Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments

Jan-Willem van de Meent[1], Jonathan E. Bronson[2], Chris H. Wiggins[3], and Ruben L. Gonzalez Jr.[2]

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

ABSTRACT Many single-molecule experiments aim to characterize biomolecular processes in terms of kinetic models that specify the rates of transitions between conformational states of the biomolecule. Estimation of these rates often requires analysis of a population of molecules, in which the conformational trajectory of each molecule is represented by a noisy, time-dependent signal trajectory. Although hidden Markov models (HMMs) may be used to infer the conformational trajectories of individual molecules, estimating a consensus kinetic model from the population of inferred conformational trajectories remains a statistically difficult task, as inferred parameters vary widely within a population. Here we demonstrate how a recently-developed empirical Bayesian method for HMMs can be extended to a enable a more automated and statistically principled approach to two widely occurring tasks in the analysis of single-molecule fluorescence resonance energy transfer (smFRET) experiments: (i) the characterization of changes in rates across a series of experiments performed under variable conditions and (ii) the detection of "degenerate" states that exhibit the same FRET efficiency but differ in their rates of transition. We apply this newly developed methodology to two studies of the bacterial ribosome, each exemplary of one of these two analysis tasks. 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 basic graphical user interface front-end are available as open source software.

INTRODUCTION

Owing to a host of technological innovations over the past two decades, single-molecule techniques are now reaching a level of maturity that makes it possible to perform detailed mechanistic investigations of some of the cell's most fundamental and complex biomolecular processes (1–5). A large class of such single-molecule experiments seeks to establish a kinetic model, defined in terms of a set of structural conformations of the molecule (hereafter referred to as 'states') and the rates of transitions between these states. This kinetic model must be inferred from a set of experimental signal versus time trajectories that report on conformational transitions in tens, hundreds, or even thousands of signal trajectories. Unfortunately, however, the analysis of large populations of trajectories presents several challenges that currently impair our ability to accurately infer such kinetic models. Specifically, it remains difficult or impossible to: (i) accurately determine the number of states that are present in each noisy signal trajectory; (ii) rigorously infer a single kinetic model that is consistent with the entire population of signal trajectories; (iii) directly compare kinetic models for populations of trajectories recorded under different experimental conditions; and (iv) confidently detect 'degenerate' states that exhibit the same signal output, but that differ in their transition rates. Overcoming these challenges, therefore, promises to increase the ease, confidence, and accuracy with which kinetic models can be inferred from this class of single-molecule experiments.

The analysis of individual, noisy signal trajectories has been greatly facilitated by the use of hidden Markov models (HMMs) (6–8). In the biophysical community, these methods were introduced within the context of patch-clamp experiments on ion channels (9–11), and have since been applied within a variety of single-molecule experimental platforms, including optical trapping (12), magnetic tweezer (13) and single-molecule fluorescence resonance energy transfer (smFRET) experiments (14–19). In HMM approaches, a statistical model defines an expected distribution of measurement values in terms of a set of parameters, such as the centers and widths of Gaussian peaks representing the signal values associated with each conformational state, and the transition probabilities between states. Given this model, maximum likelihood (ML) techniques (14, 18, 20, 21), such as those employed in the smFRET data analysis software packages HaMMy (14) and SMART (18), can determine the most likely set of parameters and conformational trajectory for each measured signal trajectory. A well-known deficiency of ML methods, however, is that the likelihood can always be improved by adding more states to the ki-

netic model, making it difficult to distinguish real conformational states from states that arise from "overfitting" the inherently noisy individual signal trajectories. Variational Bayesian (VB) techniques (15, 16, 19, 22), such as those employed in the smFRET data analysis software package vbFRET (15, 16), improve upon ML methods by introducing a prior distribution, which specifies the expected range of parameter values, allowing maximization of the "evidence", a likelihood that is averaged over this prior. Unlike the likelihood, the evidence is more likely to peak when the signal trajectory is modeled with the optimal number of states. Thus, VB methods can be used to perform model selection, that is, to determine the number of states that yields the best average agreement between the data and the model (see Methods Section for further background).

While maximization of the evidence has proven an effective model selection strategy, it does not completely eliminate overfitting, and particularly underfitting, of the signal trajectories. For example, single-molecule FRET efficiency (E_{FRET}) trajectories that are particularly noisy (*i.e.* with a standard deviation in the E_{FRET} value that is greater than $\sim$ 0.15) and/or include transitions that are fast relative to the rate of data acquisition (*i.e.* more than 1 transition every 5 time points) are particularly prone to underfitting (15). Moreover, existing ML and VB techniques have an important shortcoming that has significant theoretical and practical implications: they can only model individual signal trajectories, or multiple signal trajectories (17) only if they are modeled with the exact same parameters. For example, it is a common occurrence that the same state gives rise to a signal centered at $E_{\mathrm{FRET}} = 0.30$ in one trajectory and $E_{\mathrm{FRET}} = 0.35$ in another, due to variations in the the photophysical properties of the fluorophores, slight structural differences in the molecule, and offsetting errors in the measured fluorescence intensity. Although it might be trivial for an experimentalist to recognize that the $E_{\mathrm{FRET}} = 0.30$ and $E_{\mathrm{FRET}} = 0.35$ measurements are different manifestations of the same state, the ML and VB techniques described above cannot model this situation. From a theoretical perspective, it is unsatisfying that the existing algorithms cannot account for such a fundamental component of all real experiments that is obvious to the human eye. From a practical perspective, this shortcoming means that, rather than simultaneously modeling a large population of signal trajectories to naturally infer a single kinetic model that is most consistent with the entire population, the experimentalist must instead individually model each trajectory and subsequently perform a significant amount of post-processing to infer and validate the single, consensus kinetic model.

Recently, we have developed an empirical Bayesian (EB) technique (23, 24) that improves upon VB methods by inferring the features of the prior distribution, which in VB methods must be specified by the experimentalist. In EB estimation, the variation in parameter values predicted by the prior is matched to the variation in inferred parameter values over the population of trajectories, enabling a single, consensus kinetic model to be learned from the simultaneous analysis of a large population of signal trajectories (see the Methods section for a more detailed introduction). We have benchmarked this EB technique using computer-simulated data, demonstrating that, relative to both ML and VB methods, it exhibits a greater resistance to both over- and underfitting of signal trajectories, and have provided a basic example showing that this EB technique can be used to analyze experimental E_{FRET} trajectories (25).

In this paper, we use experimental smFRET data reporting on the mechanism of protein synthesis by the bacterial ribosome to demonstrate how our previously developed EB method (25) can be extended to perform two very frequently encountered smFRET data analysis tasks: (i) the comparison of the number of states, their occupancy, and associated transition rates, across experiments recorded for the same biomolecular system, but under different experimental conditions (e.g. in the absence, presence, and/or varying concentrations of a particular buffer or biomolecular component), and (ii) the detection of states that exhibit the same E_{FRET} value, but that have different transition rates. Currently, most experimentalists treat these problems by performing inference on the individual trajectories, deciding how many states they believe are in the data via a separate assessment (e.g., via a transition density plot (14) or similar (26) metric) and then binning the inference results in an ad hoc post-processing step. This process is time consuming, may be prone to user bias, and lacks metrics for assessing the accuracy of the outcomes. The two extensions of EB estimation presented here, in contrast, allow users to quickly perform analysis in a more automated, statistically rigorous, and reproducible manner, greatly reducing the potential for user bias.

Collectively, the results of these analyses highlight the considerable advantages of EB methods over ML and VB methods and demonstrate how the simultaneous analysis of large populations of signal trajectories using EB methods uniquely enables: (i) automated identification of a common set of states across various experimental conditions; (ii) detection of small, but statistically significant, differences in a single state across different experimental conditions; (iii) characterization of the dependence of the thermodynamic and kinetic properties of states on experimental conditions; and (iv) identification of kinetically distinct subpopulations within a single experiment.

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 observed data. In the context of single-molecule studies, these unknown variables are a set of model parameters

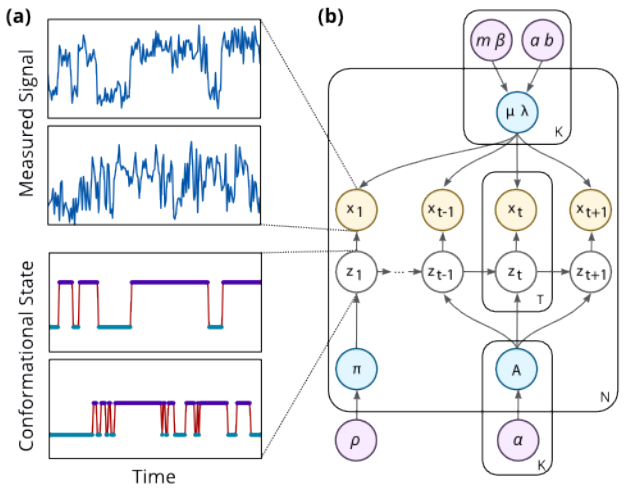

Figure 1: Graphical model for the coupled Bayesian HMM used in EB and VB methods. **(a)** smFRET signals and sequence of latent states for two trajectories in an experiment. **(b)** Graphical model showing a HMM for N trajectories with K states. The parameters $\theta_n = \{\mu_{n,k}, \lambda_{n,k}, A_{n,kl}, \pi_{n,k}\}$ of each trajectory are distributed to according to a distribution $p(\theta\,|\,\psi)$ with hyperparameters $\psi = \{m_k, \beta_k, a_k, b_k, \alpha_{kl}, \rho_k\}$. ML methods use a non-Bayesian variant of this HMM, which omits the hyperparameters ψ.

θ and a state sequence z_t, whereas the observations are a signal trajectory x_t. A graphical model defines a statistical relationship between these variables that can commonly be factored into two terms

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

The two distributions $p(x\,|\,z, \theta)$ and $p(z, \theta\,|\,\psi_0)$, known as the likelihood and prior respectively, describe our assumptions about the model. The likelihood describes the measurement signal we expect to see given the state trajectory z_t of the molecule and a set of emission model parameters, that describe the distribution of measurement values associated with each state. The prior encodes our expectations about the transition probabilities and emission model parameters. Based on these assumptions, the goal of Bayesian inference is now to reason about the so-called posterior probability of the state trajectory (z_t) and model parameters (θ) in in light of a set of measurements (x_t). Bayes' rule states that this posterior probability $p(z, \theta\,|\,x, \psi)$ can be expressed as

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

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 likelihood $p(x\,|\,z, \theta)$ depends on the type of experimental technique considered. In the case of smFRET experiments, a common approach (14–16, 18, 25) is to model the signal for each state k as 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 trajectory 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 as a directed acyclic graph (22, 27). In this network, the nodes are individual variables and edges signify dependencies. Such a graphical model for a coupled HMM on N trajectories with K states is shown in Fig. 1. The dependency structure between variables in this model reflects three fundamental assumptions about the data: (1) at each time, there is a fixed probability of entering into a given state, that depends only on the current state, and has no memory of earlier parts of the state trajectory; (2) observations associated with a given state are independent and identically distributed.; and (3) the parameters θ_n of each trajectory 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 $p(z, \theta\,|\,x, \psi_0)$ can typically not be calculated directly. This is because the normalizing term $p(x\,|\,\psi_0)$ in Eq. 2, known as the evidence, involves an intractable integral. In the EB approach used here we approximate the evidence $p(x\,|\,\psi)$ with the same techniques as those employed in VB estimation: we use a pair of distributions $q(z)$ and $q(\theta\,|\,\psi)$ to approximate the posterior with a factorized form

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

Whereas ML methods obtain a point estimate for the optimal parameters θ, this approach 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

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

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

The statistics $\mathcal{T} = \{\gamma, \xi, \Gamma, X, U\}$ summarize the information contained in each trajectory in terms of the amount of time spent in each state Γ_k, the number of transitions between states ξ_{kl}, the mean X_k/Γ_k measurement value for each state, and its variance $U_k/\Gamma_k - (X_k/\Gamma_k)^2$.

The posterior parameters can be calculated directly from the sufficient statistics and the prior parameters (see Section S3.3 of the Supplementary Material for details). For example, the posterior for the transition probabilities $q(A \mid \alpha)$

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

is simply the sum of the number of transitions ξ that we believe we have seen in the trajectory, 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 (23–25) extends VB estimation to perform simultaneous inference on populations of trajectories. To do so, we learn N approximate posterior distributions $q(\theta_n \mid \psi_n)$ for each trajectory x_n. The prior $p(\theta \mid \psi_0)$ 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). In a mathematical sense, this estimation procedure approximates the log evidence $\log p(x \mid \psi_0)$ with a lower bound L

$$L = \sum_n E_{q(z_n)q(\theta_n \mid \psi_n)} \left[\log \frac{p(x_n, z_n, \theta_n \mid \psi_0)}{q(z_n) q(\theta_n \mid \psi_n)} \right] , \tag{7}$$

by iteratively finding solutions to the equations

$$\frac{\delta L}{\delta q(z_n)} = 0 \,, \qquad \frac{\delta L}{\delta q(\theta_n \mid \psi_n)} = 0 \,, \qquad \frac{\partial L}{\partial \psi_0} = 0 \,. \tag{8}$$

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.

In summary, the EB approach to kinetic analysis uses hidden Markov models to calculate two sets of quantities. For each trajectory we obtain a set of trajectory statistics $\mathcal{T}_n$, which report on the occupancy, transitions and measurement values associated with each state. The second quantity is a set of prior parameters $\psi_0 = \psi(\mathcal{T}_0)$, which represent the characteristics that all signal trajectories have in common. Finally, a set of posterior parameters $\psi_n = \psi(\mathcal{T}_n + \mathcal{T}_0)$ encodes what we know about the parameters of individual trajectories in light of the measured signal. Note that the prior parameters ψ_0 that can be equivalently defined in terms of a set of prior statistics $\mathcal{T}_0$, whereas the posterior statistics are simply the sum of the prior statistics and the trajectory statistics.

We reiterate that EB estimation differs from VB estimation only in the fact that the hyperparameters ψ_0 are not chosen by the user, and held fixed, but are set to the values that maximize the evidence as part of the inference procedure. This allows for more accurate inference, as knowledge of "typical" parameter values results in better estimates of $\mathcal{T}_n$. Moreover, since the learned EB prior is typically less broadly peaked than the postulated prior in VB methods, the effective number of observations for each posterior is larger, resulting in tighter confidence bounds on parameter estimates for individual trajectories (25). Indeed, past analysis of simulated data, for which the true state sequence is known, has shown that EB inference systematically outperforms VB and ML methods, both in terms of parameter estimation and in model selection tasks (25).

Analysis of labeled and unlabeled subpopulations of signal trajectories

In this section we extend of the EB method to perform commonly occurring advanced analysis tasks, which we illustrate in the next sections using two experimental smFRET studies that each 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). The goal of analysis in the first example is to coherently detect the set of states that can be sampled across experiments performed in the presence and absence of other biomolecular components, and subsequently separately estimate the transition rates for each experiment. In the second example, our goal is to extend the EB method to detect subpopulations of trajectories that sample the same two states, but do so with different transition rates.

The common denominator in both these analysis tasks is that we seek to use measurements of large populations of trajectories to identify a common set of states and determine how transition rates differ for subpopulations of molecules within this aggregate data. In the case of the first set of experiments, we have 'labeled' subpopulations consisting of sets of signal trajectories recorded under identical experimental conditions, and we simply wish to obtain per-experiment estimates of the transition rates based on a shared definition of states. In the case of the second study, each experiment contains two 'unlabeled' subpopulations and the set of signal trajectories associated with each subpopulation must be inferred from the data.

To allow more straightforward analysis of labeled and unlabeled subpopulations, we will extend the EB estimation procedure in the following manner. Rather than estimate a single set of prior parameters ψ_0 from the trajectory 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 signal trajectory. This variable is simply a binary indicator that is 1 if trajectory 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

$$p(\theta \mid \psi_{0m}) \simeq \sum_n y_{nm} q(\theta \mid \psi_n) / \sum_n y_{nm} \,. \tag{9}$$

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 $p(x_n \mid \psi_{0m})$, where we assume a discrete prior $p(y \mid \phi)$ on the subpopulation membership. The evidence can now be expressed as a marginal over all possible y values

$$p(x \mid \psi_0) = \sum_y p(x \mid y, \psi_0) p(y \mid \phi) \,, \tag{10}$$

$$= \sum_n \sum_{y_n} \prod_m p(x \mid \psi_{0m})^{y_{nm}} \phi_m^{y_{nm}} \,. \tag{11}$$

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

$$L = E_{q(z \mid y) q(\theta \mid y) q(y)} [\log p(x, y, z, \theta \mid \psi_0)] \,. \tag{12}$$

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

$$\omega_{nm} = \frac{\exp(L_{nm}) \phi_m}{\sum_{m'} \exp(L_{nm'}) \phi'_m} \,. \tag{13}$$

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)

$$p(\theta \mid \psi_{0m}) \simeq \sum_n \omega_{nm} q(\theta \mid \psi_{nm}) / \sum_n \omega_{nm} \,. \tag{14}$$

Software Implementation

All analysis algorithms are implemented in MATLAB, with essential inner components (*i.e.* the forward-backward and viterbi algorithms) written in C as MEX files. Our implementation uses multiple processors when available. We performed a simple benchmark in Matlab 2013a on a Macbook equipped with a 4-core 2.3GHz Core i7 processor, using a computer-simulated dataset with $N = 350$ trajectories of average length $T = 112$. Analysis with 2-6 states required 240 s using 8 nodes and 600 s using a single node. In comparison, our previously released vbFRET software (15) required 1500 s to analyze the same dataset on the same machine.

A line-by-line derivation of the implemented EB estimation algorithm and its extensions can be found in the Supplementary Material on-line. A command-line version of the source code used in this publication, along with a GUI frontend for basic EB estimation tasks, are available at http://ebfret.github.io. This software supports a new single-modecule data (SMD) format that has been designed in collaboration with the Herschlag group at Stanford to enable exchange of data and analysis results between research groups (28).

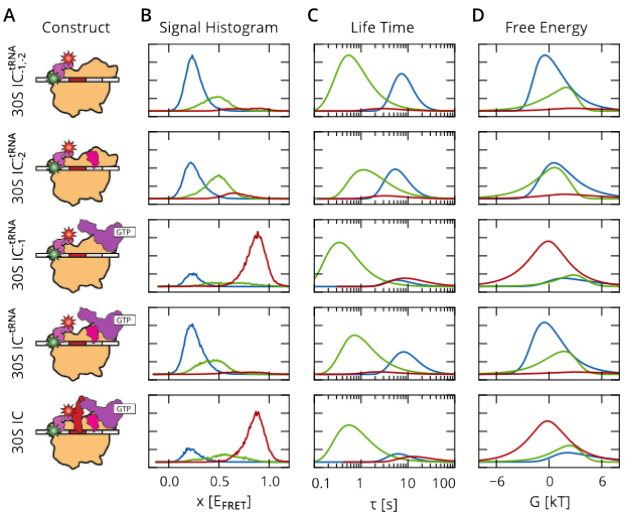

Figure 2: smFRET study of IF3 conformational dynamics on the 30S initiation complex of the bacterial ribosome. (A) Schematic illustrations of experimental contsructs: 30S IC$_{-1,-2}^{-tRNA}$, 30S IC$_{-2}^{-tRNA}$, 30S IC$_{-1}^{-tRNA}$, 30S IC^{-tRNA} and 30S IC^{fMet}. (B) Per-state observation histograms. (C) Life time distributions. (D) Free-energy distributions. The colors blue, green and red are used to label states 1, 2 and 3 respectively in each plot.

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 (29) and our EB based analysis of labeled subpopulations.

RESULTS

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

We begin by showing how the extended EB estimation procedure described by Equation 9 can be used to characterize the dependence of conformational state occupancies, emission model parameters and transition probabilities on experimental conditions. We do so by analyzing a set of previously published smFRET (29) experiments that investigate the role of initiation factor (IF) 3 in regulating the fidelity with which the bacterial ribosome initiates translation at the triplet-nucleotide start codon of the mRNA to be translated.

During bacterial translation initiation, the small, or 30S, ribosomal subunit, IF1, IF2, IF3, a specialized formylme-

thionyl initiator transfer RNA (fMet-tRNA$^{\mathrm{fMet}}$), and the mRNA to be translated form a 30S initiation complex (30S IC) in which the triplet-nucleotide anticodon of fMet-tRNA$^{\mathrm{fMet}}$ is base-paired to the mRNA start codon within the peptdiyl-tRNA binding (P) site of the 30S subunit (30). Subsequent joining of the large, or 50S, ribosomal subunit to the 30S IC results in the formation of a translation elongation-competent 70S initiation complex (70S IC). Because errors in fMet-tRNA$^{\mathrm{fMet}}$ or start codon selection can result in mistranslation of the mRNA sequence, regulating the fidelity of initiation is crucial to protein synthesis and cellular fitness. Thus, the major role of IF1, IF2, and IF3 during translation initiation is to control the fidelity of this process by, among other mechanisms, coupling the 50S subunit joining step of the initiation process to the correct selection of fMet-tRNA$^{\mathrm{fMet}}$ and the start codon; the role of IF3 in this mechanism is to prevent 50S subunit joining until fMet-tRNA$^{\mathrm{fMet}}$ and the start codon have been correctly selected into the P site.

Here we present analysis of smFRET experiments investigating the role that IF3 conformational dynamics play in coupling correct fMet-tRNA$^{\mathrm{fMet}}$ and start codon selection to 50S subunit joining (29). IF3 is composed of two globular domains connected by a flexible linker. When these domains are labeled with FRET donor and acceptor fluorophores, the value of $E_{\mathrm{FRET}} = I_{\mathrm{A}}/(I_{\mathrm{D}} + I_{\mathrm{A}})$, where I_{A} and I_{D} are the emission intensities of the acceptor and donor fluorophores, respectively) provides a noisy measure of the intramolecular distance between the two domains. Histograms of the observed E_{FRET} values (Fig. 2A) show two dominant peaks, corresponding to a low-FRET 'extended' conformational state, and a high-FRET 'compact' conformational state of 30S IC-bound IF3, whose relative occupancies depend on the presence of the other IFs and fMet-tRNA$^{\mathrm{fMet}}$ on the IC. In addition to these two states, there appear to be one or more 'intermediate' conformational 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) that obtains VB estimates for each individual E_{FRET} trajectory. In this particular set of experiments, most trajectories 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, within individual trajectories, 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 trajectory 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, that either lack IFs, lack fMet-tRNA$^{\mathrm{fMet}}$, contain an incorrect elongator aminoacyl-tRNA; or contain an incorrect near-start codon (29).

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.

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$^{\mathrm{fMet}}$ (a correctly assembled 30S IC). This aggregate dataset contained 4233 trajectories with $4.0 \cdot 10^5$ total data points. Three states were used in order to facilitate comparison with the previous results based on VB analysis. After inference, separate parameter distributions were estimated from the sufficient statistics of each individual experiment as described in Equation 9. The results of this analysis, which does not require that the user manually assign the E_{FRET} means of states in each trajectory to empirically chosen bins, are in excellent 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 of each state relative to the other states (see Section S5 of the Supplementary Material for the definitions 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 the VB-based results.

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

We now demonstrate that the extended EB estimation procedure described by Equation 14 can be used to identify kinetically distinct subpopulations of states and estimate the transition rates for each subpopulation of states. As an example of this use case, we perform analysis of a set of smFRET experiments investigating the role of elongation factor (EF) G, a member of the guanosine triphosphatase (GTPase) family of translation factors, during translation elongation.

After the addition of each amino acid to the nascent polypeptide chain during translation elongation, EF-G binds to the ribosomal pre-tranlsocaiton (PRE) complex and hydrolyzes one molecule of guanosine triphosphate (GTP) as it promotes the movement of the ribosome along the mRNA by precisely one triplet-nucleotide codon, a process termed

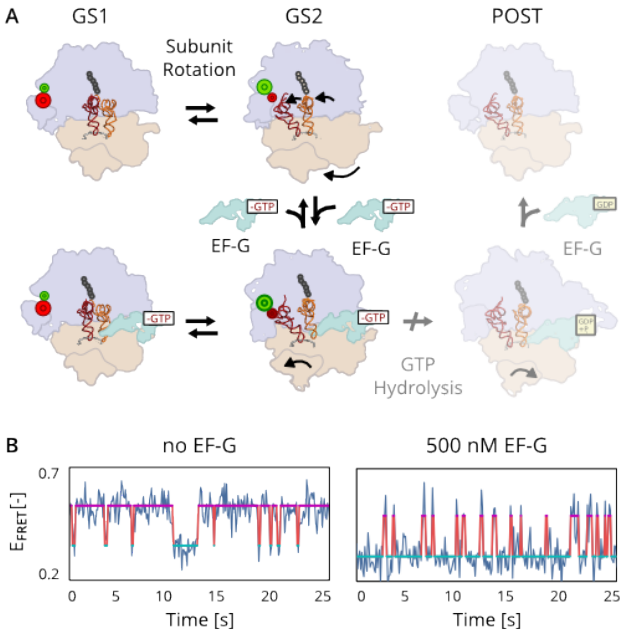

Figure 3: smFRET experiments (31) measuring the influence of EF-G on the GS1-GS2 equilibrium in the bacterial ribosome. (A) The kinetic pathway for translocation is believed to have three steps: A reversible rotation of the two subunits (purple and orange), followed by the binding of EF-G (green) which 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.

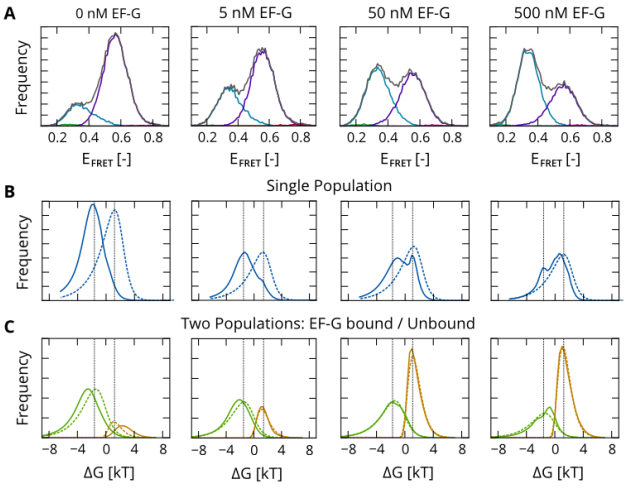

Figure 4: Analysis of GS1-GS2 equilibrium as a function of EF-G concentration. (A) Histogram of aggregate measurements, split by inferred state. (B) EB prior (dashed) and mean posterior (solid) on the free-energy difference $\Delta G = G_{GS1} - G_{GS2}$. A bi-modal signature in the posterior is visible 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.

translocation (Fig. 3A). The overall process of translocation can be broken up into three, smaller, multi-step processes. The first of these is a thermally driven, reversible transition between two global states (denoted as GS1 and GS2) of the ribosomal pre-translocation (PRE) complex. The overall process of translocation can be broken up into three, smaller, multi-step processes. This conformational transition is followed by binding of EF-G to the PRE complex, resulting in a transient stabilization of the GS2 state of the PRE complex that is long enough to enable the third step, a GTP hydrolysis-driven movement of the ribosome along the mRNA. The effect that binding of EF-G has on the dynamic equilibrium between the GS1 and GS2 states of the PRE complex can be studied using smFRET by labeling two ribosomal structural elements with a FRET donor-acceptor pair and substituting GTP with a non-hydrolyzable analogue (GDPNP) that prevents GTP hydrolysis and the associated movement of the ribosome along its mRNA template.

Fig. 3B shows two E_{FRET} trajectories that exhibit thermally driven, reversible transitions between GS1 and GS2. The first trajectory is from an experiment that was recorded in the absence of EF-G and shows a preference for the GS1 state. The second trajectory, from an experiment that was recorded in the presence of 500 nM EF-G and 1 mM GDPNP, shows a dramatic shift of the equilibrium towards the GS2 state. Qualitative comparison of these two trajectories suggests that EF-G destabilizes the GS1 state and stabilizes the GS2 state in the subpopulation of EF-G-bound PRE complexes. In order to quantify this difference in transition rates and characterize its dependence on EF-G concentration, we must obtain separate estimates for the distribution on transition rates for the EF-G-free and EF-G-bound subpopulations of PRE complexes 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. The aggregate data for 7 different EF-G concentrations contained 2472 trajectories with $2.3 \cdot 10^5$ total data points. 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) naturally reveals a bimodal signature in the posterior (solid lines) that hints at 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 very limited number of transitions between GS1 and GS2 can be observed in any one signal trajectory before one of the fluorophores photobleaches, it is not possible to obtain a precise estimate of the transition rates for each individual PRE complex. 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 signal trajectory using any form of binning approach. This ambiguity of subpopulation membership is

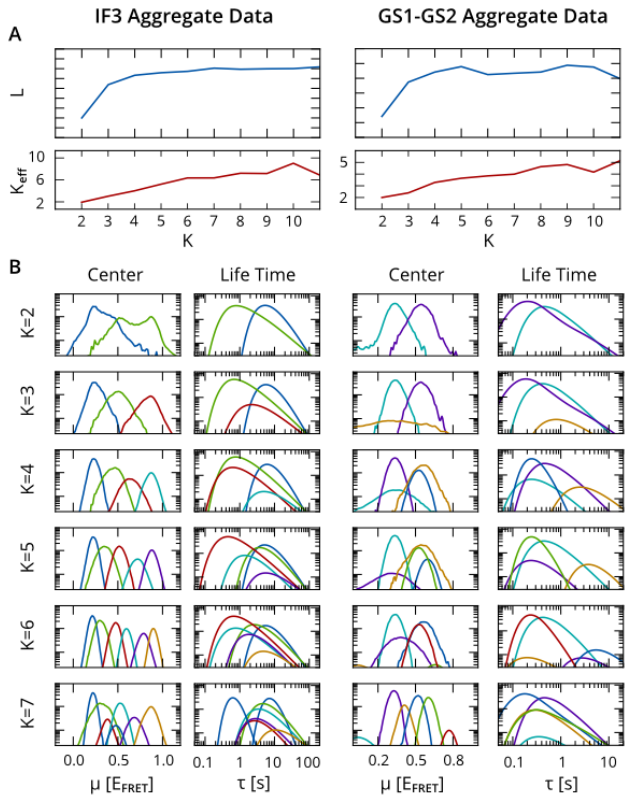

Figure 5: EB analysis of IF3 and GS1-GS2 aggregate data for increasing number of states K. (A) Evidence lower bound L and effective number of populated states K_{eff} as a function of K. (B) Averaged posterior on state centers μ and life times τ.

greatly reduced when using the subpopulation analysis technique described in the previous section (see also Section S4.5 of the Supplementary Material), which produces two much better resolved peaks (Fig. 4C). 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 EF-G-bound subpopulation increases as the concentration of EF-G increases.

Model Selection

One of the stated advantages of VB and EB methods is that they optimize a lower bound for the log evidence, a quantity that may be used to decide among analysis results with different numbers of states. Previous benchmarks using computer-simulated data have shown that EB estimation systematically outperforms VB and ML methods in model selection tasks (25). Not only does EB estimation more accurately determine the number of states in individual trajectories, preventing both under- and overfitting, but the method can also determine the correct number of states starting from a larger number of candidate states, leaving superfluous states unpopulated.

In practice, experimental data differ from simulated data in that they are never in precise agreement with a given statistical model. In smFRET experiments, for example, we assume a Gaussian distribution of E_{FRET} values for each state. With one exception (17), all HMM approaches for analysis of (time-binned) smFRET data make this same assumption (14–16, 18). In reality, however, the E_{FRET} value exhibits a sigmoidal dependence on the distance between the fluorophores, resulting in a distribution of E_{FRET} values that is skewed towards the middle of the spectrum and exhibits a subtle, but systematic, deviation from the idealized Gaussian shape assumed in the model. Distributions of E_{FRET} values further show heavy tails, that likely arise from artifacts such as intermittent photo-blinking of fluorophores (32), incorrect detection of the photobleaching transition, and errors in determining the background fluorescence intensity of individual trajectories.

In general, systematic discrepancies and artifacts can cause a statistical algorithm to "correct" for the fact that observed measurement values are not precisely distributed according to the assumed model by populating additional states, as was found to be the case in our initial analysis of experimental data (25). In Figure. 5 we revisit this notion by examining results obtained by estimating models with 2-10 states on the same two datasets that were analyzed in the previous sections. As in previous work (25), we calculate an effective number of states $K_{\text{eff}} = \exp[-\sum_{k=1}^{K} \zeta_k \log \zeta_k]$ in terms of $\zeta_k = \sum_n \Gamma_{nk} / \sum_{nk} \Gamma_{nk}$, the fraction of time points assigned to each state. When performing analysis on simulated data there is typically a range solutions for different K that yield the same (correct) K_{eff} value and leave any additional states empty (25). Consistent with our previous study (25), the results in Fig. 5A show that K_{eff} steadily increases with the number of candidate states, and it is not clear that there is an optimum K_{eff} value beyond which the lower bound L decreases. In other words, the "fit" of experimental data to the model can be improved by adding incremental, low-occupancy states that capture outliers in the data, even when using model selection criteria. This is undesirable behavior, as such outlier states are more likely to be indicative of measurement artifacts than actual conformational states of interest. However, it is important to note that this behavior is different from the typical overfitting that is associated with ML estimation. ML methods obtain a better fit by assigning natural statistical variations to separate 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 plotted 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 is generally difficult 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.

For this reason, we suggest that users do not indiscriminately rely on the lower bound for model selection; thus, some prudent decision-making with regard to model selection may still be required on the part of the experimentalist. One rule of thumb is to treat states observed in less than 5% of the trajectories with some caution. Additional states may simply: (i) capture artifacts, such as intermediate points between a transition (15), (ii) split a single state into a short-lived and long-lived variant (which may mean that a subpopulation as described in Section is necessary), or (iii) isolate the non-Gaussian tails of actual states. Moreover, any decreases in the lower bound indicate that the method has converged to a local maximum, rather than the globally optimal result, since adding an empty state to the previous result should result in the same, larger, L value. In this case, the user may either opt to perform additional restarts with random initializations of ψ_0, to make it more likely that the global optimum is found for each number of candidate states, or accept the point where L begins to decrease as a bound on the number of states that can be confidently inferred, given computational limitations. As an example, the GS1/GS2 experiment shows a decrease in L at $K = 6$, whereas the life time plot for the blue state falls off the scale at $K = 5$, suggesting that $K = 4$ is the largest number of states that is credible. Also note that these 4 states form two pairs with similar E_{FRET} values but different life times, which is consistent with our knowledge that this experiment in fact does contain kinetically distinct subpopulations. Finally, we note that the conformational trajectory can be inferred with more confidence when more transitions are observed, as it allows the inference procedure to more confidently situate one state relative to others. In cases such as the IF3/30S IC experiment, where the majority of trajectories do not exhibit transitions, analysis results could be improved by shuttering the excitation source to, optimally, obtain a state life time of order 10 time points.

In summary, while EB methods provide model selection criteria that are superior to those employed in ML and VB estimation (when applied to computer-simulated data), a methodological caveat in any statistical 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 typically use precisely the same Gaussian distribution of measurement values and suffer from the same defects. It is merely the case that these issues are obfuscated when signal trajectories are analyzed individually, since an individual signal trajectory rarely contains enough data points to make discrepancies between the data and the model apparent, and the experimentalist makes a judgement call as to how many conformational states they think are required as part of the data inference post-processing. The advantage of the EB methodology is that analyzing all trajectories at once allows us to identify systematic deviations between data and model, allowing us to assess whether there is sufficient agreement between the data and the model for model selection criteria to be effective.

DISCUSSION

While HMMs have proven an immensely popular and effective tool for inferring states and transition rates from individual signal trajectories, combining results from the analysis of multiple trajectories has remained a difficult task. Typically, users manually specify a set of bin intervals, as was done in our previous, VB-based analysis of the IF3 data (29), that allow states identified in individual signal trajectories to be clustered according to their inferred parameter values. In contrast, the EB method uniquely enables simultaneous inference on multiple signal trajectories in a statistically robust manner that eliminates the need for user-defined bin intervals, and is consequently less prone to user bias.

By exploiting the advantages of simultaneously analyzing multiple E_{FRET} trajectories using the EB method, we have developed estimation procedures that uniquely enable us to automate widely encountered tasks in the analysis of smFRET experiments. The first of these tasks is exemplified by our analysis of the IF3 experiments, which demonstrates how E_{FRET} trajectories from a large number of experiments recorded under different experimental conditions can first be simultaneously analyzed to identify a common set of states and then be subsequently reanalyzed to calculate a separate prior distribution for each experiment, allowing characterization of how the state occupancies and transition rates vary between experiments. The second task is exemplified by our analysis of the GS1/GS2 experiments, which demonstrates how the simultaneous analysis of an entire population of E_{FRET} trajectories can be used to automatically identify and characterize subpopulations of molecules occupying functionally and/or conformationally distinct states that exhibit similar E_{FRET} values but that differ in the rates of transitions between states.

For each set of experiments, the results of the EB-based analysis are largely consistent with previous results based on VB methods. However, while the previous VB-based analysis required the use of experiment-specific post-processing procedures that are time consuming to implement, subject to user bias, and difficult to validate, our EB method can be used to obtain results rapidly and with little to no manual intervention by the user. Moreover, the EB approach optimizes a well-defined, statistical, model-selection criterion, the lower bound for the log evidence, which in principle can

be used to compare and decide among different analyses of the same data.

Our EB-based analysis of smFRET data also demonstrates that comparing the prior and posterior distributions can often provide useful qualitative diagnostics that indicate whether a given model is appropriate for the data. In the case of the GS1/GS2 experiments, for example, we are able to calculate a posterior distribution on the free energy difference between states that reveals a systematic mismatch between the single population of PRE complexes that is assumed in conventional EB analysis and the two subpopulations of PRE complexes that are actually present in the experiment (i.e., EF-G-free and EF-G-bound). This mismatch is resolved when we extend our EB method to identify the two subpopulations within the set of multiple E_{FRET} trajectories. Similarly, combining results from multiple trajectories using our EB method allows us to see that the distribution of E_{FRET} values associated with a given conformational state often exhibits heavy tails and is skewed relative to the Gaussian distribution that is typically assumed in HMM analyses of smFRET data. Whereas discrepancies between the data and the statistical model will always exist, they are much more difficult to detect in individual trajectories (e.g., in ML- and VB-based HMM analyses of smFRET data). An important advantage of the EB method, therefore, is that it can tease out such discrepancies, which inform us how our assumptions about the data need to be adjusted in the next iteration of statistical model design.

We conclude by noting that the EB estimation framework is applicable to a wide range of single-molecule techniques. Although here we have analyzed smFRET experiments exclusively, our approach is by no means restricted to this platform. Adaptation of the EB algorithm presented here to the analysis of optical trapping and magnetic tweezers experimental data is possible with minimal modifications and we have recently collaborated to adapt the EB algorithm presented here to the analysis of tethered particle motion experiments (33).

ACKNOWLEDGEMENTS

The authors would like to thank Margaret Elvekrog, Kevin Emmett, Jingyi Fei, Jason Hon, Daniel MacDougall, Jordan McKittrick, and the reviewers, for their comments on this manuscript. It is also our pleasure to acknowledge helpful discussions, Martin Lindén, 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. Liu, Y., J. Park, K. A. Dahmen, Y. R. Chemla, and T. Ha, 2010. A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis. *J. Phys. Chem. B* 114:5386–403.

18. 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.

19. 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.

20. 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.

21. 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.

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

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

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

25. 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.

26. Blanco, M., and N. G. Walter, 2010. Analysis of Complex Single-Molecule FRET Time Trajectories, volume 472. Elsevier Inc., 1 edition.

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

28. Greenfeld, M., J.-W. van de Meent, D. S. Pavlichin, H. Mabuchi, C. H. Wiggins, R. L. Gonzalez Jr., and D. Herschlag, in preparation. Single-molecule data format (SMD): A generalized storage format for raw and processed single-molecule data.

29. 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.

30. 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.

31. 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.

32. Blanchard, S. C., R. L. Gonzalez, H. D. Kim, S. Chu, and J. D. Puglisi. tRNA selection and kinetic proofreading in translation. *Nat. Struct. Mol. Biol.* 11:1008–14.

33. Johnson, S., J.-W. van de Meent, C. H. Wiggins, R. Phillips, and M. Lindén, in preparation. Multiple Lac-mediated loops revealed by Bayesian statistics and tethered particle motion.