Bayesian hidden Markov model analysis of single-molecule force spectroscopy: Characterizing kinetics under measurement uncertainty

John D. Chodera,[1,*] Bettina Keller,[2,†] Phillip J. Elms,[3,4,5,‡] Frank Noé,[2,§] Christian M. Kaiser,[6,¶] Aaron Ewall-Wice,[7] Susan Marqusee,[3,8,5] Carlos Bustamante,[3,8,5,9,10,11] and Nina Singhal Hinrichs[12,**]

[1]*Computational Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA*
[2]*DFG Research Center Matheon, FU Berlin, Arnimallee 6, 14195 Berlin, Germany*
[3]*California Institute of Quantitative Biosciences (QB3), University of California, Berkeley, CA 94720, USA*
[4]*Biophysics Graduate Group, University of California, Berkeley, CA 94720, USA*
[5]*Jason L. Choy Laboratory of Single Molecule Biophysics, Institute for Quantitative Biosciences, University of California, Berkeley, CA 94720, USA*
[6]*Department of Biology, Krieger School of Arts and Sciences, Johns Hopkins University, Baltimore, MD 21218*
[7]*University of Chicago, IL 60637, USA*
[8]*Department of Molecular & Cell Biology, University of California, Berkeley, CA 94720, USA*
[9]*Department of Physics, University of California, Berkeley, CA 94720, USA*
[10]*Department of Chemistry, University of California, Berkeley, CA 94720, USA*
[11]*Howard Hughes Medical Institute, University of California, Berkeley, CA 94720, USA*
[12]*Departments of Statistics and Computer Science, University of Chicago, IL 60637, USA*

(Dated: March 12, 2015)

Single-molecule force spectroscopy has proven to be a powerful tool for studying the kinetic behavior of biomolecules. Through application of an external force, conformational states with small or transient populations can be stabilized, allowing them to be characterized and the statistics of individual trajectories studied to provide insight into biomolecular folding and function. Because the observed quantity (force or extension) is not necessarily an ideal reaction coordinate, individual observations cannot be uniquely associated with kinetically distinct conformations. While maximum-likelihood schemes such as hidden Markov models have solved this problem for other classes of single-molecule experiments by using temporal information to aid in the inference of a sequence of distinct conformational states, these methods do not give a clear picture of how precisely the model parameters are determined by the data due to instrument noise and finite-sample statistics, both significant problems in force spectroscopy. We solve this problem through a Bayesian extension that allows the experimental uncertainties to be directly quantified, and permits uncertainty to be further reduced through the imposition of physical priors (such as the requirement that dynamics obey detailed balance). We illustrate the utility of this approach in characterizing the observed multistate kinetics of two molecules in a stationary optical trap: an RNA hairpin that exhibits three-state behavior and a protein system (apomyglobin at pH 5) that exhibits two-state behavior with highly overlapping force distributions characterizing each state.

Keywords: force spectroscopy; optical tweezers; single-molecule experiments; hidden Markov model (HMM); Bayesian analysis; statistical error; statistical uncertainty

I. INTRODUCTION

Recent advances in biophysical measurement have led to an unprecedented ability to monitor the dynamics of single biological macromolecules, such as proteins and nucleic acids [**?**]. Advances in instrumentation for optical force spectroscopy in particular have produced instruments of extraordinary stability, precision, and temporal resolution [**?** **?**] that have already demonstrated great utility in the study of biomolecules in the presence of externally perturbative forces [**?** **?** **?**]. Under external force, it becomes possible to stabilize and characterize short-lived conformational states, such as protein folding and unfolding intermediates [**?** **?** **?**].

In a typical single-molecule optical trapping experiment, a protein or nucleic acid is tethered to two polystyrene beads by dsDNA handles that serve as spacers to prevent the molecule under study from interacting with the beads (as in Figure 1). The handle-biomolecule-handle assembly—referred to as a fiber—is associated with the beads through tight noncovalent interactions, with one bead held in an optical trap and the other either suctioned to a micropipette or held in a second optical trap. During an experiment, the position of the bead within the laser trap is monitored, and either the relative displacement from the trap center or the total force on the bead is recorded, resulting in a timeseries such as the one depicted in Fig. 4. The instrument can generally be operated in several modes: a force ramp mode, in which the trap is translated rapidly enough to potentially carry the system transiently out of equilibrium; a passive mode, in which the trap is held fixed and the equilibrium fluctuations of the system are observed; and a constant force-feedback mode, in which the trap is continually repositioned to maintain a specified average constant force on the fiber. Here, we concern ourselves with passive-mode force probe experiments, though both nonequilibrium experiments and constant force-feedback experiments and their analysis remain an exciting topic of active research [**?** **?**] FN: add additional citations

* Corresponding author; john.chodera@choderalab.org
† bettina.keller@fu-berlin.de
‡ elms@berkeley.edu
§ frank.noe@fu-berlin.de
¶ kaiser.jhu.bio@gmail.com
** nshinrichs@uchicago.edu

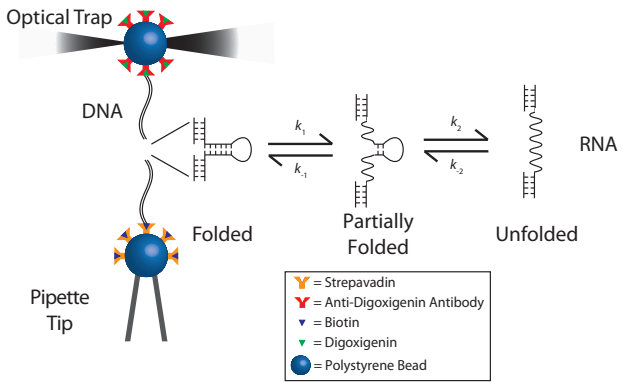

FIG. 1. Single-molecule optical trapping configuration. The biomolecule of interest—here, an RNA hairpin—is tethered to two polystyrene beads by dsDNA handles. The fluctuating force on one bead held in an optical trap is monitored, while the other bead is held suctioned to a micropipette tip. Conformational transitions of the hairpin—such as transitions among the three kinetically metastable states illustrated here—are observed indirectly through motion of the bead in the trap.

for both modes / analyses.

Often, the dynamics observed in these experiments appears to be dominated by stochastic transitions between two or more strongly metastable conformational states [**? ?**]—regions of conformation space in which the system remains for long times before making a transition to another conformational state. The observed force trace in Fig. 4, for example, demonstrates apparent hopping behavior between two or more such states. Transitions among strongly metastable states such as these are generally well-described by first-order kinetics [**?**]. While visual inspection of the dynamics may suggest the clear presence of multiple metastable states, quantitative characterization of these states is often difficult. First, the observed force or extension is unlikely to correspond to a true reaction coordinate easily able to separate all metastable states [**? ? ? ?**], and second, measurement noise may further broaden the force or extension signatures of individual states, increasing their overlap. As a result, a histogram of the observed temporal trace can have strongly overlapping state distributions, as in Fig. 4, right. Attempting to separate these states by simply dividing the observed force or extension range into regions, following current practice [**? ?**], can often lead to a high degree of state mis-assignment that results in the estimated rate constants and state distributions containing a significant amount of error [**? ?**] (see *Supplementary Material: Comparison with threshold model*).

Hidden Markov models (HMMs) [**?**], which use temporal information in addition to the instantaneous value of the observable (force or extension) to determine which conformational states the system has visited during the experiment, have provided an effective solution to the hidden state problem in many other classes of single-molecule experiments, such as ion channel currents [**? ? ? ?**], single-molecule FRET [**? ? ? ? ? ?**], and the stepping of motor proteins [**? ? ?**]. In applying hidden Markov modeling to the analysis of single-molecule force spectroscopy data, the observed force or extension trace is assumed to come from a realization of an underlying Markov chain, where the system makes history-independent transitions among a set of discrete conformational states with probabilities governed by a transition or rate matrix. Under a given set of external force conditions, each state has a distribution of forces or extensions associated with it.

Given observed timeseries data for forces or extensions, the maximum likelihood estimate (MLE) of the model parameters (transition rates and state force or extension distributions) and sequence of hidden states corresponding to the observed data (used to color the points in Fig. 4) can be determined by standard methods [**? ?**], as demonstrated in recent work [**?**]. Unfortunately, this approach has a number of significant drawbacks. The parameters estimated by the MLE might be subject to enormous uncertainty. Methods for estimating the standard error or confidence intervals from hidden Markov models [**?**] make a number of approximations that can lead to a significant underestimation of the error. Worse yet, the standard algorithms employed to find the MLE may not even find the true maximum likelihood solution, instead converging to a local maximum in likelihood that is far from optimal [**?**].

Here, we resolve this issue through the use of a Bayesian extension of hidden Markov models [**? ? ? ?**] applicable to single molecule force experiments. By sampling over the posterior distribution of model parameters and hidden state assignments instead of simply finding the most likely values, the experimenter is able to accurately characterize the correlated uncertainties in both the model parameters (transition rates and state force or extension distributions) and hidden state sequences corresponding to observed data. Additionally, prior information (either from additional independent measurements or physical constraints) can be easily incorporated. We also include a reversibility constraint on the transition matrix—in which microscopic detailed balance is imposed on the kinetics, as dictated by the physics of equilibrium systems [**?**]—which has been shown to significantly reduce statistical uncertainties in data-poor conditions [**? ?**]. The framework we present is based on Gibbs sampling [**? ?**], allowing simple swap-in replacement of models for observable distributions, extension to multiple observables, and alternative models for state transitions. Additionally, the Bayesian method provides a straightforward way to model the statistical outcome and assess the utility of additional experiments given some preliminary data, providing the experimenter with a powerful tool for assessing whether time spent collecting additional data will yield sufficient reductions in uncertainty to allow competing hypotheses to be differentiated with statistical significance.

To demonstrate the utility of the Bayesian HMM in the analysis of real force spectroscopy data, we illustrate its application to the analysis of an RNA hairpin exhibiting apparent three-state behavior (Figure 4) and a protein system that exhibits two-state behavior with highly overlapping force signatures (Figure **??**) collected in a passive (equilibrium) optical trap. In addition to maximum likelihood estimates of equilibrium state probabilities, state-to-state transition rates, and the force distributions characterizing each state, the Bayesian extension provides estimates of confidence intervals for individual parameters (Tables II and **??**) as well as the full joint distribution of parameters.

A Matlab implementation of the approach described here is also made freely available online [`http://github.com/choderalab/bhmm`].

II. HIDDEN MARKOV MODELS FOR FORCE SPECTROSCOPY

Suppose the temporal behavior of some observable $O(x)$ that is a function of molecular configuration x—here, the force or molecular extension—is observed at temporal intervals Δt to produce a timeseries o_t, where $t = 0, 1, \ldots, L$. An instantaneous observation o_t does not necessarily contain enough information to unam-

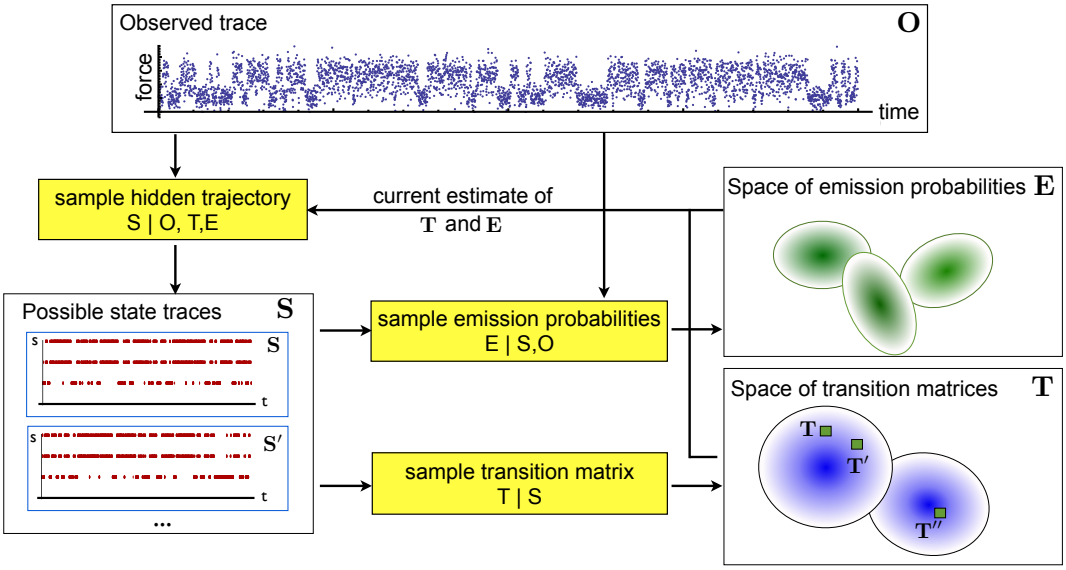

FIG. 2. Illustration of the Bayesian hidden Markov model (BHMM) sampling algorithm. During each BHMM sampling iteration, a modified forward-backward algorithm is used to sample a new hidden state trace $\mathbf{S}$, which is then used to generate a new sample of the reversible transition matrix $\mathbf{T}$ using MCMC sampling, followed by sampling new state emission probabilities $\mathbf{E}$. Collectively, these updates sample from the Bayesian posterior of the HMM and fully characterize the uncertainty in the kinetic model given the data.

biguously identify the current conformational state the molecule occupies; to infer the hidden state, we must also make use of the temporal information in the observed trace. We restrict ourselves to consideration of scalar functions $O(x)$, but the generalization to multidimensional probes (or multiple probes, such as combined force and fluorescence measurements [**?**]) and multiple observed temporal traces is straightforward.

We presume the system under study has M kinetically distinct states, in the sense that the system generally remains in a given state for several observation intervals Δt, but these states may not necessarily represent highly populated states of the system at equilibrium. We treat these conformational states as the hidden states of the model, because we cannot directly observe the identity of the metastable state in which the system resides. The hidden Markov model presumes the observed data $\mathbf{O} \equiv \{o_t\}$ was generated according to the following model dependent on the hidden states $\mathbf{S} \equiv \{s_t\}$, parameters $\mathbf{\Theta} \equiv \{\mathbf{T}, \mathbf{E}\}$, where $\mathbf{T}$ is an $M \times M$ row-stochastic transition matrix and $\mathbf{E}$ a set of emission parameters governing the observable (force or extension) distributions for each of the M hidden states, prior information about the initial state distribution $\boldsymbol{\rho}$, and emission probability distribution $\varphi(o|\mathbf{e})$,

$$\begin{aligned} \mathrm{P}(s_0) &= \rho_{s_0} \\ \mathrm{P}(s_t \mid s_{t-1}, \mathbf{T}) &= T_{s_{t-1} s_t}, \;\; t \geq 1 \\ \mathrm{P}(o_t \mid s_t, \mathbf{e}_{s_t}) &= \varphi(o_t \mid \mathbf{e}_{s_t}). \end{aligned} \tag{1}$$

In diagrammatic form, the observed state data $\{o_t\}$ and corresponding hidden state history $\{s_t\}$ can be represented

$$\begin{array}{ccccccccc} \xrightarrow{\boldsymbol{\rho}} & s_0 & \xrightarrow{\mathbf{T}} & s_1 & \xrightarrow{\mathbf{T}} & s_2 & \xrightarrow{\mathbf{T}} & \cdots & \xrightarrow{\mathbf{T}} & s_L \\ & \downarrow \varphi & & \downarrow \varphi & & \downarrow \varphi & & & & \downarrow \varphi \\ & o_0 & & o_1 & & o_2 & & & & o_L \end{array} \tag{2}$$

The initial state distribution $\boldsymbol{\rho}$ reflects our knowledge of the initial conditions of the experiment that collected data $\mathbf{O}$. In the case that the experiment was prepared in equilibrium, $\boldsymbol{\rho}$ corresponds to the equilibrium distribution $\boldsymbol{\pi}$ of the model transition matrix $\mathbf{T}$; if the experiment was prepared out of equilibrium, $\boldsymbol{\rho}$ may be chosen to reflect some other prior distribution (e.g. the uniform prior).

State transitions $(s_{t-1} \rightarrow s_t)$ are governed by the discrete transition probability $T_{s_{t-1} s_t}$. The Markov property of HMMs prescribes that the probability that a system originally in state i at time t is later found in state j at time $t + \Delta t$ is dependent only on knowledge of the state i, and given by the corresponding matrix element T_{ij} of the (row-stochastic) transition matrix $\mathbf{T}$. The choice of Δt is critical for the accuracy of the model [**?**] — Δt must be chosen to reside in a time-scale gap between the M slow relaxation processes that are resolved and faster relaxation processes that are not resolved by the model.

The dependence of the experimental observable on the hidden state is governed by the emission probability $\varphi(o_t \mid \mathbf{e}_{s_t})$, parametrized by observable emission parameters $\mathbf{e}$. For example, in the force spectroscopy applications described here, $\varphi(o \mid \mathbf{e}_s)$ is taken to be a univariate normal (Gaussian) distribution parameterized by a mean μ and variance σ^2 that characterize each state, such that $\mathbf{e}_i \equiv \{\mu_i, \sigma_i^2\}$. Other choices of observable distribution can easily be substituted in a modular way without affecting the structure of the algorithms presented here.

Given the HMM process specified in Eq. 1, the probability of observing data $\mathbf{O}$ given the model parameters $\mathbf{\Theta}$ is then,

$$\mathrm{P}(\mathbf{O} \mid \mathbf{\Theta}) = \sum_{\mathbf{S}} \rho_{s_0} \varphi(o_0 \mid \mathbf{e}_{s_0}) \prod_{t=1}^{L} T_{s_{t-1} s_t} \varphi(o_t \mid \mathbf{e}_{s_t}), \tag{3}$$

where the sum over hidden state histories $\mathbf{S}$ is shorthand for

$$\sum_{\mathbf{S}} \equiv \sum_{s_0=1}^{M} \sum_{s_1=1}^{M} \cdots \sum_{s_L=1}^{M}. \tag{4}$$

If multiple independent traces $\{o_t\}$ are available, the probability $\mathrm{P}(\mathbf{O} \mid \mathbf{\Theta})$ is simply the product of Eq. 3 for the independent traces.

A. Maximum likelihood hidden Markov model (MLHMM)

The standard approach to construct an HMM from observed data is to compute the maximum likelihood estimator (MLE) for the model parameters $\boldsymbol{\Theta} \equiv \{\mathbf{T}, \mathbf{E}\}$, which maximize the probability of the observed data $\mathbf{O}$ given the model,

$$\hat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} \mathrm{P}(\mathbf{O} \mid \boldsymbol{\Theta}), \tag{5}$$

yielding MLE estimates of transition matrix $\hat{\mathbf{T}}$ and state emission parameters $\hat{\mathbf{E}}$. Typically, determination of the model parameters $\boldsymbol{\Theta}$ is carried out using the Baum-Welch algorithm [**?**].

Once the MLE parameters $\hat{\boldsymbol{\Theta}}$ are determined, the most likely hidden state history that produced the observations $\mathbf{O}$ can be determined using these parameters:

$$\hat{\mathbf{S}} = \arg\max_{\mathbf{S}} \mathrm{P}(\mathbf{S} \mid \mathbf{O}, \hat{\boldsymbol{\Theta}}). \tag{6}$$

This is typically carried out using the Viterbi algorithm [**?**], a classic example of dynamic programming.

B. Bayesian hidden Markov model (BHMM)

Instead of simply determining the model that maximizes the likelihood of observing the data $\mathbf{O}$ given the model parameters $\boldsymbol{\Theta}$, we here propose to sample to full posterior distribution of model parameters given the observed data. Towards this end, we emply Bayes' theorem to compute the posterior distribution:

$$\mathrm{P}(\boldsymbol{\Theta} \mid \mathbf{O}) \propto \mathrm{P}(\mathbf{O} \mid \boldsymbol{\Theta})\mathrm{P}(\boldsymbol{\Theta}). \tag{7}$$

Here, $\mathrm{P}(\boldsymbol{\Theta})$ denotes a prior distribution that encodes any *a priori* information we may have about the model parameters $\boldsymbol{\Theta}$. This prior information might include, for example, physical constraints (such as ensuring the transition matrix satisfies detailed balance) or prior rounds of inference from other independent experiments.

Making use of the likelihood (Eq. 3), the model posterior is then given by,

$$\mathrm{P}(\boldsymbol{\Theta} \mid \mathbf{O}) \propto \mathrm{P}(\boldsymbol{\Theta}) \sum_{\mathbf{S}} \rho_{s_0} \varphi(o_0 \mid \mathbf{e}_{s_0}) \prod_{t=1}^{L} T_{s_{t-1} s_t} \varphi(o_t \mid \mathbf{e}_{s_t}) \tag{8}$$

Drawing samples of $\boldsymbol{\Theta}$ from this distribution will, in principle, allow the confidence with which individual parameters and combinations thereof are known, given the data (and subject to the validity of the model of Eq. 1 in correctly representing the process by which the observed data is generated). While the posterior $\mathrm{P}(\boldsymbol{\Theta}|\mathbf{O})$ is complex, we could in principle use a Markov chain Monte Carlo (MCMC) approach [**?**] to sample it. In its current form, however, this would be extremely expensive due to the sum over all hidden state histories $\mathbf{S}$ appearing in ratios involving Eq. 8. Instead, we introduce the hidden state histories $\mathbf{S}$ as an auxiliary variable, sampling from the augmented posterior,

$$\mathrm{P}(\boldsymbol{\Theta}, \mathbf{S} \mid \mathbf{O}) \propto \left[\rho_{s_0} \varphi(o_0 \mid \mathbf{e}_{s_0}) \prod_{t=1}^{L} T_{s_{t-1} s_t} \varphi(o_t \mid \mathbf{e}_{s_t}) \right] \mathrm{P}(\boldsymbol{\Theta}). \tag{9}$$

which makes it much less costly to compute the ratios required for MCMC on the augmented $(\boldsymbol{\Theta}, \mathbf{S})$ parameter space.

If we presume the prior is separable, such that $\mathrm{P}(\boldsymbol{\Theta}) \equiv \mathrm{P}(\mathbf{T})\mathrm{P}(\mathbf{E})$, we can sample from the augmented posterior (Eq. 9) using the framework of Gibbs sampling [**?**], in which the augmented model parameters are updated by cycles of sampling from conditional distributions,

$$\begin{aligned}
\mathbf{S}' &\sim \mathrm{P}(\mathbf{S} \mid \mathbf{T}, \mathbf{E}, \mathbf{O}) \propto \rho_{s_0} \varphi(o_0 \mid \mathbf{e}_{s_0}) \prod_{t=1}^{L} T_{s_{t-1} s_t} \varphi(o_t \mid \mathbf{e}_{s_t}) \\
\mathbf{T}' &\sim \mathrm{P}(\mathbf{T} \mid \mathbf{E}, \mathbf{S}', \mathbf{O}) = \mathrm{P}(\mathbf{T} \mid \mathbf{S}') \propto \mathrm{P}(\mathbf{T}) \prod_{t=1}^{L} T_{s'_{t-1} s'_t} \\
\mathbf{E}' &\sim \mathrm{P}(\mathbf{E} \mid \mathbf{S}', \mathbf{T}', \mathbf{O}) = \mathrm{P}(\mathbf{E} \mid \mathbf{S}', \mathbf{O}) \propto \mathrm{P}(\mathbf{E}) \prod_{t=0}^{L} \varphi(o_t \mid \mathbf{e}_{s'_t}).
\end{aligned} \tag{10}$$

The equalities on the second and third lines reflect the conditional independence of the hidden Markov model defined by Eq. 1. When only the model parameters $\boldsymbol{\Theta} \equiv \{\mathbf{T}, \mathbf{E}\}$ or the hidden state histories $\mathbf{S}$ are of interest, we can simply marginalize out the uninteresting variables by sampling from the augmented joint posterior for $\{\mathbf{T}, \mathbf{E}, \mathbf{S}\}$ and examine only the variables of interest.

In the applications presented here, we employ a Gaussian observable distribution model for $\varphi(o \mid \mathbf{e})$, which is expected to be a good model of the distribution of observed force or positions o for force spectroscopy experiments,

$$\varphi(o \mid \mathbf{e}) = \varphi(o \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \frac{(o-\mu)^2}{\sigma^2} \right], \tag{11}$$

where μ is the mean force or extension characterizing a particular state, and σ is the standard deviation or width of forces or extensions corresponding to that state. Other choices for the observable distribution model $\phi(o|\mathbf{e})$ can easily be substituted should they be more appropriate for analyzing particular classes of experiments, and only require substituting an appropriate parameter update step. We note that marginal posterior distributions of each mean $\mathrm{P}(\mu_i|\mathbf{O})$ reflect the statistical uncertainty in how well the mean force or position is determined, and need not correspond to the standard deviation σ_i, which may be much broader (or narrower, depending on the situation).

Detailed descriptions of the algorithms used to fit the maximum likelihood hidden Markov model (MLHMM) and sample from the posterior distribution of the Bayesian hidden Markov model (BHMM) are given in the Appendix.

III. VALIDATION USING SYNTHETIC DATA

To verify that our BHMM posterior sampling scheme reflects the true uncertainty in the model parameters, we can test that the scheme produces the expected estimates for synthetic data generated from a model with known parameters $\boldsymbol{\Theta}^*$. Given observed data $\mathbf{O}$ generated from $\mathrm{P}(\mathbf{O} \mid \boldsymbol{\Theta}^*)$, sampling from the posterior $\mathrm{P}(\boldsymbol{\Theta} \mid \mathbf{O})$ using the scheme described in *Sampling from the posterior of the BHMM* will provide us with confidence intervals $[\theta_{\mathrm{low}}, \theta_{\mathrm{high}}]$ for a specified confidence interval $\alpha \in [0, 1]$. If these computed confidence intervals are accurate, we should find that the true model parameter θ^* lies in the computed confidence interval $[\theta_{\mathrm{low}}^{(\alpha)}, \theta_{\mathrm{high}}^{(\alpha)}]$ with probability α. A 95% confidence interval ($\alpha = 0.95$), for example, should contain the BHMM estimate in 95% of such trials. This can be tested by generating synthetic observed data $\mathbf{O}$ from $\mathrm{P}(\mathbf{O} \mid \boldsymbol{\Theta}^*)$ and verifying that we find $\theta^* \in [\theta_{\mathrm{low}}^{(\alpha)}, \theta_{\mathrm{high}}^{(\alpha)}]$ in a fraction α of these synthetic experiments.

As an illustration of the computation of confidence intervals, we simulated a three-state system intended to mimic a protein with

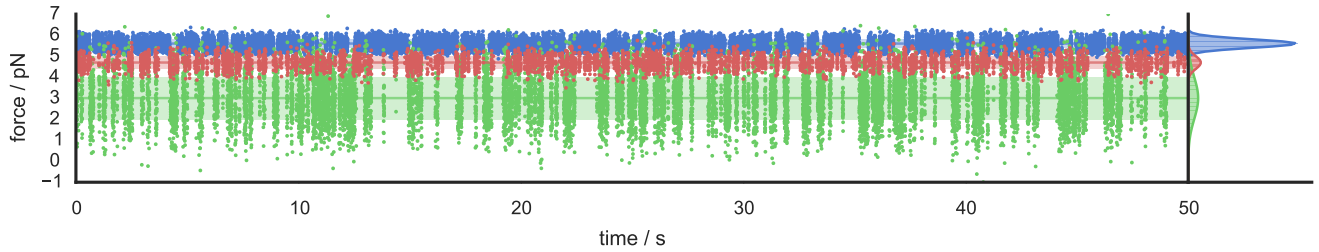

FIG. 3. Synthetic force trajectory and inferred state assignments in MLHMM. Observed samples are colored by their hidden state assignments. Dark horizontal lines terminating in triangles to the right denote state means, while lightly colored bands indicate one standard deviation on either side of the state mean. The histograms on the right side shows the state-resolved probabilities of samples, while the colored peaks show the weighted Gaussian output contribution from each state, and the black outline the weighted sum of the Gaussian output contributions from the HMM states.

			Estimated Model Parameters		
Property		**True Value**	**1 000 observations**	**10 000 observations**	**100 000 observations**
stationary probability	π_1	0.308	$0.228^{0.480}_{0.074}$	$0.318^{0.407}_{0.244}$	$0.324^{0.355}_{0.292}$
	π_2	0.113	$0.093^{0.172}_{0.042}$	$0.124^{0.155}_{0.098}$	$0.112^{0.121}_{0.104}$
	π_3	0.579	$0.679^{0.870}_{0.415}$	$0.558^{0.648}_{0.455}$	$0.564^{0.599}_{0.531}$
transition probability	T_{11}	0.980	$0.970^{0.987}_{0.945}$	$0.972^{0.978}_{0.966}$	$0.979^{0.981}_{0.978}$
	T_{12}	0.019	$0.023^{0.045}_{0.009}$	$0.026^{0.032}_{0.021}$	$0.020^{0.021}_{0.018}$
	T_{13}	0.001	$0.007^{0.018}_{0.001}$	$0.002^{0.003}_{0.001}$	$0.001^{0.001}_{0.001}$
	T_{21}	0.053	$0.054^{0.106}_{0.018}$	$0.067^{0.082}_{0.053}$	$0.057^{0.061}_{0.052}$
	T_{22}	0.900	$0.868^{0.931}_{0.790}$	$0.890^{0.907}_{0.870}$	$0.897^{0.903}_{0.892}$
	T_{23}	0.050	$0.078^{0.136}_{0.035}$	$0.043^{0.056}_{0.033}$	$0.046^{0.050}_{0.042}$
	T_{31}	0.001	$0.002^{0.006}_{0.000}$	$0.001^{0.002}_{0.000}$	$0.001^{0.001}_{0.000}$
	T_{32}	0.009	$0.010^{0.019}_{0.004}$	$0.010^{0.012}_{0.007}$	$0.009^{0.010}_{0.008}$
	T_{33}	0.990	$0.988^{0.995}_{0.978}$	$0.990^{0.992}_{0.987}$	$0.990^{0.991}_{0.989}$
state mean force (pN)	μ_1	3.000	$2.947^{3.082}_{2.812}$	$2.998^{3.033}_{2.963}$	$3.001^{3.013}_{2.990}$
	μ_2	4.700	$4.666^{4.721}_{4.612}$	$4.699^{4.716}_{4.683}$	$4.702^{4.707}_{4.696}$
	μ_3	5.600	$5.597^{5.614}_{5.583}$	$5.602^{5.607}_{5.596}$	$5.602^{5.603}_{5.600}$
state std dev force (pN)	σ_1	1.000	$1.037^{1.134}_{0.951}$	$0.992^{1.018}_{0.967}$	$0.999^{1.007}_{0.991}$
	σ_2	0.300	$0.254^{0.300}_{0.217}$	$0.287^{0.300}_{0.275}$	$0.301^{0.305}_{0.296}$
	σ_3	0.200	$0.200^{0.211}_{0.190}$	$0.203^{0.207}_{0.199}$	$0.201^{0.203}_{0.200}$

TABLE I. Estimated mean model parameters and confidence intervals for synthetic timeseries data

(1) a highly-compliance, low-force unfolded state, (2) a moderately compliant low-population intermediate at intermediate force, and (3) a low-compliance, high-force folded state. Here, the term "compliance" refers to the width of the force or extension distribution characterizing the state. Parameters of the model are given in Table I, and the observation interval was taken to be $\tau = 1$ ms. An example realization of a model trajectory mimicking a passive-mode experiment—in which the trap was stationary—is shown in Figure 3 along with the MLHMM state assignment. We generated a trajectory of 100 000 observations, and characterized the BHMM mean parameter estimate and 95% confidence intervals for a subset of this trajectory of varying lengths. The results, shown in Table I, show that the confidence intervals contract as trajectory length increases, as expected, and the BHMM-computed 95% confidence intervals contain the true model parameters with the expected statistics. In contrast, a model created from simply segmenting the observed forces into disjoint region and assigning state membership based on the force value alone estimates model parameters with significant bias even for 1 000 000 observations (see *Supporting Information*).

As a more rigorous test, we sampled 50 random models from the prior $\mathrm{P}(\boldsymbol{\Theta})$ with two to six states, generated a 10 000 observation synthetic trajectory for each, and accumulated statistics on the observed fraction of time the true model parameters were within the BHMM confidence intervals for various values of the confidence interval width α. The results of this test are depicted in *Supplementary Figure 1*. We expect that the plot traces the diagonal if the observed and expected confidence intervals are identical; an overestimate of the confidence interval will be above the diagonal, and an underestimate will fall below it. Because only a finite number of independent replicates of the experiment are conducted, there is some associated uncertainty with the observed confidence intervals. The results show that the observed confidence intervals line up with the expected confidence intervals to within statistical error, suggesting the BHMM confidence intervals neither underestimate nor overestimate the actual uncertainty in model parameters.

IV. APPLICATIONS

A. Apomyoglobin at pH 5 in a passive optical trap

[JDC: This section remains to be completed.]

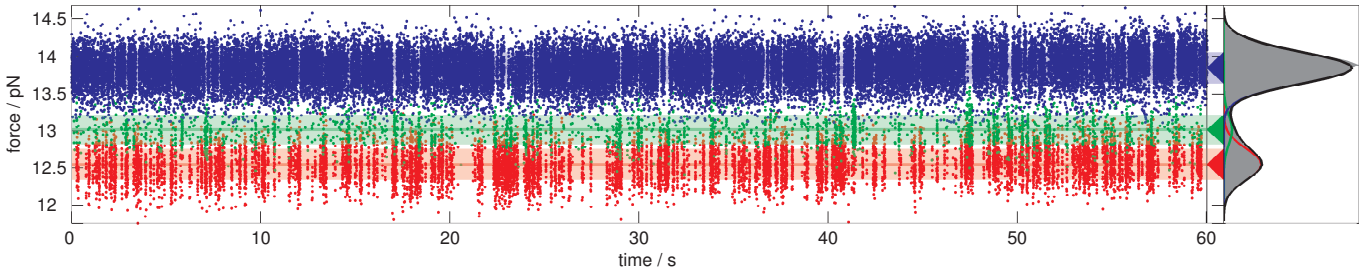

FIG. 4. Experimental force trajectory of an RNA hairpin and MLHMM state assignments. Observed samples are colored by their hidden state assignments. Dark horizontal lines terminating in triangles to the right denote state means, while lightly colored bands indicate one standard deviation on either side of the state mean. The gray histogram on the right side shows the total observed probability of samples, while the colored peaks show the weighted Gaussian output contribution from each state, and the black outline the weighted sum of the Gaussian output contributions from the HMM states.

Property		Value
Equilibrium probability	π_1	$0.215\,^{0.236}_{0.193}$
	π_2	$0.046\,^{0.050}_{0.041}$
	π_3	$0.740\,^{0.762}_{0.717}$
Transition probability ($\Delta t = 1$ ms)	T_{11}	$0.954\,^{0.959}_{0.950}$
	T_{12}	$0.033\,^{0.037}_{0.029}$
	T_{13}	$0.013\,^{0.015}_{0.011}$
	T_{21}	$0.154\,^{0.169}_{0.139}$
	T_{22}	$0.650\,^{0.673}_{0.627}$
	T_{23}	$0.196\,^{0.216}_{0.180}$
	T_{31}	$0.004\,^{0.004}_{0.003}$
	T_{32}	$0.012\,^{0.013}_{0.011}$
	T_{33}	$0.984\,^{0.985}_{0.983}$
State force mean (pN)	μ_1	$12.549\,^{12.552}_{12.544}$
	μ_2	$13.016\,^{13.027}_{13.006}$
	μ_3	$13.849\,^{13.852}_{13.848}$
State force std dev (pN)	σ_1	$0.210\,^{0.213}_{0.207}$
	σ_2	$0.201\,^{0.208}_{0.193}$
	σ_3	$0.213\,^{0.214}_{0.211}$
Transition rate (s^{-1})	k_{12}	$41.4\,^{46.6}_{36.3}$
	k_{13}	$9.1\,^{11.3}_{7.2}$
	k_{21}	$194.7\,^{216.7}_{173.1}$
	k_{23}	$243.7\,^{271.5}_{219.0}$
	k_{31}	$2.6\,^{3.2}_{2.1}$
	k_{32}	$15.0\,^{16.6}_{13.4}$
State mean lifetime (ms)	τ_1	$21.9\,^{24.1}_{20.0}$
	τ_2	$2.9\,^{3.1}_{2.7}$
	τ_3	$63.1\,^{68.5}_{58.4}$

TABLE II. BHMM model estimates for an RNA hairpin data.

B. RNA hairpin in a passive optical trap

We illustrate the BHMM approach applied to real force spectroscopy data by characterizing the average forces and transition rates among kinetically distinct states of an RNA hairpin in an optical trap under passive (equilibrium) conditions.

A sample of p5ab RNA hairpin from *Tetrahymena thermophilia* utilized in a previous study [**?**] was provided by Jin-Der Wen, and prepared as previously described [**?**]. Within the population of RNA hairpin molecules in the examined sample, there were two chemically distinct species, exhibiting either apparent two-state (as reported previously [**?**]) or three-state behavior (studied here). For the purposes of testing this method, we examined a fiber that appeared to consistently exhibit three-state behavior upon visual inspection of the force timeseries data.

The instrument used in this experiment was a dual-beam counter-propagating optical trap with a spring constant of 0.08 pN/nm. A piezoactuator controlled the position of the trap and allowed position resolution to within 0.5 nm [**?**]. The instrument was used in passive (equilibrium) mode, in which the trap remained stationary relative to the pipette into which one bead of the tether was suctioned (corresponding to the geometry depicted in Fig. 1). The voltage on the position-sensitive detectors, reporting on the force on the bead held in the optical trap, was recorded at 50 kHz using a 32-bit ADC operating with ∼600 ADC units/pN sensitivity. Each recorded experimental trace was 60 s in duration. From averaging forces over 1 s intervals at the initial and final parts of each trace, average force drift in the instrument was estimated to be 0.0015(5) pN/s. The appropriate observation interval (1 ms) for BHMM analysis was determined by examination of force autocorrelation functions, which suggested that intrastate motion had completely decayed by 1 ms (see *Supplementary Material: Choice of observation interval*). [FN: Update the selection of the HMM lag time. I suggest fixing the lag time τ first, running a maximum-likelihood HMM with N states, and then select $M < N$ based on how many relaxation timescales are larger than τ] The observed 50 kHz force traces were therefore subsampled to 1 kHz for analysis by retaining every 50th sample.

A single observed force trajectory at a fixed trap position adequate to cause hopping among multiple states is shown in Figure 4. The most likely state trajectory from the MLHMM fit with three states is shown by coloring the observations most likely to be associated with each state, with bands of color indicating the mean and standard deviation about the mean force characterizing each state.

[Describe how/why you selected three over two states in this model] Table II lists the BHMM posterior means and confidence intervals characterizing the three-state model extracted from this single 60 s observed force trace. Several things are notable about the estimated model parameters. Surprisingly, while there is a clearly-resolved intermediate-force state (*state 2*) through which most of the flux from the high- and low-force states passes (as seen from large K_{12} and K_{23}), there are nontrivial rate constants connecting the high and low force states directly (K_{13}), indicating that while a sequential mechanism involving passing through the intermediate state is preferred, it may not be an obligatory step in hairpin formation under these conditions. While the state mean forces are clearly distinct, the state standard deviations—which reflect the width of the observed force distribution characterizing each

state, rather than the uncertainty in state means—possess overlapping confidence intervals. These standard deviations reflect not only contributions from both the distribution of extensions sampled by the hairpin in each conformational state, but also from fluctuations in the handles and beads, and other sources of mechanical and electrical noise in the measurement. The overlapping confidence intervals in these standard deviations suggest there is no significant change in the overall stiffness of the tether upon unfolding.

Finally, the lifetime of the intermediate-force state is significantly shorter than for the low- and high-force states by nearly an order of magnitude, and only a few times longer than the observation interval of 1 ms—despite this, the lifetime appears to be well-determined, as indicated by the narrow confidence intervals.

V. DISCUSSION

We have described an approach to determining the first-order kinetic parameters and observable (force or extension) distributions characterizing conformational states in single-molecule force spectroscopy. By use of a Bayesian extension of hidden Markov models, we are able to characterize the experimental uncertainty in these parameters due to instrument noise and finite-size datasets. The use of a detailed balance constraint additionally helps reduce the experimental uncertainty over standard hidden Markov models, as both transitions into and out of conformational states provide valuable information about state kinetics and populations in data-poor conditions [**?** **?**]. Additionally, the Gibbs sampling framework used to sample from the Bayesian posterior can be easily extended to incorporate additional nuisance parameters, such as stochastic models of measurement noise or laser power fluctuations.

Drift has historically been a challenging problem for force spectroscopy measurements, with experiments using previous-generation instruments resorting to short force-ramp experiments to minimize the cumulative effects of drift in recorded traces [**?**]. If the drift only affects the measured signal (e.g. force or bead position), it is possible to correct for this drift during the analysis procedure using various approaches, such as the introduction of nuisance parameters characterizing linear drift (along with associated priors), or heuristic corrections applied directly to the observed data [**?**]. However, compensating for significant drift in the physical trap-pipette or trap-trap distance in analysis is difficult, as changes in the relative optical trap position will change the external potential felt by the molecule, and hence alter the kinetics and thermodynamics. Without imposing a model of how those change with this distance, as well as a physical or empirical model for the dynamics of drift, it is impossible to include these effects during analysis.

Fortunately, obtaining many drift-free measurement traces of the duration presented here with current-generation miniaturized optical traps that employ a pipette to suction one of the beads in the tether (such as the one employed here [**?**]), the collection of data sufficiently free of drift is regularly achievable [**?**]. Moreover, drift is even less problematic for newer generations of instrument that utilize two optical traps instead of a pipette [**?** **?**].

We have opted to make use of a reversible transition matrix to describe the statistical kinetic behavior between the observation intervals Δt, but we note that it is possible to use a reversible rate matrix instead by substituting a rate matrix sampling scheme [**?**] in the appropriate stage of the Gibbs sampling updates. Because the use of a rate matrix instead of a transition matrix involves a parametric change of variables, an appropriate prior must be selected, or a Jacobian required to obtain the same posterior distribution as we obtain sampling the reversible transition matrix.

While the experimenter must currently choose the number of conformational states by hand, a number of extensions of Bayesian hidden Markov models can be used to automatically determine the number of states best supported by the data, including reversible-jump schemes [**?** **?**] and variational Bayes methods [**?** **?**].

We note that the experimenter in principle has access to the full posterior distribution of models given the observed data, so that instead of looking at the confidence of single parameters, confidence intervals in more complex functions of parameters—such as the rates or lifetimes in Table II—can be computed, or joint posterior distributions of multiple parameters examined. It is also possible to generate synthetic data from the current model, or family of models, to examine how the collection of additional data will further reduce uncertainties or allow discrimination among particular hypotheses. The field of Bayesian experimental design [**?**] holds numerous possibilities for selecting how future experiments can maximize information gain, and whether the information gain from the collection of additional data will be of sufficient utility to justify the expense.

VI. ACKNOWLEDGMENTS

The authors thank Sergio Bacallado (Stanford University) and anonymous referees for helpful feedback on an earlier version of this manuscript, and Steve Presse (UCSF) for engaging discussions on this topic. JDC acknowledges support from a QB3-Berkeley Distinguished Postdoctoral Fellowship and the Memorial Sloan Kettering Cancer Center. BGK gratefully acknowledges the support by the German Research Foundation (DFG) within the collaborative research center SFB 765 (associated project). FN acknowledges funding by DFG Grant 825/2 and ERC starting grant “pcCell”. This work was supported in part by a grant from the NSF (SM).

VII. APPENDIX: ALGORITHMS

A. Generating an initial model

To initialize either computation of the MLHMM or sampling from the posterior for the BHMM, an initial model that respects any constraints imposed in the model prior $\mathrm{P}(\boldsymbol{\Theta})$ must be selected. Here, we employ a Gaussian observable distribution model for $\varphi(o \mid \mathbf{e})$ (Eq. 11) and enforce that the transition matrix $\mathbf{T}$ satisfy detailed balance.

[JDC: We need to revise this section to discuss the new randomized multi-start initialization.]

1. *Observable parameter estimation*

We first initialize the observed distributions of each state by fitting a Gaussian mixture model with M states to the pooled observed data $\mathbf{O}$, ignoring temporal information:

$$\mathrm{P}(\mathbf{O} \mid \boldsymbol{\pi}, \mathbf{E}) = \prod_{t=0}^{L} \sum_{m=1}^{M} \pi_m \varphi(o_t \mid \mu_m, \sigma_m^2), \tag{12}$$

where the state observable emission probability vector $\mathbf{E} \equiv \{\mathbf{e}_1, \ldots, \mathbf{e}_M\}$ and $\mathbf{e}_m \equiv \{\mu_m, \sigma_m^2\}$ with μ_m denoting the observable mean and σ_m^2 the variance for state m for the Gaussian mixture model. The vector $\boldsymbol{\pi}$ is composed of equilibrium state populations $\{\pi_1, \ldots, \pi_M\}$ with $\pi_m \geq 0$ and $\sum_{m=1}^{M} \pi_m = 1$.

A first approximation to $\boldsymbol{\pi}$ and $\mathbf{E}$ is computed by pooling and sorting the observed o_t, and defining M indicator functions $h_m(o)$ that separate the data into M contiguous regions of the observed range of o of roughly equal population. Let $N_m \equiv \sum_{t=0}^{L} h_m(o_t)$ denote the total number of observations falling in region m, and $N_{\text{tot}} = \sum_{m=1}^{M} N_m$. The initial parameters are then computed as,

$$\begin{aligned} \pi_m &= N_m / N_{\text{tot}} \\ \mu_m &= N_m^{-1} \sum_{t=0}^{L} o_t \, h_m(o_t) \qquad (13) \\ \sigma_m^2 &= N_m^{-1} \sum_{t=0}^{L} (o_t - \mu_m)^2 \, h_m(o_t). \qquad (14) \end{aligned}$$

To try to maximize the likelihood in Eq. 12, this approximation is then improved upon by iterating the expectation-maximization procedure described by Bilmes [**?**],

$$\begin{aligned} \pi'_m &= N_{\text{tot}}^{-1} \sum_{t=0}^{L} \chi_m(o_t, \mathbf{E}, \boldsymbol{\pi}) \\ \mu'_m &= (\pi'_m N_{\text{tot}})^{-1} \sum_{t=0}^{L} o_t \, \chi_m(o_t, \mathbf{E}, \boldsymbol{\pi}) \\ \sigma'^2_m &= (\pi'_m N_{\text{tot}})^{-1} \sum_{t=0}^{L} (o_t - \mu'_m)^2 \, \chi_m(o_t, \mathbf{E}, \boldsymbol{\pi}) \qquad (15) \end{aligned}$$

where the function $\chi_m(o, \mathbf{E}, \boldsymbol{\pi})$ is given by the fuzzy membership function,

$$\chi_m(o, \mathbf{E}, \boldsymbol{\pi}) = \frac{\pi_m \, \varphi(o \mid \mathbf{e}_m)}{\sum_{l=1}^{M} \pi_l \, \varphi(o \mid \mathbf{e}_l)}. \qquad (16)$$

The iterative procedure is terminated at iteration j when the change in the parameters $\{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2\}$ falls below a certain relative threshold, such as $\|\boldsymbol{\pi}^{[j]} - \boldsymbol{\pi}^{[j-1]}\|_2 / \|\boldsymbol{\pi}^{[j]}\|_2 < 10^{-4}$, where $\| \ldots \|_2$ denotes the euclidean L2 norm.

2. *Transition matrix estimation*

Once initial state observable emission parameters $\mathbf{E}$ are determined, an initial transition matrix is estimated using an iterative likelihood maximization approach that enforces detailed balance [**?**]. First, a matrix of fractional transition counts $\mathbf{C} \equiv (c_{ij})$ is estimated using the membership function:

$$c_{ij} = \sum_{t=1}^{L} \chi_i(o_{t-1}, \mathbf{E}, \boldsymbol{\pi}) \, \chi_j(o_t, \mathbf{E}, \boldsymbol{\pi}) \qquad (17)$$

A symmetric $M \times M$ matrix $\mathbf{X} \equiv (x_{ij})$ is initialized by

$$x_{ij} = x_{ji} = c_{ij} + c_{ji}. \qquad (18)$$

The iterative procedure described in Algorithm 1 of [**?**] is then applied to update the transition matrix in a manner that preserves the detailed balance constraint while maximizing the likelihood function $\mathrm{P}(\mathbf{T}|\mathbf{S})$. This algorithm proceeds as follows.

For each update iteration, we first update the diagonal elements of $\mathbf{X}$:

$$x'_{ii} = \frac{c_{ii}(x_{i*} - x_{ii})}{c_{i*} - c_{ii}} \;\; ; \;\; c_{i*} = \sum_{j=1}^{M} c_{ij} \;\; ; \;\; x_{i*} = \sum_{j=1}^{M} x_{ij}, \qquad (19)$$

followed by the off-diagonal elements:

$$x'_{ij} = x'_{ji} = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad (20)$$

where the quantities a, b, and c are computed from $\mathbf{X}$ and $\mathbf{C}$,

$$\begin{aligned} a &\equiv c_{i*} - c_{ij} + c_{j*} - c_{ji} \\ b &\equiv c_{i*}(x_{j*} - x_{ji}) + c_{j*}(x_{i*} - x_{ij}) \\ &\quad - (c_{ij} + c_{ji})(x_{i*} - x_{ij} + x_{j*} - x_{ji}) \\ c &\equiv -(c_{ij} + c_{ji})(x_{i*} - x_{ij})(x_{j*} - x_{ji}). \qquad (21) \end{aligned}$$

Once a sufficient number of iterations j have been completed to compute a stable estimate of $\mathbf{X}$ (such as the relative convergence criteria $\|\mathbf{X}^{[j]} - \mathbf{X}^{[j-1]}\|_2 / \|\mathbf{X}^{[j]}\|_2 < 10^{-4}$), the maximum likelihood transition matrix estimate $\mathbf{T}$ is computed as

$$T_{ij} = x_{ij} / x_{i*}. \qquad (22)$$

Note that the equilibrium probability vector $\boldsymbol{\pi}$ computed during the Gaussian mixture model fitting is not respected during this step. [FN: clarify which of the two π-vectors is used in the subsequent MLHMM]

B. Fitting a maximum likelihood HMM

The HMM parameters $\boldsymbol{\Theta} \equiv \{\mathbf{T}, \mathbf{E}\}$ are fit to the observed data $\mathbf{O}$ through use of the expectation-maximization (EM) algorithm [**?**]. This is an iterative procedure, where the model parameters are subsequently refined through successive iterations, according to the following scheme:

$$\mathbf{S}' = \arg\max_{S} \mathrm{P}(S \mid \mathbf{T}, \mathbf{E}, \mathbf{O}) \qquad (23)$$

$$\mathbf{T}' = \arg\max_{T} \mathrm{P}(T \mid \mathbf{S}, \mathbf{O}) \qquad (24)$$

$$\mathbf{E}' = \arg\max_{E} \mathrm{P}(E \mid \mathbf{S}, \mathbf{O}) \qquad (25)$$

The initial HMM is usually quick to compute, and can give the experimenter a rough idea of the model parameters, as well as providing a useful starting point for sampling models from the Bayesian posterior.

During each iteration, the Baum-Welch algorithm [**?**] is used to compute $\boldsymbol{\Xi} \equiv (\xi_{tij})$, which represents the probability that the system transitions from hidden state i at time $t-1$ to hidden state j at time t, and γ_{ti}, the probability that the system occupied state i at time t. This is accomplished by first executing the forward algorithm,

$$\alpha_{tj} = \begin{cases} \rho_j \, \varphi(o_0 \mid \mathbf{e}_j) & t = 0 \\ \varphi(o_t \mid \mathbf{e}_j) \sum_{i=1}^{M} \alpha_{(t-1)i} T_{ij} & t = 1, \ldots, L \end{cases} \qquad (26)$$

followed by the backward algorithm,

$$\beta_{ti} = \begin{cases} 1 & t = L \\ \sum_{j=1}^{M} T_{ij} \varphi(o_{t+1} \mid \mathbf{e}_j) \beta_{(t+1)j} & t = (L-1), \ldots, 0 \end{cases} \qquad (27)$$

The $L \times M \times M$ matrix $\boldsymbol{\Xi}$ is then computed for $t = 0, \ldots, (L-1)$ as,

$$\xi_{tij} = \alpha_{ti} \varphi(o_{t+1} \mid \mathbf{e}_i) T_{ij} \beta_{(t+1)j} / \sum_{i=1}^{M} \alpha_{Ti} \tag{28}$$

$$\gamma_{ti} = \sum_{j=1}^{M} \xi_{tij} \tag{29}$$

In practice, the logarithms of these quantities are computed instead to avoid numerical underflow.

The aggregate matrix of expected transition counts $\mathbf{C} \equiv (c_{ij})$ is then computed from $\boldsymbol{\Xi}$ as,

$$c_{ij} = \sum_{t=0}^{L-1} \xi_{tij}. \tag{30}$$

This count matrix is used to update the maximum-likelihood transition matrix $\mathbf{T}$ using the method of Prinz et al. [**?**] described in the previous section.

The state observable distribution parameters $\mathbf{E}$ are then updated from the γ_{ti}. For the univariate normal distribution applied to force spectroscopy data here, we update the mean μ_i and variance σ_i^2 for state i using the scheme,

$$\mu_i' = \frac{\sum_{t=0}^{L} o_t \gamma_{ti}}{\sum_{t=0}^{L} \gamma_{ti}} \;\; ; \;\; {\sigma'}_i^2 = \frac{\sum_{t=0}^{L} (o_t - \mu_i')^2 \gamma_{ti}}{\sum_{t=0}^{L} \gamma_{ti}}. \tag{31}$$

Once the model parameters have been fitted by iteration of the above update procedure to convergence (which may only converge to a local maximum of the likelihood), the most likely hidden state sequence can be determined given the observations $\mathbf{O}$ and the MLE model $\hat{\boldsymbol{\Theta}}$ using the Viterbi algorithm [**?**]. Like the forward-backward algorithm employed in the Baum-Welch procedure, the Viterbi algorithm also has a forward recursion component,

$$\epsilon_{jt} = \begin{cases} \rho_j \varphi(o_t \mid \mathbf{e}_j) & t = 0 \\ \varphi(o_t \mid \mathbf{e}_j) \max_i \epsilon_{i(t-1)} T_{ij} & t = 1, \ldots, L \end{cases} \tag{32}$$

$$\Phi_{jt} = \begin{cases} 1 & t = 0 \\ \arg\max_i \epsilon_{i(t-1)} T_{ij} & t = 1, \ldots, L \end{cases}$$

as well as a reverse reconstruction component to compute the most likely state sequence $\hat{\mathbf{S}}$,

$$\hat{s}_t = \begin{cases} \arg\max_i \epsilon_{it} & t = L \\ \Phi_{\hat{s}_{t+1}(t+1)} & t = (L-1), \ldots, 0 \end{cases} \tag{33}$$

C. Sampling from the posterior of the BHMM

Sampling from the posterior of the BHMM (Eq. 8) proceeds by rounds of Gibbs sampling, where each round consists of an update of the augmented model parameters $\{\mathbf{T}, \mathbf{E}, \mathbf{S}\}$ by sampling

$$\begin{aligned} \mathbf{S}' \mid \mathbf{T}, \mathbf{E}, \mathbf{O} &\sim \mathrm{P}(\mathbf{S}' \mid \mathbf{T}, \mathbf{E}, \mathbf{O}) \\ \mathbf{T}' \mid \mathbf{S}' &\sim \mathrm{P}(\mathbf{T}' \mid \mathbf{S}') \\ \mathbf{E}' \mid \mathbf{S}', \mathbf{O} &\sim \mathrm{P}(\mathbf{E}' \mid \mathbf{S}', \mathbf{O}) \end{aligned}$$

where the conditional probabilities are given by Eq. 10.

1. *Updating the hidden state sequences*

We use a modified form of the Viterbi process to generate an independent sample of the hidden state history $\mathbf{S}$ given the transition probabilities $\mathbf{T}$, state observable distribution parameters $\mathbf{E}$, and observed data $\mathbf{O}$. Like the Viterbi scheme, a forward recursion is applied to each observation trace $\mathbf{o}$, but instead of computing the most likely state history on the reverse pass, a new hidden state history $\mathbf{S}$ is drawn from the distribution $\mathrm{P}(\mathbf{S} \mid \mathbf{O}, \mathbf{T}, \mathbf{E})$. The forward recursion uses the same forward algorithm as used in Baum-Welch [**?**],

$$\alpha_{tj} = \begin{cases} \rho_j \, \varphi(o_0 \mid \mathbf{e}_j) & t = 0 \\ \varphi(o_t \mid \mathbf{e}_j) \sum_{i=1}^{M} \alpha_{(t-1)i} T_{ij} & t = 1, \ldots, L \end{cases} \tag{34}$$

In the reverse recursion, we now sample a state sequence by sampling each hidden state from the conditional distribution $s_t \sim \mathrm{P}(s_t \mid s_{t+1}, \ldots, s_L)$ starting from $t = L$ and proceeding down to $t = 0$, where the conditional distribution is given by,

$$\begin{aligned} &\mathrm{P}(s_t = i \mid s_{t+1}, \ldots, s_L) \\ &\propto \begin{cases} \alpha_{ti} / \sum_{j=1}^{M} \alpha_{tj} & t = L \\ \alpha_{ti} T_{i s_{t+1}} / \sum_{j=1}^{M} \alpha_{tj} T_{j s_{t+1}} & t = (L-1), \ldots, 0 \end{cases} \end{aligned} \tag{35}$$

It is straightforward to show the result of these sampling steps reconstitutes the probability distribution $\mathrm{P}(\mathbf{S}|\mathbf{T}, \mathbf{E}, \mathbf{O})$ (see *Supplementary Material: Proof of state history sampling scheme*).

2. *Updating the transition probabilities*

If we assume each row of the transition matrix $\mathbf{T}$ is independent with Dirichlet prior $\mathrm{P}(\mathbf{T})$, then it is possible to sample from the transition matrix posterior Dirichlet distribution $\mathrm{P}(\mathbf{T}' \mid \mathbf{S}')$, given the transition counts calculated from the sampled state sequence $\mathbf{S}'$ [**?**]. However, because physical systems in the absence of energy input through an external driving force should satisfy detailed balance, we make use of this constraint in updating our transition probabilities, since this has been demonstrated to substantially reduce parameter uncertainty in the data-limited regime [**?**].

Here we employ the reversible transition matrix sampling algorithm derived in [**FN: reference to reversible T-matrix estimation preprint when available**]. We first define a coordinate transform to unconditional transition probabilities:

$$x_{ij} = \pi_i T_{ij} \tag{36}$$

and sample from their posterior

$$p(\mathbf{X} \mid \mathbf{C}) = p(\mathbf{C} \mid \mathbf{X}) p(\mathbf{X}) \tag{37}$$

using the prior

$$p(\mathbf{X}) = \prod_{i \geq j} x_{ij}^{-1}. \tag{38}$$

This choice of prior ensures that the maximum likelihood estimator, i.e. the maximum of $p(\mathbf{C} \mid \mathbf{X})$ coincides with the mean of the posterior $p(\mathbf{X} \mid \mathbf{C})$. This behavior is desirable in order to have uncertainty intervals envelop the maximum likelihood estimator. Without a suitable choice of prior this property is lost, as described in [**?**].

It turns out that the posterior (37) can be conveniently sampled by considering yet another set of variables $\mathbf{V} = [v_{ij}]$, which are related to $\mathbf{X}$ by the scaling $x_{ij} = v_{ij} / \sum_{k,l} v_{kl}$. Thus, we can still

recover transition matrices from the v_{ij} variables by row normalization:

$$T_{ij} = \frac{v_{ij}}{\sum_k v_{ik}} \tag{39}$$

Using the v_{ij} variables, the sampling of posterior (37) is implemented by iterating the following Gibbs sampling procedure. Initially, when given an initial transition matrix $\mathbf{T}^{(0)}$, we set

$$v_{ij}^{(0)} = x_{ij}^{(0)} = \pi_i T_{ij}^{(0)} \tag{40}$$

When starting from scratch we initialize

$$v_{ij}^{(0)} = (c_i j + c_j i)/2 \sum_{k,l} c_{kl}. \tag{41}$$

In each Gibbs sampling cycle we iterate through all $i \geq j$. If $c_{ij} + c_{ji} = 0$, the conditional probability density is concentrated at $v_{ij} = 0$ and we do not perform any action. If $c_{ij} + c_{ji} > 0$, one of the following two steps is executed for either diagonal or off-diagonal elements:

Case 1: diagonals: If $i = j$, we sample $v_{ii}^{(k+1)}$ according to the conditional

$$\mathbb{P}(v_{ii}|\{v_{kl}\}_{k\geq l}\backslash\{v_{ii}\}, \mathbf{C}) \propto v_{ii}^{c_{ii}-1}(v_i - v_{ii})^{-c_i}. \tag{42}$$

$$\propto T_{ii}^{c_{ii}-1}(1 - T_{ii})^{c_i - c_{ii} - 1}. \tag{43}$$

where we have introduced the row sums

$$v_i = \sum_j v_{ij}. \tag{44}$$

The conditional (43) can be sampled by drawing Beta-distributed variables

$$T_{ii} \sim \mathrm{Beta}(\alpha, \beta) \tag{45}$$

with parameters $\alpha = c_{ii}$ and $\beta = c_i - c_{ii}$ (see [**?**]). The new element $v_{ii}^{(k+1)}$ can be obtained from the sampled T_{ii} by

$$v_{ii}^{(k+1)} = \frac{T_{ii}}{1 - T_{ii}}(v_i^{(k)} - v_{ii}^{(k)}). \tag{46}$$

Case 2: off-diagonals: If $i \neq j$, we have to sample a conditional density which is not easily tractable. Therefore, we approximate the conditional density as follows:

$$v_{ij}^{(c_{ij}+c_{ji}-1)}(v_i^{(k)} - v_{ij}^{(k)} + v_{ij})^{-c_i}(v_j^{(k)} - v_{ji}^{(k)} + v_{ij})^{-c_j} \approx q(v_{ij} \mid \alpha, \beta, f_0) \tag{47}$$

where employ the approximate function

$$q(v_{ij} \mid \alpha, \beta, f_0) = (v_{ij})^{-1} \exp\left(\alpha \log v_{ij} - \beta v_{ij} + f_0\right) \tag{48}$$

which is, up to a constant factor, a Gamma distribution with parameters α, β which can be efficiently sampled [**?**]. The parameters are given by

$$\alpha = -f''(\bar{v}_{ij})\bar{v}_{ij}^2 \tag{49a}$$

$$\beta = -f''(\bar{v}_{ij})\bar{v}_{ij} \tag{49b}$$

$$f_0 = f(\bar{v}_{ij}) + f''(\bar{v}_{ij})\bar{v}_{ij}^2(\log \bar{v}_{ij} - 1), \tag{49c}$$

with the maximum point $\bar{v}_{ij}$ as the positive root of a quadratic equation:

$$\bar{v}_{ij} = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{50}$$

and a, b, c given by:

$$a = c_i + c_j - c_{ij} - c_{ji} \tag{51a}$$

$$b = c_i(v_i^{(k)} - v_{ij}^{(k)}) + c_j(v_j^{(k)} - v_{ji}^{(k)}) - (c_{ij} + c_{ji})(v_i^{(k)} - v_{ij}^{(k)} + v_j^{(k)} - v_{ji}^{(k)}) \tag{51b}$$

$$c = -(c_{ij} + c_{ji})(v_i^{(k)} - v_{ij}^{(k)})(v_j^{(k)} - v_{ji}^{(k)}) \tag{51c}$$

The approximate conditional q in (48) is not necessarily a bounding envelope for the true conditional (47) so that instead of rejection sampling we use the approximation q as a proposal density in a Metropolis sampling step. The acceptance probability for the step is then given as $\min\{1, p_{acc}\}$ with

$$p_{acc} = \frac{q(v_{ij}^{(k)})v_{ij}^{(k)} \exp f(v_{ij}^{(k+1)})}{q(v_{ij}^{(k+1)})v_{ij}^{(k+1)} \exp f(v_{ij}^{(k)})}. \tag{52}$$

For each element sampled, symmetry of the $\mathbf{V}$ matrix is enforced by setting $v_{lk} = v_{kl}$. After each cycle of the Gibbs sampling procedure, we rescale $\mathbf{V}$ by

$$v_{ij}^{(k+1)} \leftarrow \frac{v_{ij}^{(k+1)}}{\sum_{l,m} v_{lm}^{(k+1)}} \tag{53}$$

In brief, we can summarize the above procedure by the algorithm:

Input: $C, V^{(k)}$
Output: $V^{(k+1)}$
for $i = 1, \ldots, n$ **do**
  for $j = 1, \ldots, i$ **do**
    if $c_{ij} + c_{ji} > 0$ **then**
      if $i = j$ **then**
        Sample $v_{ii}^{(k+1)}$ according to (43), (**??**)
      else
        Obtain parameters a, b, c from (51a), (51b), (51c)
        Compute $\bar{v}_{ij}$ according to (50)
        Compute parameters α, β, f_0 using (49a), (49b),(49c)
        Sample v'_{ij} according to (48)
        Compute p_{acc} from (52)
        U from $[0, 1]$ uniformly distributed
        if $U < \min\{1, p_{acc}\}$ **then**
          $v_{ij}^{(k+1)} = v'_{ij}$
        else
          $v_{ij}^{(k+1)} = v_{ij}^{(k)}$
        end
        $v_{ji}^{(j+1)} = v_{ij}^{(k+1)}$
      end
    end
  end
end

Algorithm 1: Reversible sampling algorithm

3. *Updating the observable distribution parameters*

Following the update of the transition matrix $\mathbf{T}$, the observable distribution parameters $\mathbf{E}$ are updated by sampling $\mathbf{E}$ from the

conditional probability $\mathrm{P}(\mathbf{E}' \mid \mathbf{S}', \mathbf{O})$. The conditional probability for the observable distribution parameters for state m, denoted $\mathbf{e}_m$, is given in terms of the output model $\varphi(o \mid \mathbf{e})$ by Bayes' theorem,

$$\mathrm{P}(\mathbf{E} \mid \mathbf{O}, \mathbf{S}) = \left[\prod_{t=0}^{L} \varphi(o_t \mid \mathbf{e}_{s_t})\right] \mathrm{P}(\mathbf{E}). \tag{54}$$

An important choice must be made with regards to the prior, $\mathrm{P}(\mathbf{E})$. If the prior is chosen to be composed of independent priors for each state, as in

$$\mathrm{P}(\mathbf{E}) = \prod_{m=1}^{M} \mathrm{P}(\mathbf{e}_m), \tag{55}$$

then the full BHMM posterior (Eq. 8) will be invariant under any permutation of the states. This behavior might be undesirable, as the states may switch labels during the posterior sampling procedure; this will require any analysis of the models sampled from the posterior to account for the possible permutation symmetry in the states. Here we choose to order the states according to their value in the experimental observable (force or extension), thus removing the ambiguity due to exchangeable state labels. Note that this practice can artificially restrict the confidence intervals of the states when the distributions of the experimental observable overlap strongly in neighboring states

Here, we make the choice that the prior be separable (Eq. 55), which has the benefit of allowing the conditional probability for $\mathbf{E}$ (Eq. 54) to be decomposed into a separate posterior for each state. For each state m, collect all the observations o_t whose updated hidden state labels ${s_t}' = m$ into a single dataset $\mathbf{o} \equiv \{o_n\}_{n=1}^{N_m}$, where N_m is the total number of times state m is visited, for the purposes of this update procedure. Then, the observable parameters $\mathbf{e}$ for this state are given by

$$\mathrm{P}(\mathbf{e} \mid \mathbf{o}) = \mathrm{P}(\mathbf{o} \mid \mathbf{e})\mathrm{P}(\mathbf{e}) = \left[\prod_{n=1}^{N_m} \varphi(o_n \mid \mathbf{e})\right] \mathrm{P}(\mathbf{e}). \tag{56}$$

In the application presented here, we use a Gaussian output model (Eq. 11) for the state observable distributions $\mathrm{P}(o \mid \mathbf{e})$, where $\mathbf{e} \equiv \{\mu, \sigma^2\}$, with μ the state mean observable and σ^2 the variance (which will include both the distribution of the observable characterizing the state and any broadening from measurement noise). Other models (including multidimensional or multimodal observation models) are possible, and require replacing only the observation model $\varphi(o \mid \mathbf{e})$ and corresponding prior $\mathrm{P}(\mathbf{e})$.

We use the (improper) Jeffreys prior [**?**] which has the information-theoretic interpretation as the prior that maximizes the information content of the data [**?**], (suppressing the state index subscript m),

$$\mathrm{P}(\mathbf{e}) \propto \sigma^{-1}, \tag{57}$$

which produces the posterior

$$\mathrm{P}(\mathbf{e} \mid \mathbf{o}) \propto \sigma^{-(N+1)} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(o_n - \mu)^2\right], \tag{58}$$

where we remind the reader that here and in the remainder of this section, the symbols $\mathbf{e}$, $\mathbf{o}$, σ, μ, and N refer to $\mathbf{e}_m$, $\mathbf{o}_m$, σ_m, μ_m, and N_m, respectively.

Updating $\{\mu, \sigma^2\}$ also proceeds by a Gibbs sampling scheme, alternately updating μ and σ, as earlier described in Ref. [**?**],

$$\begin{aligned} \mu &\sim \mathrm{P}(\mu \mid \sigma^2, \mathbf{o}) \\ \sigma^2 &\sim \mathrm{P}(\sigma^2 \mid \mu, \mathbf{o}) \end{aligned} \tag{59}$$

The conditional distribution of the mean μ is then given by

$$\mathrm{P}(\mu \mid \sigma^2, \mathbf{o}) \propto \exp\left[-\frac{1}{2(\sigma^2/N)}(\mu - \hat{\mu})^2\right] \tag{60}$$

where $\hat{\mu}$ is the sample mean for $\mathbf{o}$, the samples in state m,

$$\hat{\mu} \equiv \frac{1}{N}\sum_{n=1}^{N} o_n \tag{61}$$

This allows us to update μ according to

$$\mu' \sim \mathcal{N}(\hat{\mu}, \sigma^2/N) \tag{62}$$

The conditional distribution of the variance σ^2 is given by

$$\mathrm{P}(\sigma^2 \mid \mu, \mathbf{o}) \propto \sigma^{-(N+1)} \exp\left[-\frac{N\hat{\sigma}^2}{2\sigma^2}\right] \tag{63}$$

where the quantity $\hat{\sigma}^2$, which is not in general identical to the sample variance, is given by

$$\hat{\sigma}^2 \equiv \frac{1}{N}\sum_{n=1}^{N}(o_n - \mu)^2. \tag{64}$$

A convenient way to update $\sigma^2 \mid \mu, \mathbf{o}$ is to sample a random variate y from the chi-square distribution with $N-1$ degrees of freedom,

$$y \sim \chi^2(N-1) \tag{65}$$

and then update σ^2 as

$${\sigma'}^2 = N\hat{\sigma}^2/y. \tag{66}$$

Note that μ and σ^2 can be updated in either order, but the updated values of μ or σ^2 must be used in sampling the not-yet-updated σ^2 or μ, and vice-versa.

Other output probabilities, such as mixtures of normal distributions or other distributions, can be substituted by simply changing $\mathrm{P}(\mathbf{E} \mid \mathbf{O}, \mathbf{S})$ and the scheme by which $\mathbf{E}$ is updated.