix, we estimate $\Delta x_{\text{tot}} = 18.2$ nm. For the arHMM, $F_{1/2} = 10.5$ pN and $\Delta G = 46.6\ k_B T$. For the wnHMM, $F_{1/2} = 10.1$ pN and $\Delta G = 44.8\ k_B T$. In comparison, the MFOLD webserver predicts $\Delta G = 44\ k_B T$ for our sequence and experimental

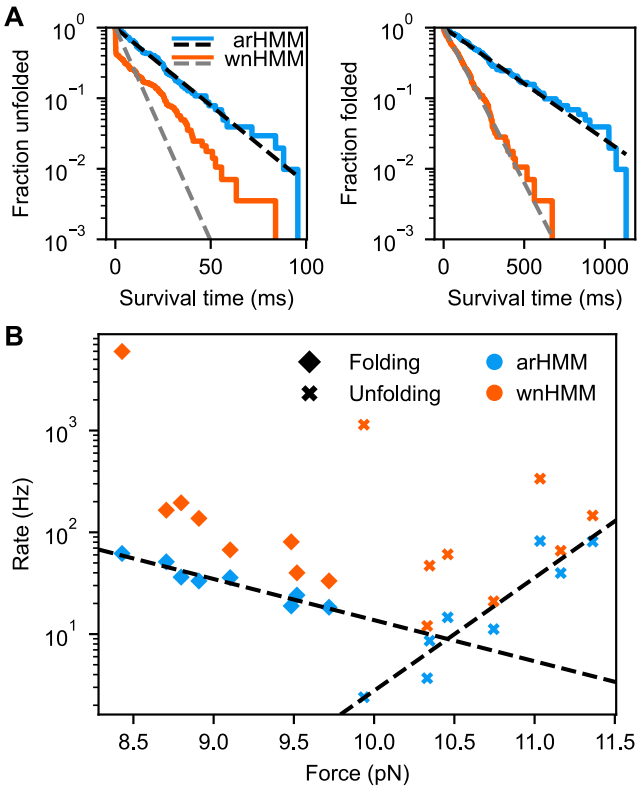

Figure 4: Folding and unfolding rates predicted by arHMM and wnHMM. *(A)* Survival fraction of the unfolded (left) and folded (right) states, resulting from the folding and unfolding rates respectively, as predicted by arHMM and wnHMM. Rate constants were extracted from fitting the HMMs, from which an exponential decay can be predicted (dashed lines). The HMMs were then smoothed to predict the most likely state at each timepoint, from which a histogram of state lifetimes can be obtained (solid lines). The arHMM predicts longer survival times and thus slower folding and unfolding rates. *(B)* Folding (diamonds) and unfolding (xs) rates obtained by holding at different trap distances. Rates are obtained by fitting the arHMM (blue) or wnHMM (orange). Data are from two replicates which were each held at 4 different distances. Dashed lines indicate an exponential fit to the arHMM rates. wnHMM consistently predicts faster transition rates.

conditions (25). While these wnHMM and arHMM results are comparable for ΔG, this interpretation of the wnHMM was only possible after removal of the outliers in fig. 4B. Overall, these measurements demonstrate the enhanced performance of the arHMM in extracting realistic molecular parameters, including the transition state and free-energy state difference, from OT experimental data.

CONCLUSION

In conclusion, we show that wnHMM models, often used for OT data analysis, consistently overestimate molecular transition rates in single molecule OT measurements. We derive analytically how to modify wnHMM models to incorporate the autocorrelation inherent to OT data series which we demonstrate here. The resulting arHMM model captures critical features of real data which wnHMM does not. Importantly, we show that an arHMM produces accurate and consistent measurements of molecular reaction rates when applied to real OT time-series data of a known DNA hairpin construct. Our results suggest that arHMM is a significant improvement over standard wnHMM approaches and should be the default analysis technique of OT data series for single molecule kinetic rate measurements. The method we develop is freely available as an annotated Python code on Github. We emphasize that our work here not only improves the reliability of molecular parameters extracted from single molecule force spectroscopy measurements, but also extends the reach of these methods to new users who can now use the simple passive mode technique in combination with arHMM analysis code to perform state-of-the-art measurements.

AUTHOR CONTRIBUTIONS

B.D. and M.K conceived the project. B.D. conducted the experiments and analyzed data. B.D. and M.K. wrote the article.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under award #2117585 and by Arrakis Therapeutics.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Nobel Prize Outreach AB. *The Nobel Prize in Physics 2018*. URL: https://www.nobelprize.org/prizes/physics/2018/summary/.
2. Carlos J. Bustamante et al. “Optical tweezers in single-molecule biophysics”. In: *Nature Reviews Methods Primers 2021 1:1* 1.1 (2021), pp. 1–29. ISSN: 2662-8449. DOI:

10.1038/s43586-021-00021-6. URL: https://www.nature.com/articles/s43586-021-00021-6.

3. A Ashkin et al. "Observation of a Single-Beam Gradient Force Optical Trap for Dielectric Particles". English. In: *Optics Letters* 11.5 (1986), pp. 288–290. ISSN: 01469592.
4. Karel Svoboda et al. "Direct observation of kinesin stepping by optical trapping interferometry". In: *Nature* 365.6448 (1993), pp. 721–727. ISSN: 1476-4687. DOI: 10.1038/365721a0. URL: https://doi.org/10.1038/365721a0.
5. Krishna Neupane et al. "Direct observation of transition paths during the folding of proteins and nucleic acids". In: *Science* 352.6282 (2016), pp. 239–242. ISSN: 10959203. DOI: 10.1126/science.aad0637.
6. J. Liphardt et al. "Reversible unfolding of single RNA molecules by mechanical force". In: *Science* 292.5517 (2001), pp. 733–737. ISSN: 00368075. DOI: 10.1126/science.1058498. URL: https://www.science.org/doi/10.1126/science.1058498.
7. Andrew H. Mack et al. "Kinetics and thermodynamics of phenotype: Unwinding and rewinding the nucleosome". In: *Journal of Molecular Biology* 423.5 (2012), pp. 687–701. ISSN: 10898638. DOI: 10.1016/j.jmb.2012.08.021.
8. Vishnu Chandra et al. "Single-molecule analysis reveals multi-state folding of a guanine riboswitch". In: *Nature Chemical Biology 2016 13:2* 13.2 (2016), pp. 194–201. ISSN: 1552-4469. DOI: 10.1038/nchembio.2252. URL: https://www.nature.com/articles/nchembio.2252.
9. Matthew J. Lang et al. "An Automated Two-Dimensional Optical Force Clamp for Single Molecule Studies". In: *Biophysical Journal* 83.1 (2002), pp. 491–501. ISSN: 0006-3495. DOI: https://doi.org/10.1016/S0006-3495(02)75185-0. URL: https://www.sciencedirect.com/science/article/pii/S0006349502751850.
10. Lorin S. Milescu et al. "Extracting Dwell Time Sequences from Processive Molecular Motor Data". In: *Biophysical Journal* 91.9 (2006), pp. 3135–3150. ISSN: 0006-3495. DOI: https://doi.org/10.1529/biophysj.105.079517. URL: https://www.sciencedirect.com/science/article/pii/S0006349506720276.
11. Ying Gao, George Sirinakis, and Yongli Zhang. "Highly Anisotropic Stability and Folding Kinetics of a Single Coiled Coil Protein under Mechanical Tension". In: *Journal of the American Chemical Society* 133.32 (2011). PMID: 21707065, pp. 12749–12757. DOI: 10.1021/ja204005r. eprint: https://doi.org/10.1021/ja204005r. URL: https://doi.org/10.1021/ja204005r.
12. Johannes Stigler and Matthias Rief. "Hidden Markov Analysis of Trajectories in Single-Molecule Experiments and the Effects of Missed Events". In: *ChemPhysChem* 13.4 (2012), pp. 1079–1086. DOI: https://doi.org/10.1002/cphc.201100814. eprint: https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/cphc.201100814. URL: https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/cphc.201100814.
13. Sheyum Syed et al. "Improved Hidden Markov Models for Molecular Motors, Part 2: Extensions and Application to Experimental Data". In: *Biophysical Journal* 99.11 (2010), pp. 3696–3703. ISSN: 0006-3495. DOI: https://doi.org/10.1016/j.bpj.2010.09.066. URL: https://www.sciencedirect.com/science/article/pii/S0006349510012506.
14. John F Beausang and Philip C Nelson. "Diffusive hidden Markov model characterization of DNA looping dynamics in tethered particle experiments". In: *Physical Biology* 4.3 (2007), p. 205. DOI: 10.1088/1478-3975/4/3/007. URL: https://dx.doi.org/10.1088/1478-3975/4/3/007.
15. Maurizio Righini et al. "Full molecular trajectories of RNA polymerase at single base-pair resolution". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.6 (2018), pp. 1286–1291. ISSN: 10916490. DOI: 10.1073/PNAS.1719906115/-/DCSUPPLEMENTAL.
16. Kirstine Berg-Sørensen and Henrik Flyvbjerg. "Power spectrum analysis for optical tweezers". In: *Rev. Sci. Instrum* 75 (2004), pp. 594–612. DOI: 10.1063/1.1645654. URL: https://doi.org/10.1063/1.1645654.
17. Zi-Qiang Wang et al. "Calibration of optical tweezers based on an autoregressive model". In: *Opt. Express* 22.14 (2014), pp. 16956–16964. DOI: 10.1364/OE.22.016956. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-22-14-16956.
18. William J. Greenleaf et al. "Passive all-optical force clamp for high-resolution laser trapping". In: *Physical Review Letters* 95.20 (2005), p. 208102. ISSN: 00319007. DOI: 10.1103/PHYSREVLETT.95.208102/FIGURES/3/MEDIUM. URL: https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.95.208102.
19. Lorenz Rognoni et al. "Force-dependent isomerization kinetics of a highly conserved proline switch modulates the mechanosensing region of filamin". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.15 (2014), pp. 5568–5573. ISSN: 10916490. DOI: 10.1073/PNAS.1319448111/-/DCSUPPLEMENTAL. URL: /pmc/articles/PMC3992639/%20/pmc/articles/PMC3992639/?report=abstract%

20https://www-ncbi-nlm-nih-gov.ezproxy.bu.edu/pmc/articles/PMC3992639/.

20. G. E. Uhlenbeck and L. S. Ornstein. “On the Theory of the Brownian Motion”. In: *Phys. Rev.* 36 (5 1930), pp. 823–841. DOI: 10.1103/PhysRev.36.823. URL: https://link.aps.org/doi/10.1103/PhysRev.36.823.

21. James D Hamilton. “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle”. In: *Econometrica* 57 (2 1989), pp. 357–384. ISSN: 00129682, 14680262. DOI: 10.2307/1912559. URL: http://www.jstor.org/stable/1912559.

22. Chang Jin Kim. “Dynamic linear models with Markov-switching”. In: *Journal of Econometrics* 60 (1-2 1994), pp. 1–22. ISSN: 0304-4076. DOI: 10.1016/0304-4076(94)90036-1.

23. Ignacio Tinoco Jr and Carlos Bustamante. “The effect of force on thermodynamics and kinetics of single molecule reactions”. In: *Biophysical Chemistry* 101-102 (2002). Special issue in honour of John A Schellman, pp. 513–533. ISSN: 0301-4622. DOI: https://doi.org/10.1016/S0301-4622(02)00177-1. URL: https://www.sciencedirect.com/science/article/pii/S0301462202001771.

24. Michael T. Woodside et al. “Direct Measurement of the Full, Sequence-Dependent Folding Landscape of a Nucleic Acid”. In: *Science* 314.5801 (2006), pp. 1001–1004. DOI: 10.1126/science.1133601. eprint: https://www.science.org/doi/pdf/10.1126/science.1133601. URL: https://www.science.org/doi/abs/10.1126/science.1133601.

25. Michael Zuker. “Mfold web server for nucleic acid folding and hybridization prediction”. In: *Nucleic Acids Research* 31.13 (2003), pp. 3406–3415. ISSN: 0305-1048. DOI: 10.1093/nar/gkg595. eprint: https://academic.oup.com/nar/article-pdf/31/13/3406/9487491/gkg595.pdf. URL: https://doi.org/10.1093/nar/gkg595.