Population analysis of single-molecule FRET experiments through Empirical Bayes estimation on Hidden Markov Models.

Jan-Willem van de Meent[1,2], Chris H. Wiggins[1], and Ruben L. Gonzalez Jr.[2]

[1]Dept. of Applied Physics and Applied Mathematics, Columbia University, New York, NY
[2]Dept. of Chemistry, Columbia University, New York, NY

ABSTRACT Single-molecule experiments characterize biomolecular processes in terms of the kinetics of transitions between conformational states. Estimation of these kinetic rates often requires analysis of a population of molecules. For each molecule a time-dependent measurement is a noisy representation of the conformational state-space trajectory. Hidden Markov models (HMMs) may be used to infer kinetic trajectories for individual molecules, but because learned parameters vary widely within a population, estimating a consensus kinetic model remains a statistically difficult task. Here we demonstrate how Bayesian analysis of coupled HMMs, which represent each state with a distribution of parameters, can greatly simplify this analysis problem. We present analysis of two single-molecule fluorescence energy transfer (smFRET) studies of the bacterial ribosome, that shows how this coupled analysis can identify a common set of conformational states, characterize the dependence of kinetic equilibria on experimental conditions, and detect kinetically distinct subpopulations within a single experiment. We conclude with a discussion of model selection techniques for determination of the appropriate number of conformational states. The code used to perform this analysis and a graphical user interface are available as open source software.

INTRODUCTION

Owing to a host of technological innovations over the past two decades, fluorescence- and force-based single-molecule techniques are now reaching a level of maturity that makes it possible to perform detailed investigations of some of the cell's most fundamental and complex biochemical reactions (1–5). Many single-molecule experiments seek to characterize the kinetics for biological reactions or pathways by identifying a set of conformational states, the transitions that can occur between them, and estimating the associated kinetic rates. While atomic or near-atomic structural techniques such as X-ray crystallography and cryogenic electron microscopy can provide static snapshots of at least some of the biomolecular conformational states of interest, the strength of single-molecule biophysical approaches is that they enable direct observation of conformational transitions within individual molecules, yielding insights into the mechanisms and regulation of biochemical reactions that cannot be obtained from static approaches alone.

As the field advances, it is increasingly common that single experiments yield data for populations of hundreds or even thousands of molecules, allowing the study of rare transition events and characterization of the heterogeneity of kinetic rates within a population. At the same time, statistical analysis of experimental results is becoming a bottleneck in many studies. To interpret measurements, an experimentalist first has to accurately determine the number of conformational states and infer stochastic conformational transitions from a noisy time series measurement, such as a fluorescence intensity, bead position, or conductance. Calculation of average kinetic rates for a population of molecules typically involves a number of ad-hoc analysis steps, such as defining signal thresholds for each state, and separating time series into 'fluctuating' and 'static' populations. Determining which of these steps are appropriate is a time-consuming process of trial and error that requires both a detailed understanding of the experimental setting and a solid grounding in statistics. As a result it can be difficult for third parties to verify the validity of the resulting analysis. Moreover, learned best practices rarely generalize to new experiments. For these reasons, the development of statistically principled, well-tested techniques that may be used to identify states and estimate kinetic rates in a more automated manner is crucial to further advances in single-molecule approaches to understanding biological mechanisms.

In this paper we demonstrate a statistically robust yet straightforward approach for the analysis of populations of time series that does not require prior knowledge of the characteristics of each state. Our methodology applies Bayesian analysis to hidden Markov models (HMMs), which are one of the tools of choice in machine learning applications that seek to infer state-space trajectories from noisy time series (6–8). In the biophysical community HMM techniques were first introduced in the analysis of patch-clamp experiments in ion channels (9–11), and have since been applied in a variety of single-molecule platforms, including optical trapping (12), magnetic tweezers (13) and single-molecule fluorescence resonance energy transfer (smFRET) experiments

(14–18). In these approaches, a statistical model defines the distribution of measurement values we expect to see based on a set of parameters, such as the centers and widths of Gaussian peaks associated with each state, and the transition probabilities between states. Given this model, statistical inference techniques can determine the most likely set of parameters in light of the measured single-molecule signal. While HMM approaches often work well in the analysis of individual time series, combining information obtained from a population of molecules remains a significant statistical challenge. Currently, analysis of multiple time series often implicitly assumes an 'ergodic' population, where all molecules behave in a stochastic yet statistically indistinguishable manner. Under this assumption, observation of a single molecule over a longer interval is equivalent to observation of several molecules over shorter intervals with the same combined length. An equivalent statement is that a single set of HMM parameters accurately represents data from different molecules. In practice, parameters estimated with HMM analysis vary significantly within a population owing to a combination of physical heterogeneities, thermal noise, and signal processing artifacts. Similar, yet not identical, states must therefore be mapped to a set of 'consensus' states, a process which typically requires some form of binning approach that explicitly postulates intervals in the measurement histogram associated with each state.

We improve upon this current practice by using empirical Bayes (EB) estimation on coupled HMMs (19) to describe conformational states in terms of a distribution on parameters, rather than a single point estimate for their optimal values. By explicitly representing the heterogeneity of a population with a parameter distribution that is learned from the data, this algorithm can characterize states with variable parameters in an experiment-agnostic manner, while eliminating the need for an ergodicity assumption. The result is an an almost fully automated inference procedure that requires only a broad initial guess for the parameters, produces more accurate estimates for the state-space trajectory, does not suffer from statistical biases associated with a given choice of prior, and eliminates the need for ad-hoc post-processing steps. For data that is sufficiently well-described by a given statistical model, EB estimation can accurately determine the number of distinct states and will even leave superfluous states empty when provided (19). Finally, representing each state by a distribution on parameters makes it possible to obtain statistically robust estimates for quantities that derive from the model parameters, such as state life times or free energies, that may be used to compare experiments performed under different conditions in more detail.

To demonstrate the effectiveness of the EB approach, we present analysis of two smFRET studies. Both investigate aspects of translation, the mechanism by which the bacterial ribosome synthesizes the protein that is encoded by a messenger RNA (mRNA) template (see (1) for a review). Each study presents population analysis tasks that are exemplary for single-molecule experiments. The first (20) investigates the mechanisms that regulate the initiation stage of translation, during which a translation-competent ribosomal complex is assembled at the start codon of mRNA template. Specifically, these experiments look at the role of the initiation factor 3 (IF3) in regulating 30S initiation complex assembly and 50S ribosomal subunit joining (see the Results Section for further details). The IF3 molecule accesses at least three distinct conformational states whose occupancy depends on the presence of other initiation factors, the mRNA start codon and the initiator transfer RNA (tRNA) required to start translation. Coherently detecting states across experiments is challenging, particularly when some states are sampled infrequently under certain experimental conditions. We show how coupled analysis on the aggregate data from a range of experiments allows identification of a common set of conformational states, after which analysis results may be separated by experiment to obtain individual estimates for the kinetic rates.

As a second example, we look at experiments (21) that investigate the function of the elongation factor G (EF-G) during translocation, the process through which the ribosome moves along its mRNA template by precisely one codon after incorporation each amino acid into the nascent peptide chain (see the Results Section for details). The concentration of EF-G is varied to control the equilibrium between two 'global' conformational states, denoted GS1 and GS2 respectively, which is shifted towards the GS2 state as the EF-G concentration increases. Because EF-G will not bind to all ribosomal complexes, we expect to see two kinetically distinct subpopulations, corresponding to those ribosomes that have associated with EF-G and those that have not. We show how to extend the EB method to infer the population membership and kinetic rates for each subpopulation.

The common denominator in both analysis tasks is that we seek to use measurements of large populations of molecules to identify a common set of states and determine how kinetic rates differ for subpopulations within this aggregate data. In the case of the IF3 experiment, we have *labeled subpopulations* consisting of sets of time series recorded under identical experimental conditions, and we simply wish to obtain per-experiment estimates of the kinetic rates based on a shared definition of states. In the case of the EF-G study, each experiment contains two *unlabeled subpopulations* and the set of time series associated with each population fraction must be inferred from the data.

The remainder of this paper is organized as follows. We begin with an overview of Bayesian analysis techniques for HMMs and show how EB estimation can be used to infer states with similar yet variable parameters from a population of time series. We demonstrate how EB analysis can be extended to obtain separate estimates of kinetic rates for labeled and unlabeled subpopulations. Finally, we present analysis of the IF3 and EF-G experiments, and conclude with a discussion of the results. A full derivation of our algo-

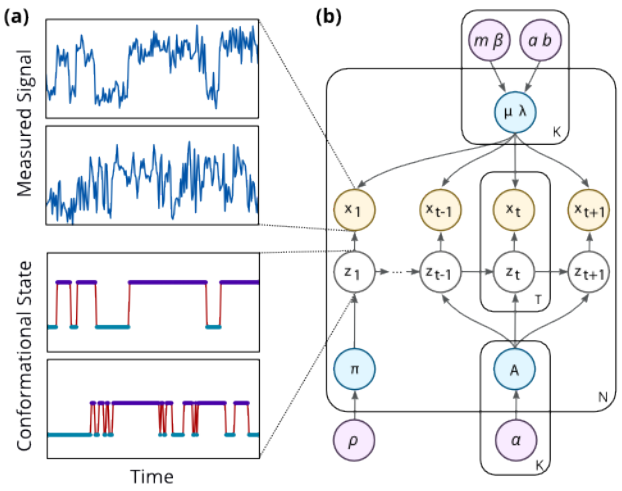

Figure 1: Graphical model for a coupled HMM. **(a)** smFRET signals and sequence of latent states for two time series in an experiment. **(b)** Graphical model showing a coupled HMM for N time series with K states. The parameters $\theta_n = \{\mu_{n,k}, \lambda_{n,k}, A_{n,kl}, \pi_{n,k}\}$ of each time series are distributed to according to a distribution $p(\theta \,|\, \psi)$ with hyperparameters $\psi = \{m_k, \beta_k, a_k, b_k, \alpha_{kl}, \rho_k\}$.

rithm can be found in Sections S1-S4 of the Supplementary Material. The source code used in our analysis, accompanied by an user-friendly graphical interface, is available at http://ebfret.github.io.

METHODS

Bayesian inference in coupled HMMs

Simply stated, Bayesian inference seeks to determine the probability of a set of unknown variables in light of a set of observations. In the context of single-molecule studies, these unknown variables are a set of model parameters θ and a state sequence z_t, whereas the observations are a time series x_t. A graphical model defines a probabilistic relationship between these variables, that can commonly be factored into two terms

$$p(x, z, \theta \,|\, \psi_0) = p(x \,|\, z, \theta)\; p(z, \theta \,|\, \psi_0)\;. \tag{1}$$

Here $p(x \,|\, z, \theta)$ defines the probability of seeing the observed data given known variables, whereas $p(z, \theta \,|\, \psi_0)$ defines a prior probability for these variables in the absence of data, parameterized by some set of hyperparameters ψ_0. Based on these distributions we wish to determine the posterior probability

$$p(z, \theta \,|\, x, \psi_0) = p(x, z, \theta \,|\, \psi_0) \;/\; p(x \,|\, \psi_0)\;. \tag{2}$$

The two distributions $p(x \,|\, z, \theta)$ and $p(z, \theta \,|\, \psi_0)$ describe our assumptions about the data. The first tells us what type of observations we expect to see given the state of the molecule, whereas the second determines the underlying dynamics of the system. Based on these assumptions we now hope to reason about what state sequences and parameter are most probable in light of a set of measurements.

The prior for a HMM can be written as $p(z, \theta \,|\, \psi_0) = p(z \,|\, \theta)\; p(\theta \,|\, \psi_0)$, where the probability $p(z \,|\, \theta)$ depends on two model parameters. The first is a transition matrix A_{kl} that specifies the probability of entering state l from state k at any given time. The second is a set of probabilities π_k that specify the likelihood of starting in state k. The form of the observation distribution $p(x \,|\, z, \theta)$ depends on the type of experimental technique considered. In the case of smFRET experiments the signal for each state k is typically described by a Gaussian peak with center μ_k and width σ_k, or precision $\lambda_k = 1/\sigma_k^2$. The parameters that describe any given time series are therefore $\theta = \{\mu, \lambda, A, \pi\}$. The prior distribution $p(\theta \,|\, \psi_0)$ on the parameters can itself be defined in terms of a set of hyperparameters $\psi_0 = \{m_0, \beta_0, a_0, b_0, \alpha_0, \rho_0\}$ (see supplementary material).

The structure of the probabilistic relationships that define a HMM can be represented as a network, or more precisely a directed acyclic graph (22, 23). In this network variables are the nodes and edges signify dependencies between variables. Such a graphical model for a coupled HMM on N time series with K states is shown in Fig. 1. The dependency structure between variables in this model reflects three fundamental assumptions about the data. The first is that at each time there is a fixed probability of entering into a given state, that depends only on the current conformational state, and has no memory of earlier parts of the state space trajectory. The second is that observations associated with a given conformational state are identically distributed but otherwise uncorrelated in time. The final assumption is that the parameters θ_n of each time series are 'coupled' through a shared prior $p(\theta_n|\psi_0)$, whose distribution reflects the variability of parameter values in an experiment.

The main difficulty in Bayesian inference is that the posterior $p(z, \theta \,|\, \psi_0)$ is typically not analytically or even numerically tractable. Inference strategies for HMMs therefore rely on some form of mathematical approximation. Maximum likelihood (ML) approaches (24, 25) simplify the analysis problem by removing the prior on the parameters $p(\theta \,|\, \psi_0)$and calculating a posterior $p(z \,|\, x, \theta^*)$, along with a point-estimate θ^* that maximizes the likelihood $p(x \,|\, \theta^*)$. Two well-known deficiencies of ML estimation are that it is prone to overfitting and can produce numerical divergences in cases where a state is assigned only a single data point (22). However ML techniques are both computationally efficient and easy to implement, and remain popular in single-molecule applications (14, 17) for this reason.

Variational Bayesian (VB) techniques (15, 16, 18, 23) improve on ML methods by calculating an approximation for the evidence $p(x \,|\, \psi_0)$. One advantage of maximizing the evidence is that it can be used to prevent overfitting by identifying the most appropriate number of states, a problem known in the statistical community as model selection. The like-

lihood, much like in curve fitting procedures, can typically always be increased by using more parameters. However the evidence, which can be thought of as the average likelihood over the range of parameter values encoded by the prior $p(\theta \,|\, \psi_0)$, does not exhibit this behavior, since models that overfit the data will typically only obtain a good fit for a very narrow range of parameter values (22, 26).

In VB estimation, a pair of distributions $q(z)$ and $q(\theta \,|\, \psi)$ approximate the posterior with a factorized form

$$p(z, \theta \,|\, x, \psi_0) \simeq q(z)\, q(\theta \,|\, \psi)\ . \tag{3}$$

Rather than a point-estimate θ^*, this yields a distribution $q(\theta \,|\, \psi)$ defined in terms of a set of posterior parameters ψ. The relationship between ψ and ψ_0 reflects an important principle of Bayesian statistics. The posterior parameters have the same form as the prior parameters, but define a more tightly peaked distribution that reflects our increased knowledge in light of the measurements. More precisely put, ψ can be calculated from a set of 'sufficient statistics' $\mathcal{T}$ (see Section S2 in the Supplementary Material). For a HMM these statistics are given by

$$\gamma_{tk} = E_{q(z)}\big[z_{tk}\big]\ , \qquad \xi_{kl} = \textstyle\sum_t E_{q(z)}\big[z_{(t+1)l} z_{tk}\big]\ , \tag{4}$$

$$\Gamma_k = \textstyle\sum_t \gamma_{tk}, \qquad X_k = \textstyle\sum_t \gamma_{tk} x_t, \qquad U_k = \textstyle\sum_t \gamma_{tk} x_t^2\ . \tag{5}$$

In other words $\mathcal{T} = \{\gamma, \xi, \Gamma, X, U\}$ defines the amount of time spent in each state Γ_k, the number of transitions between states ξ_{kl}, as well as the mean X_k/Γ_k and variance $U_k/\Gamma_k - (X_k/\Gamma_k)^2$ of the observations associated with each state. The posterior parameters can now be calculated from these statistics and the prior parameters (see Section S3.3 of the Supplementary Material for details). For example, the posterior for the transition probabilities $q(A \,|\, \alpha)$

$$\alpha_{kl} = \xi_{kl} + \alpha_{0,kl}\ , \tag{6}$$

is simply the sum of the number of transitions ξ that we believe we have seen in the time series, and the equivalent number of transitions of the prior α_0. In general, placing a prior on the parameters is equivalent to assuming that one has already seen a number of data points with statistics $\mathcal{T}_0$ before seeing the measurements x_t. The number of equivalent observations associated with $\mathcal{T}_0$ determine how quickly the posterior will change in light of new observations.

Empirical Bayes (EB) estimation (19, 27, 28) extends VB estimation to perform coupled inference on populations of time series. To do so we learn N approximate posterior distributions $q(\theta_n \,|\, \psi_n)$ for each time series x_n. The prior $p(\theta \,|\, \psi_0)$ is subsequently chosen by way of a self-consistency requirement; the range of θ_n values predicted by the posterior distributions should match that of the prior. This is equivalent to choosing a set of prior parameters whose distribution is as close as possible to the average posterior (see Section S4 of the Supplementary material)

$$p(\theta \,|\, \psi_0) \simeq \frac{1}{N} \sum_n q(\theta \,|\, \psi_n)\ . \tag{7}$$

In a mathematical sense, this estimation procedure approximates the evidence $\log p(x \,|\, \psi_0)$ with a lower bound L

$$L = \sum_n E_{q(z_n) q(\theta_n \,|\, \psi_n)} \left[\log \frac{p(x_n, z_n, \theta_n \,|\, \psi_0)}{q(z_n) q(\theta_n \,|\, \psi_n)} \right]\ , \tag{8}$$

by iteratively finding solutions to the equations

$$\frac{\delta L}{\delta q(z_n)} = 0\ , \qquad \frac{\delta L}{\delta q(\theta_n \,|\, \psi_n)} = 0\ , \qquad \frac{\partial L}{\partial \psi_0} = 0\ . \tag{9}$$

A full derivation of each of these update steps in this algorithm can be found in Sections S3 and S4 of the Supplementary Material of this paper. The maximization of this lower bound is equivalent to minimizing a 'distance' between distributions, expressed in terms of a quantity known as the Kullback-Leibler (KL) divergence (22). The first two updates obtain a solution to Equation 3 by minimizing the KL divergence between the posterior $p(z_n, \theta_n \,|\, x_n, \psi_0)$ and the variational form $q(\theta_n \,|\, \psi_n)$ and $\psi(\theta_n \,|\, \psi_0)$ (19).

In summary, the empirical Bayes approach to kinetic analysis uses hidden Markov models to calculate two sets of quantities. For each time series we obtain a set of posterior statistics $\mathcal{T}_n$, which report on the occupancy, transitions and measurement values associated with each conformational state. The second quantity is a set of prior parameters ψ_0, which is itself equivalent to a set of prior statistics $\mathcal{T}_0$ that represent the characteristics that are common to all time series in the population.

In cases where we wish to perform inference on multiple time series with similar kinetics, EB estimation has several important advantages over traditional VB and ML approaches. The first is that it provides a straightforward mechanism to determine whether states identified in individual time series derive from the same ensemble state, without requiring post-processing heuristics or binning of similar states. Learning ψ_0 from the data also eliminates subtle and unanticipated sources of bias that may arise from any particular choice of prior. Whereas ψ_0 must be chosen by the user in VB estimation, EB methods provide an almost fully-automated approach that only requires a rough initial guess for ψ_0. Learning the prior also allows more accurate parameter estimates. In VB estimation the prior is typically chosen in such a manner that the statistics $\mathcal{T}_0$ have an equivalent number of observations that is much smaller than the typical number of data points in a time series. The learned prior in EB estimation is more informative in the sense that it encodes a narrower range of parameters. This knowledge of 'typical' parameter values increases the effective number of observations for each posterior estimate, resulting in tighter confidence bounds (19). Finally, learning the prior is also beneficial when determining the correct number of states. Just like VB methods, the EB procedure calculates an estimate for the log evidence, which may be compared for models with different numbers of states. However unlike VB methods, which often implicitly assumes all states

are equally populated, EB infers the occupancy for each state. For sufficiently clean data EB estimation can automatically detect the correct number of states, even when given the wrong number of states to begin with, in the sense that the algorithm will leave superfluous states unpopulated. As we will discuss in the last section of this paper, measurement data often deviates from the idealized form assumed in a statistical model in a number of ways. In practice such systematic discrepancies limit the usefulness of purely statistical criteria for determining the true number of conformational states. However, on simulated data for which the true state sequence is known, EB inference systematically outperforms VB and ML methods (19).

Population Analysis

In our analysis of labeled and unlabeled populations, we will extend the EB estimation procedure in a straightforward manner. Rather than estimate a single set of prior parameters ψ_0 from the posterior statistics $\mathcal{T}_n$, we split our population in into M fractions with prior parameters ψ_{0m}. We introduce a new variable y_{nm} for the population membership of each time series. This variable is simply a binary indicator that is 1 if time series n is part of population m. For labeled populations the values for y are known, and we can estimate distributions for individual populations from the restricted set of posterior distributions

$$p(\theta \mid \psi_{0m}) \simeq \sum_n y_{nm} q(\theta \mid \psi_n) / \sum_n y_{nm} \,. \tag{10}$$

In the case of unlabeled subpopulations, y must be inferred from the data. In order to do so we generalize the EB approach to a mixture of distributions $p(x_n \mid \psi_{0m})$. The evidence can now be expressed as a marginal over all possible y values

$$p(x \mid \psi_0) = \sum_y p(x \mid y, \psi_0) p(y \mid \phi) \,, \tag{11}$$

$$= \sum_n \sum_{y_n} \prod_m p(x \mid \psi_{0m})^{y_{nm}} \phi_m^{y_{nm}} \,. \tag{12}$$

An expectation maximization algorithm over this mixture can be be constructed by introducing a variational posterior $q(y)$ and maximizing the lower bound

$$L = E_{q(z \mid y) q(\theta \mid y) q(y)} [\log p(x, y, z, \theta \mid \psi_0)] \,. \tag{13}$$

We can now estimate the statistic $\omega_{nm} = E_{q(y)}[y_{nm}]$ from the lower bounds $L_{nm} \geq \log p(x_n \mid \psi_{0m})$

$$\omega_{nm} = \frac{\exp(L_{nm}) \phi_m}{\sum_{m'} \exp(L_{nm'}) \phi'_m} \,. \tag{14}$$

In the resulting EB procedure the expectation values with respect to the approximate posteriors are now weighted by the population weights (see Section S4.5 of the Supplementary Material)

$$p(\theta \mid \psi_{0m}) \simeq \sum_n \omega_{nm} q(\theta \mid \psi_{nm}) / \sum_n \omega_{nm} \,. \tag{15}$$

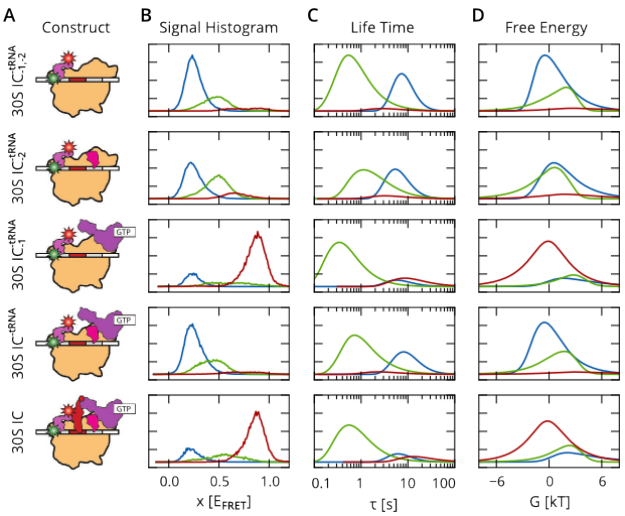

Figure 2: smFRET study of IF3 conformational dynamics on the 30S initiation complex of the bacterial ribosome. (A) Schematic illustrations of experimental contsructs: 30S IC$_{-1,-2}^{-tRNA}$, 30S IC$_{-2}^{-tRNA}$, 30S IC$_{-1}^{-tRNA}$, 30S IC^{-tRNA} and 30S IC^{fMet}. (B) Per-state observation histograms. (C) Life time distributions. (D) Free-energy distributions.

Construct	VB + Binning			EB		
	ext.	*int.*	*cpt.*	*ext.*	*int.*	*cpt.*
30S IC$_{-1,-2}^{-tRNA}$	0.54	0.40	0.06	0.63	0.30	0.07
30S IC$_{-2}^{-tRNA}$	0.52	0.45	0.03	0.47	0.43	0.10
30S IC$_{-1}^{-tRNA}$	0.23	0.11	0.66	0.14	0.15	0.72
30S IC^{-tRNA}	0.56	0.42	0.02	0.60	0.34	0.06
30S IC^{fMet}	0.15	0.17	0.68	0.15	0.21	0.64

Table 1: Relative occupancies of the 'extended', 'intermediate' and 'compact' states of IF3 obtained from VB analysis performed with vbFRET (20) and our EB based analysis of labeled subpopulations.

RESULTS

Labeled subpopulations: The role of IF3 conformational dynamics in regulating translation initiation

During the initiation stage of translation in bacteria the start code of an mRNA template and the anticodon of an fMet-tRNA^{fMet} are positioned into the peptidyl-tRNA binding site of the small, or 30S, ribosomal subunit (29). Once this 30S initiation complex (30S IC) has been formed, the large, or 50S, ribosomal subunit joins onto the 30S IC forming an elongation competent 70S initiation complex (70S IC), trigerring the start of the translation elongation cycle. Because errors in fMet-tRNA^{fMet} or start codon selection can result in mis-translation of an mRNA sequence, the fidelity of initiation is crucial to protein synthesis. Three protein initiation factors, known as IF1, IF2, and IF3, control

the fidelity of initiation by, among other mechanisms, coupling 50S subunit joining to the correct selection of fMet-tRNA$^{\mathrm{fMet}}$ and the start codon. The role of IF3 in this process is to prevnt 50S subunit joining in the event of incorrect tRNA or codon selection.

Here we present analysis of data from experiments that investigate the role of IF3 conformational dynamics in ensuring translation initiation fidelity (20). IF3 is composed of two globular domains connected by a flexible linker. When these domains are labeled with FRET donor and acceptor fluorophores, the ratio of intensities $E_{\mathrm{FRET}} = I_{\mathrm{A}}/(I_{\mathrm{D}} + I_{\mathrm{A}})$ provides a noisy measure of the intramolecular distance between the two domains. Histograms of the E_{FRET} ratio (Fig. 2A) show two dominant peaks, corresponding to a low-FRET 'extended' state, and a high-FRET 'compact' state, whose relative occupancies depend on the presence of the other initiation factors and the fMet-tRNA$^{\mathrm{fMet}}$ on the 30S subunit. In addition to these two states there appear to be one or more 'intermediate' states, which tend to be shorter-lived and have E_{FRET} values that are less well-defined.

Previous analysis was performed with the vbFRET software (15) which performs VB estimation on each individual time series. In this particular set of experiments most time series are 'static', i.e. no conformational transitions are observed before the fluorophores photobleach. This makes it more difficult to distinguish between intermediate and extended or compact states, since there are few transitions that reveal the location of a state relative to others. For this reason the resulting E_{FRET} means of states in each time series were assigned to three empirically-chosen bins with intervals $[0, 0.3)$, $[0.3, 0.7)$ and $[0.7, 1.0)$, where all potential intermediate states were grouped into the middle interval. The compact state was found to be highly populated in a correctly assembled 30S IC , whereas the extended state is highly populated in incorrectly assembled or incomplete 30S ICs, either lacking IFs, containing an incorrect elongator tRNA or an incorrect, near-start codon (20).

In our analysis, we first performed EB inference on the aggregate data from five experiments that were recorded under different conditions: 30S IC$^{-\mathrm{tRNA}}_{-1,-2}$ (lacking IF1, IF2 and tRNA), 30S IC$^{-\mathrm{tRNA}}_{-2}$ (lacking IF2 and tRNA), 30S IC$^{-\mathrm{tRNA}}_{-1}$ (lacking IF1 and tRNA), 30S IC$^{-\mathrm{tRNA}}$ (lacking tRNA) and 30S IC$^{\mathrm{fMet}}$. Three states were used in order to facilitate comparison with previous results.After inference, separate parameter distributions were estimated from the sufficient statistics of each individual experiment in the manner outlined in Equation 10. The results of this analysis are in good agreement with previous results based on explicitly defined bin intervals. Fig. 2 shows observation histograms for each state, as well as distributions of the life time and free energy relative to other states (see Section S5 of the Supplementary Material for a discussion of the calculation of these quantities). The width of each distribution provides us with a confidence interval on each of the parameters. The fractional occupancies obtained for each experiment (Table 1) similarly show a close correspondence to the values obtained with VB analysis. While the results of EB analysis are largely in agreement with previous results, a significant advantage the EB method is that it eliminates the need for post-processing analysis based on manually defined bin intervals, which is both time-consuming and potentially prone to user bias, by providing an automated analysisa that can be performed less than an hour on a single machine without any advance detailed knowledge of the experimental system.

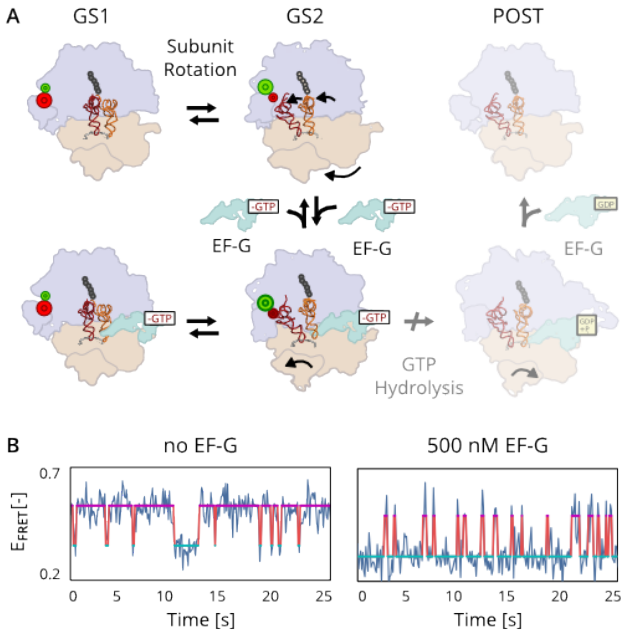

Figure 3: smFRET experiments (21) measuring the influence of EF-G on the GS1-GS2 equilibrium in the bacterial ribosome. (A) The kinetic pathway for translocation is believed to have three steps: A reversible rotation of the two subunits (purple and orange), followed by the binding of EF-G (green) which stabilizes the rotated GS2 state long enough for a GTP-driven transition to the post-translocation (POST) complex, blocked here by substitution of GTP by a non-hydrolyzable analogue. (B) smFRET signals reporting on the GS1-GS2 transition show a shift of the equilibrium towards the GS2 state (magenta) in the presence of EF-G.

Unlabeled subpopulations: The influence of EF-G binding on the GS1-GS2 equilibrium

In this section we address the analysis case of unlabeled populations, where we show that the extended EB estimation procedure described by Equation 15 can identify kinetically distinct subpopulations and estimate the kinetic rates for each population fraction. We present analysis of a series of experiments that investigate the role of EF-G in the mechanism of translocation (Fig. 3A), the movement of the ribosome along its mRNA template by precisely one triplet-nucleotide codon. The process of translocation can be broken up into 3 multi-step mechanistic phases. The first is a

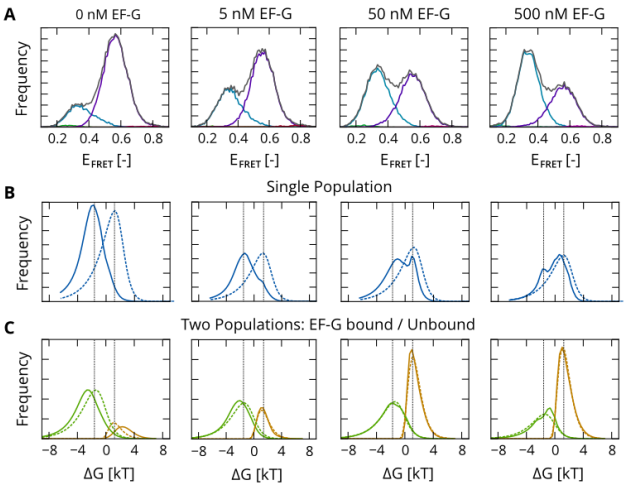

Figure 4: Analysis of GS1-GS2 equilibrium as a function of EF-G concentration. (A) Histogram of aggregate measurements, split by inferred state. (B) EB prior (dashed) and mean posterior (solid) on the free-energy difference $\Delta G = G_{GS1} - G_{GS2}$. A bi-modal signature in the posterior is visible in experiments where EF-G is present. (C) Prior and posterior after unlabeled subpopulation analysis, showing an increasing occupancy of the bound fraction (orange) relative to the non-bound fraction (green) as a function of EF-G concentration.

EF-G	**0 nM**	**5 nM**	**50 nM**	**500 nM**	**1000 nM**
$\rho_{+\mathrm{EF-G}}$	0.13	0.30	0.56	0.65	0.67
$\Delta G_{+\mathrm{EF-G}}$	1.7	1.2	1.3	1.4	1.4
$\Delta G_{-\mathrm{EF-G}}$	-2.4	-1.7	-0.8	-0.4	-0.4

Table 2: EF-G concentration dependence in unlabeled subpopulation analysis of GS1-GS2 equilibrium, showing the bound fraction $\rho_{+\mathrm{EF-G}}$, and the free energy difference ΔG between the GS1 and GS2 state for each subpopulation.

thermally driven, reversible transition between two global conformational states of the of the ribosomal complex (denoted here as GS1 and GS2). This transition is followed by the binding the translational GTPase EF-G, which transiently stabilizes the GS2 state long enough to enable the third step, a GTP hydrolysis-driven movement of the ribosome along its mRNA template. The kinetic equilibrium between the GS1 and GS2 states is studied experimentally by labeling characteristic domains with a fluorophore pair and substituting GTP with a non-hydrolysable analogue (GDPNP), preventing the third and final phase of translocation from taking place.

Fig. 3B shows two time series that exhibit thermally driven, reversible transitions between GS1 and GS2. The first is recorded in the absence of EF-G and shows a preference for the GS1 state. The second time series, taken from an experiment where 500 nM EF-G was added to the imaging buffer, shows a shift of the equilibrium towards the GS2 state. Qualitative comparison of these two time trajectories suggests that EF-G destabilizes the GS1 state and stabilizes the GS2 state in subpopulation of EF-G-bound ribosomal complexes. In order to quantify this difference in kinetic rates and characterize its dependence on EF-G concentration, we must obtain separate estimates for the distribution on kinetic rates for the EF-G bound and unbound subpopulations in an experiment.

EB analysis of a series of experiments performed at increasing EF-G concentrations is shown in Fig. 4. As with the previous experiment we first analyze the aggregate data to identify two states. As can be seen in the observation histograms (Fig. 4A), the occupancy of the GS2 state (magenta) increases with the EF-G concentration. Conventional EB analysis with a single population (Fig. 4B) reveals a bimodal signature in the posterior (solid lines) that reveals the existence of two (unlabeled) subpopulations. This signature is absent from the prior (dashed lines) since EB analysis assumes all transition probabilities are governed by the same prior distribution. Because a limited number of transitions between GS1 and GS2 can be observed before one of the fluorophores photobleaches, it is not possible to obtain a precise estimate of the kinetic rates for each individual molecule. As a result, the two peaks in Fig. 4B have a very high degree of overlap, showing that it would be difficult to determine the population membership for each time series using any form of binning approach. The subpopulation analysis technique described in the previous section (see Section S4.5 of the Supplementary Material) produces two much better resolved peaks (Fig. 4C). This result is a consequence of the fact that EB estimation with a single population produces a broadly peaked prior on the transition probabilities, reflecting the heterogeneity in the dataset. When two components are used, each of the subpopulations is more homogeneous, resulting in more tightly peaked prior and posterior estimates. Table 2 lists the population fraction and free energy difference obtained from EB estimation with unlabeled subpopulations. As should be expected, the relative size of the bound subpopulation increases with the EF-G concentration.

The advantage of the population-level analysis of these experiments is that it allows us to calculate derivative quantities, such as free energy differences, which act as diagnostics that allow us to determine whether we need to a more sophisticated model to describe the data. Our inference algorithm for unlabeled subpopulations is such a more sophisticated approach, that both provides a more principled and less laborious method for separating populations in to kinetically distinct fractions, vastly simplifying the analysis process for these types of data.

Model Selection

An ever present challenge in statistical modeling is the determination of the appropriate number of degrees of freedom, which in the case of HMM analysis is simply the number of

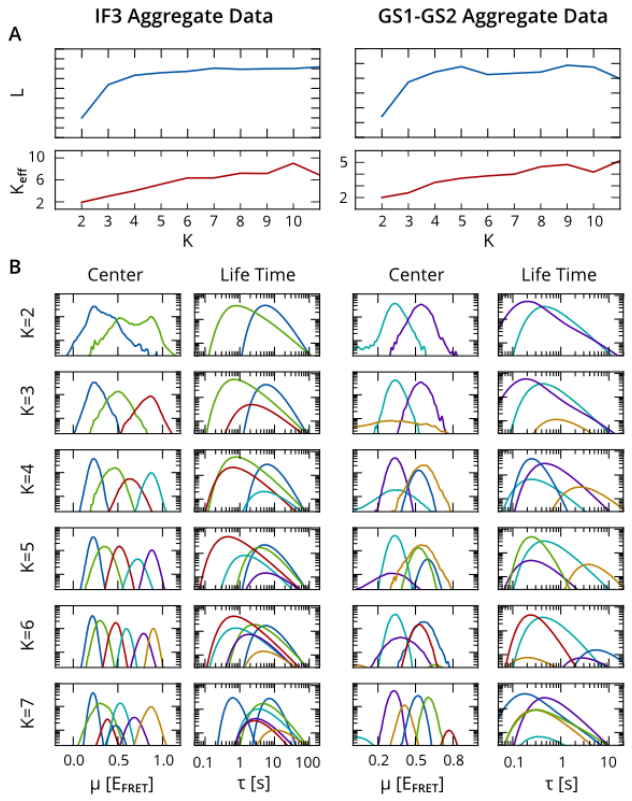

Figure 5: EB analysis of IF3 and GS1-GS2 aggregate data for increasing number of states K. (A) Evidence lower bound L and effective number of populated states K_{eff} as a function of K. (B) Averaged posterior on state centers μ and life times τ.

conformational states. Existing approaches employ model selection criteria such as the lower bound evidence (in VB estimation), or cross-validation and the Bayesian information criterion (for ML methods) to determine the number of states in individual time series (15–18). EB inference improves upon existing methods by obtaining the best solution for the number of states at the population level, which is not possible with ML and VB methods.

As we discussed in the Methods section, both VB and EB calculate a lower-bound estimate for the log evidence L. From previous results on simulated data we know that this quantity depends on the effective number of states K_{eff} populated by the algorithm (19), which can be expressed as $K_{\text{eff}} = \exp[-\sum_k \zeta_k \log \zeta_k]$, where $\zeta_k = \sum_n \Gamma_{nk} / \sum_{nk} \Gamma_{nk}$ is the average normalized state occupancy for the population. Typically there is a range solutions for different K that yield the same (correct) K_{eff} value. At higher K we may observe some over-fitting as a result of imperfect convergence, but this overfitting can be detected through a decrease in L (19). Moreover the EB method more accurately predicts the number of states in individual time series when compared to ML and VB approaches (19). In short, for computer-simulated data, where each variable is distributed precisely as is assumed by the model, EB estimation systematically outperforms existing methods. It obtains more accurate results in individual time series, can perform model selection at the population level, and typically even obtains the correct result when run with more states than are present in the data.

Unfortunately real data is rarely in perfect agreement with a given statistical model. In smFRET experiments for example, we assume a Gaussian distribution of measurement values for each conformational state. All HMM approaches for smFRET data make this same assumption, which is necessary for mathematical tractability of the model. In reality the E_{FRET} ratio exhibits a sigmoidal dependence on the distance between the fluorophores, resulting in peaks that are typically slightly skewed towards the middle of the spectrum. Measurements are further corrupted by a number of artifacts such as intermittent photo-blinking of fluorophores, or incorrect detection of photobleaching, and errors in determination of the background fluorescence intensity. In general, such systematic discrepancies and artifacts will cause any statistical algorithm to populate extra states.

As a consequence of this disagreement between data and model, analysis of experimental data (Fig. 5A) typically shows steady increasing K_{eff} values, which are not matched by a decrease in L. While this may appear to be an instance of 'overfitting' in the sense that EB estimation identifies states that do not correspond to actual physical conformations, it is important to note that this is behavior is very different from normal overfitting. ML approaches obtain a better fit by assigning natural statistical variations to seperate states, and will do so even for simulated data that is in perfect agreement with the hypothesized model. EB analysis generally obtains the correct result on simulated data, but uncovers 'unnatural' variations in experimental data that are 'real' from a statistical point of view, but do not contain useful information about actual conformational transitions.

Examples of these systematic discrepancies can be seen in Fig. 5B, which shows the averaged posterior distribution on the state centers μ_{nk} and state dwell times τ_{nk} obtained by analyzing the aggregate datasets from the previous sections with increasing number of states. When plot on a logarithmic scale, a Gaussian distribution will have a parabolic shape. The curves for μ_{nk} clearly show both asymmetries and aberrant tails that deviate from this idealized form. As a result it can be fairly obvious when too few states are used, such as in the $K = 2$ analysis of the IF3 experiments, but it is generally impossible to say whether too many states are used, since the curves obtained at higher K do show a closer agreement with the shape assumed in the model. So even with this more sophisticated approach, it is generally advisable to use the smallest number of states possible when interpreting results, and exercise caution when attributing conclusions to states observed in less than 5% of the population. Practitioners may also find it useful to note that EB inference becomes easier and more accurate when more is known about the life times of states. In platforms where the data ac-

quisition rate can be controlled, one should therefore ideally pick a measurement interval that yields dwell times on the order of 10 time points.

The practical lesson in this analysis is that model selection criteria are only as accurate as the representation of the measurement data in the model. We emphasize that this limitation is by no means unique to EB analysis. ML and VB approaches use precisely the same Gaussian distribution, for the observations, and suffer from the same defects. It is merely the case that these issues are obfuscated when time series are analyzed individually, since a single time series rarely contains enough data points to make discrepancies relative to model apparent. The advantage of the EB methodology is that it at least puts any disagreement between data and model in plain view so that it may be addressed in a later iteration of model design. In this sense, EB estimation not only improves upon existing approaches through better model selection critera, but also provides diagnostics that tell us whether there is sufficient agreement between data and model for such criteria to be effective.

DISCUSSION

Our results show that empirical Bayes estimation on coupled hidden Markov models both greatly simplifies and enhances kinetic analysis of populations of single-molecule time series. The EB algorithm performs HMM analysis on individual time series and learns a distribution on the model parameters in order to represent the heterogeneity of parameters within a population. A desirable feature of the EB estimation is that it can identify conformational states and learn associated kinetic rates in an almost fully automated manner, requiring only a minimal amount of user input in the form of an initial guess for the hyperparameters. Analysis of IF3 and GS1-GS2 transition data is largely consistent with previous results based on VB inference. The main advantage of this more automated estimation is that it eliminates potential sources of bias arising from manually set prior parameters, post-analysis binning of states, or a separation of time series into static and dynamic fractions.

The population-based analysis also makes it possible to assess the degree of heterogeneity within a population, quantify how kinetic equilibria are affected by experimental conditions and detect kinetically distinct subpopulations within a single experiment. In the case of the IF3 study, consensus states may first be identified by analyzing the aggregate dataset of time series obtained with different complexes, after which kinetic rates for each individual experiment can be obtained almost trivially by re-estimating the prior from the sufficient statistics of each individual dataset. The analysis results for the GS1-GS2 experiment show that that we can extend the EB to perform inference over a mixture of unlabeled subpopulations.

Learning prior and posterior distributions can furthermore provide useful diagnostics that indicate whether a given model is appropriate to the data. Consider, for example, the analysis of the GS1-GS2 experiments. When a prior based on a single population is learned from the aggregate data, the differences between prior and posterior clearly exhibit a trend as a function of the EF-G concentration, and a bi-modal signature of the posterior suggests that the data may contain two kinetically distinct populations. The discrepancies between model and data also prove central to the determination of the correct number of states. Previous approaches have advocated model selection criteria such as the lower bound evidence (in VB estimation), or cross-validation and the BIC (for ML methods). While such criteria are indeed a useful guide, a practical lesson learned from this more sophisticated analysis is that these criteria are only reliable when the statistical model can accurately describe the data, including all artifacts particular to a given choice of experimental platform. In this sense, an important advantage of the EB approach is precisely that it can tease out discrepancies between the data and the chosen statistical model, which inform us how our assumptions about the data need to be adjusted in the next iteration of statistical model design.

Finally, the EB estimation framework is applicable to a wide range of single-molecule techniques. Although we here have analyzed single-molecule FRET experiments exclusively, our approach is by no means restricted to this platform. An adaptation of our algorithm to the analysis of optical trapping and magnetic tweezers experiments is possible with minimal modifications and we have recently collaborated to develop an application to analysis of tethered particle motion experiments (30).

In summary, the EB methodology presented here is general, extensible, and both simplifies and improves upon existing methodologies in every respect. While no idealized statistical model can describe experimental data perfectly, this population-based methodoloy provides estimates that are as accurate as possible given the fidelity of the underlying model, and makes it far easier to identify the discrepancies to be eliminated in the next iteration of model design.

SUPPORTING MATERIAL

A full derivation of the EB estimation algorithm can be found in the Supplementary Material on-line. The source code used in this publication, accompanied by a GUI frontend for interactive analysis, is available at http://ebfret.github.io.

The authors would like to thank Margaret Elvekrog, Kevin Emmett, Jingyi Fei, Jason Hon, Daniel MacDougall and Jordan McKittrick for comments on this manuscript. It is also our pleasure to acknowledge helpful discussions with Jonathan Bronson, Martin Linden, Frank Wood, Matt Hoffman and David Blei. This work was supported by an NSF CAREER Award (MCB 0644262) and an NIH-NIGMS grant (R01 GM084288) to R.L.G.; a NIH National Centers for Biomedical Computing grant (U54CA121852) to

C.H.W.; and a Rubicon fellowship (680-50-1016) from the Netherlands Organization for Scientific Research (NWO) to J.W.M.

Bibliography

1. Tinoco, I., and R. L. Gonzalez, 2011. Biological mechanisms, one molecule at a time. *Genes. Dev.* 25:1205–31.

2. Joo, C., H. Balci, Y. Ishitsuka, C. Buranachai, and T. Ha, 2008. Advances in single-molecule fluorescence methods for molecular biology. *Ann. Rev. Biochem.* 77:51–76.

3. Borgia, A., P. M. Williams, and J. Clarke, 2008. Single-molecule studies of protein folding. *Ann. Rev. Biochem.* 77:101–25.

4. Neuman, K. C., and A. Nagy, 2008. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nat. Methods* 5:491–505.

5. Cornish, P. V., and T. Ha, 2007. A survey of single-molecule techniques in chemical biology. *ACS Chem. Biol.* 2:53–61.

6. Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *P. IEEE* 77:257–286.

7. Eddy, S. R., 1996. Hidden Markov models. *Curr. Opin. Struc. Biol.* 6:361–5.

8. Bilmes, J., 1998. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *Int. Comp. Sci. Inst.* 1198.

9. Chung, S. H., J. B. Moore, L. G. Xia, L. S. Premkumar, and P. W. Gage, 1990. Characterization of single channel currents using digital signal processing techniques based on Hidden Markov Models. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 329:265–85.

10. Qin, F., A. Auerbach, and F. Sachs, 1997. Maximum likelihood estimation of aggregated Markov processes. *Proc. R. Soc. Lond. B Biol. Sci.* 264:375–83.

11. Qin, F., A. Auerbach, and F. Sachs, 2000. A direct optimization approach to hidden Markov modeling for single channel kinetics. *Biophys. J.* 79:1915–1927.

12. Smith, D. A., and R. M. Simmons, 2001. Models of Motor-Assisted Transport of Intracellular Particles. *Biophys. J.* 80:45–68.

13. Kruithof, M., and J. van Noort, 2009. Hidden Markov analysis of nucleosome unwrapping under force. *Biophys. J.* 96:3708–15.

14. McKinney, S. a., C. Joo, and T. Ha, 2006. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* 91:1941–51.

15. Bronson, J. E., J. Fei, J. M. Hofman, R. L. Gonzalez, and C. H. Wiggins, 2009. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* 97:3196–205.

16. Bronson, J. E., J. M. Hofman, J. Fei, R. L. Gonzalez, and C. H. Wiggins, 2010. Graphical models for inferring single molecule dynamics. *BMC Bioinformatics* 11 Suppl 8:S2.

17. Greenfeld, M., D. S. Pavlichin, H. Mabuchi, and D. Herschlag, 2012. Single Molecule Analysis Research Tool (SMART): An Integrated Approach for Analyzing Single Molecule Data. *PloS One* 7:e30024.

18. Okamoto, K., and Y. Sako, 2012. Variational Bayes Analysis of a Photon-Based Hidden Markov Model for Single-Molecule FRET Trajectories. *Biophys. J.* 103:1315–24.

19. van de Meent, J.-W., J. E. Bronson, F. Wood, R. L. Gonzalez, and C. H. Wiggins, 2013. Hierarchically-coupled hidden Markov models for learning kinetic rates from single-molecule data. *Proc. Int. Conf. Mach. Learn.* 28:361–369.

20. Elvekrog, M. M., and R. L. Gonzalez, 2013. Conformational selection of translation initiation factor 3 signals proper substrate selection. *Nat. Struct. Mol. Biol.* 20:628–33.

21. Fei, J., J. E. Bronson, J. M. Hofman, R. L. Srinivas, C. H. Wiggins, and R. L. Gonzalez, 2009. Allosteric collaboration between elongation factor G and the ribosomal L1 stalk directs tRNA movements during translation. *P. Nat. Acad. Sci. USA* 106:15702–7.

22. Bishop, C. M., 2006. Pattern recognition and machine learning. Springer, New York.

23. Jordan, M., Z. Ghahramani, and T. Jaakkola, 1999. An introduction to variational methods for graphical models. *Mach. Learn.* 233:183–233.

24. Dempster, A., N. Laird, and D. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* 39:1–38.

25. Baum, L., T. Petrie, G. Soules, and N. Weiss, 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Ann. Math. Stat.* 41:164–171.

26. Beal, M. J., 2003. Variational Algorithms for Approximate Bayesian Inference. Phd thesis, University College London.

27. Berger, J., 1982. Bayesian Robustness and the Stein Effect. *J. Am. Stat. Assoc.* 77:358–368.

28. Kass, R., and D. Steffey, 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Stat. Assoc.* 84.

29. Laursen, B. S., H. P. Sorensen, K. K. Mortensen, and H. U. Sperling-Petersen, 2005. Initiation of protein synthesis in bacteria. *Microbiol Mol Biol Rev* 69:100–101.

30. Lindén, M., S. Johnson, J.-W. van de Meent, C. H. Wiggins, and R. Phillips, in preparation. Interconversion map of multiple Lac-mediated DNA loops from Bayesian analysis of tethered particle motion.