

Gene expression

Inferring single-cell gene expression mechanisms using stochastic simulation

Bernie J. Daigle, Jr.¹, Mohammad Soltani², Linda R. Petzold³ and Abhyudai Singh^{2,*}

¹Institute for Collaborative Biotechnologies, University of California, Santa Barbara, CA 93106, ²Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716 and ³Department of Computer Science, University of California, Santa Barbara, CA 93106, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on June 11, 2014; revised on November 5, 2014; accepted on January 4, 2015

Abstract

Motivation: Stochastic promoter switching between transcriptionally active (ON) and inactive (OFF) states is a major source of noise in gene expression. It is often implicitly assumed that transitions between promoter states are memoryless, i.e. promoters spend an exponentially distributed time interval in each of the two states. However, increasing evidence suggests that promoter ON/OFF times can be non-exponential, hinting at more complex transcriptional regulatory architectures. Given the essential role of gene expression in all cellular functions, efficient computational techniques for characterizing promoter architectures are critically needed.

Results: We have developed a novel model reduction for promoters with arbitrary numbers of ON and OFF states, allowing us to approximate complex promoter switching behavior with Weibull-distributed ON/OFF times. Using this model reduction, we created bursty Monte Carlo expectation-maximization with modified cross-entropy method ('bursty MCEM²'), an efficient parameter estimation and model selection technique for inferring the number and configuration of promoter states from single-cell gene expression data. Application of bursty MCEM² to data from the endogenous mouse glutaminase promoter reveals nearly deterministic promoter OFF times, consistent with a multi-step activation mechanism consisting of 10 or more inactive states. Our novel approach to modeling promoter fluctuations together with bursty MCEM² provides powerful tools for characterizing transcriptional bursting across genes under different environmental conditions.

Availability and implementation: R source code implementing bursty MCEM² is available upon request.

Contact: absingh@udel.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The process of gene expression—whereby the information contained in a DNA sequence is converted into RNA and proteins—plays an essential role in the execution of nearly all cellular functions. As a result, the misregulation of this process underlies a large number of human diseases including cancer, diabetes and neurological disorders (Lee and Young, 2013). Despite its importance, the

mechanistic details of gene expression are still not well understood. In particular, we lack a comprehensive molecular-level explanation for expression 'bursts'—periods of intense RNA and protein production separated by periods of quiescence—observed in pro- and eukaryotes (Cai *et al.*, 2006; Raj *et al.*, 2006). Precise characterization of the mechanisms underlying expression bursts is profoundly important, as the properties of these bursts have been implicated in

disease-related processes such as bacterial phenotype switching (Choi *et al.*, 2008) and HIV activation (Singh *et al.*, 2010).

Recent advances in single-cell monitoring and single-molecule detection have made possible the experimental characterization of gene expression bursts (Dar *et al.*, 2012; Golding *et al.*, 2005; Singh, 2014; So *et al.*, 2011; Suter *et al.*, 2011; Yu *et al.*, 2006). Specifically, Suter *et al.* (2011) have quantified transcriptional bursts from 11 endogenous mouse promoters, demonstrating that each observed expression pattern can be approximated using a stochastic two-state model of promoter architecture. This commonly used ‘random telegraph’ model assumes that each promoter can exist in one of two states—‘ON’ or ‘OFF’—with synthesis of RNA only possible in the ON state. Because of intrinsic noise exhibited by the small numbers of molecules involved in transcription (e.g. 1–2 copies of DNA, few available copies of transcriptional regulators) (Raser and O’Shea, 2005), the promoter produces expression bursts by switching randomly between the transcriptionally active (ON) and inactive (OFF) states according to kinetic parameters (rate constants) that can be estimated from single-cell time series data (Suter *et al.*, 2011).

Although conceptually useful and amenable to analytical characterization, the random telegraph model is an oversimplification of the architecture of most promoters. Because of simultaneous regulation by multiple transcription factors as well as chromatin modifications, the effective number of states for most promoters is thought to be larger than two (Zhang *et al.*, 2012). A recent study of the human prolactin gene supports this assertion, where the distribution of time its promoter spent in an inactive state was inferred to be strongly non-exponential and thus indicative of multiple, sequential OFF states (Harper *et al.*, 2011). Classical examples of multi-state promoters include that of P_{RM} in phage lambda, whose complex mechanism of regulation gives rise to 128 regulatory states (Sanchez *et al.*, 2013) and the *Endo16* gene in sea urchin, whose cis-regulatory domain contains >30 binding sites for 15 different proteins that perform combinatorial regulation (Yuh *et al.*, 1998). In light of these observations and the increasing availability of single-cell expression data, computational methods are needed for characterizing complex promoter architectures and efficiently simulating their behavior.

An ideal candidate for such a method would (i) leverage the intrinsic noise of the system to better identify underlying mechanisms (Munsky *et al.*, 2009), (ii) simultaneously infer both the configuration of promoter states and the associated kinetic parameters needed for predictive simulation and (iii) provide computationally efficient performance for a wide range of architectures. Currently existing methods satisfy only a subset of these requirements. Suter *et al.* (2011) performed hidden Markov model parameter inference for two- and three-state promoter architectures, but their models assume constant (noise-free) promoter activity and RNA levels between discretely observed time points and they do not provide an efficient means to characterize architectures with larger numbers of states. We previously developed Monte Carlo expectation-maximization (MCEM) with modified cross-entropy method (MCEM²), which uses statistically exact stochastic simulations to infer kinetic parameters from single-cell time series data (Daigle *et al.*, 2012). However, the original version of MCEM² does not enable characterization of promoter architecture. Toni *et al.* (2009) developed an approximate Bayesian computation-based method for inferring both parameters and model structure using stochastic simulations. Unfortunately, when using this method to discriminate between promoter models with increasing numbers of states, the addition of each state increases the number of unknown kinetic

parameters (e.g. switching rates). In the presence of limited amounts of experimental data, this quickly renders more complex (and thus more realistic) models non-identifiable. We note that this drawback applies to any inference method that represents transitions between individual promoter states explicitly. Finally, stochastic simulation of multi-state promoter architectures suffers from a linear increase in computational cost with the addition of each promoter state, making the study of more complex models difficult.

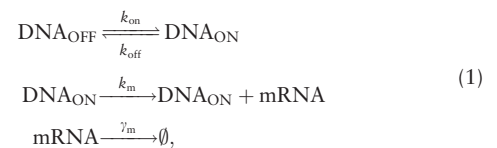
Because of the limitations described above, our goal in this work was to develop a computationally efficient method for characterizing gene expression bursts by inferring the number and configuration of promoter states from single-cell time series data.

2 Results

Our results are structured as follows: we first devise a novel model reduction strategy that represents arbitrary numbers of promoter states by a single state accompanied by a time-dependent switching rate. As we demonstrate below, this strategy enables both efficient simulation and parameter inference. Next, we incorporate this model reduction into MCEM² and augment the method to perform model selection for configuration of promoter states. We demonstrate the resulting approach—‘bursty MCEM²’—by inferring the effective numbers of promoter states underlying simulated single-cell expression data. In addition, we use bursty MCEM² to characterize the architecture of the endogenous mouse glutaminase promoter given experimentally observed time series data (Suter *et al.*, 2011).

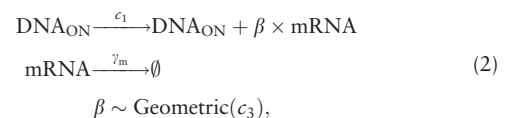
2.1 Model reduction for multi-state promoters

The random telegraph model of transcription can be represented by the following four chemical reactions:



where the promoter randomly switches between OFF and ON states according to rates k_{on} and k_{off} . RNA synthesis occurs at rate k_m from the ON state and expressed mRNAs live for an exponentially distributed time interval with mean lifetime $1/\gamma_m$, where γ_m is the mRNA degradation rate.

A simplified version of (1) that is often used to model transcriptional bursting of mRNAs is:



where c_1 and β denote the burst frequency and size, respectively. In the stochastic formulation of (2) [see Gillespie (2007) for background information], mRNA bursts arrive at exponentially distributed time intervals with rate c_1 . Each expression event generates a geometrically distributed number of transcripts β with mean value $(1 - c_3)/c_3$ (Evans *et al.*, 2000). Model (2) provides an increasingly accurate approximation of (1) as $k_{\text{off}} \rightarrow \infty$, with $c_1 = k_{\text{on}}$ and $c_3 = k_{\text{off}}/(k_{\text{off}} + k_m)$. A sample trajectory of mRNA

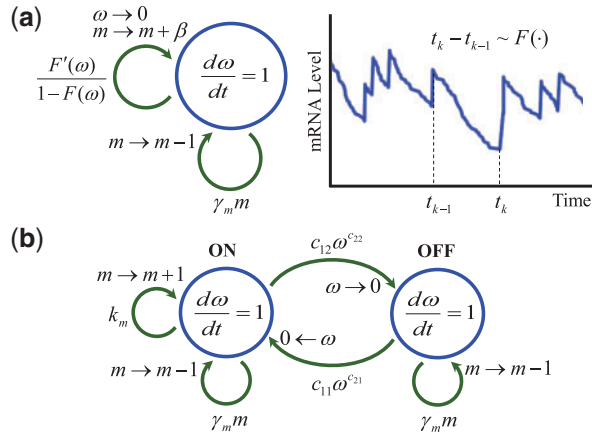
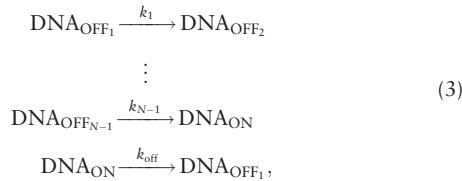


Fig. 1. Schematic of bursty gene expression models. (a) Transcriptional bursting model where each expression event generates a geometrically distributed burst of mRNAs. Each mRNA degrades at a constant rate γ_m . Bursting and decay reactions are represented by arrows with propensity functions directly adjacent (e.g. $\gamma_m m$). The state changes at the end of the arrows (e.g. $m \rightarrow m-1$) denote the change in $m(t)$ (mRNA level at time t) and/or $\omega(t)$ (timer state) when the reactions occur. The timer is set to zero when mRNA bursts are produced, and it increases as $\omega(t) = 1$ between bursts. An example of mRNA trajectory is displayed in the inset plot. The desired CDF $F(\cdot)$ of the inter-burst arrival time ($t_k - t_{k-1}$ in the plot) is obtained by setting the burst event propensity function to $F'(\omega)/(1-F(\omega))$. (b) Promoter-switching model where the promoter randomly toggles between ON and OFF states. Choosing the ON-OFF transition propensity functions as monomials in $\omega(t)$ results in Weibull-distributed ON and OFF times

copy numbers with transcriptional bursts followed by decay is illustrated in Figure 1.

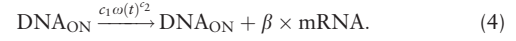
Given the evidence discussed above that promoters exhibit multiple, sequential OFF states, a more realistic representation of the reversible reaction in (1) is the following:



where the promoter exhibits a multi-step OFF to ON transition with $N-1$ OFF states and a single ON state. In contrast to (1), the distribution of time spent in OFF states between transcription events is now non-exponential. Specifically, these time increments follow the hypoexponential distribution, which approaches an Erlang distribution (Evans et al., 2000) when the switching rates (k_1, \dots, k_{N-1}) are roughly equal.

The construction of a simplified transcriptional bursting model of (3) that mirrors (2) requires the generation of inter-burst arrival times from an arbitrary distribution. This can be realized by making the burst frequency time dependent (Fig. 1a). Let $\omega(t)$ denote the state of a timer at time t measuring time elapsed since the last burst event: $\omega(t)$ is reset to zero whenever bursts occur and $d\omega/dt = 1$ in between burst events. The desired inter-burst arrival time cumulative distribution function (CDF) $F(\cdot)$ is obtained by setting the propensity function of the transcription reaction to $F'(\omega)/(1-F(\omega))$, a quantity also known as the hazard function or failure rate (Evans et al., 2000). Thus, the probability that an mRNA burst event will occur in the next time interval $[t, t+dt)$ is $F'(\omega)/(1-F(\omega))dt$.

In this work, we consider transcription propensity functions of the form $c_1\omega(t)^{c_2}$, where $c_1 > 0$ and $c_2 \geq 0$. The corresponding simplified transcription reaction is represented as:



For this class of propensity functions, the inter-burst arrival time distribution can be obtained as follows:

$$\frac{dF(\omega)}{d\omega} \frac{1}{1-F(\omega)} = c_1\omega^{c_2} \quad (5)$$

$$\Rightarrow F(\omega) = 1 - \exp\left(-\frac{c_1}{c_2+1}\omega^{c_2+1}\right). \quad (6)$$

Equation (6) is the CDF of the Weibull distribution with shape parameter k_W and scale parameter λ_W , where $c_1 = k_W/\lambda_W^{k_W}$ and $c_2 = k_W - 1$ (Evans et al., 2000). The mean (represented by $E(\cdot)$) and coefficient of variation squared ($\text{CV}^2 \equiv \text{variance}/\text{mean}^2$) of this distribution can be expressed as follows:

$$E(\omega(t_k)) = \left(\frac{c_2+1}{c_1}\right)^{\frac{1}{c_2+1}} \Gamma\left(\frac{c_2+2}{c_2+1}\right) \quad (7)$$

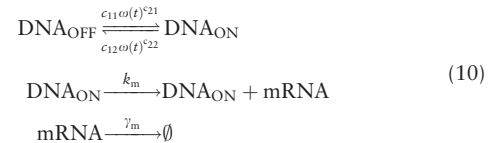
$$\text{CV}^2(\omega(t_k)) = \frac{\Gamma\left(\frac{c_2+3}{c_2+1}\right)}{\Gamma^2\left(\frac{c_2+2}{c_2+1}\right)} - 1, \quad (8)$$

respectively, where t_k denotes the k th mRNA burst time and $\Gamma(\cdot)$ is the gamma function. When $c_2 = 0$ (time-independent burst frequency), the Weibull distribution reduces to an exponential distribution and $\text{CV}^2 = 1$. As c_2 increases, CV^2 monotonically decreases to zero. Thus, for large values of c_2 , mRNA bursts arrive at deterministic time intervals. The Weibull distribution is known to provide an accurate approximation of the Erlang distribution (Malhotra and Reibman, 1993). This property allows us to mimic the behavior of an N -state promoter with roughly equal switching rates (3) using a simplified model of transcription with the propensity function $c_1\omega(t)^{c_2}$. Specifically, by increasing the value of c_2 , we can simulate expression from promoters with larger numbers of states. The relationship between c_2 and N is given by the following expression, which is obtained by equating the CV^2 of the Weibull and Erlang distributions:

$$\frac{\Gamma\left(\frac{c_2+3}{c_2+1}\right)}{\Gamma^2\left(\frac{c_2+2}{c_2+1}\right)} = \frac{N}{N-1}, \quad (9)$$

where N is one greater than the Erlang shape parameter (k_E in the Erlang CV^2 formula: $1/k_E$).

In model (2) and its modification (4), we assume that mRNA bursts are instantaneous. This assumption can be relaxed by generalizing (2) to:



where (c_{11}, c_{21}) and (c_{12}, c_{22}) contribute to the first and second reaction propensity, respectively, and the timer $\omega(t)$ resets to zero each time the promoter transitions between the OFF and ON states (Fig. 1b). By choosing the transition rates to be monomials in $\omega(t)$, the promoter resides in each state for a Weibull-distributed time interval. Setting $c_{21} = c_{22} = 0$ recovers the standard random telegraph model where promoter ON and OFF times are exponentially

distributed random variables. Note that when $c_{22} = 0$, in the limit of large c_{12} (i.e. promoter ON state is unstable), (10) reduces to the instantaneous mRNA burst model [modification (4) of (2)]

2.2 Bursty MCEM²

Single-cell time course datasets are necessarily incomplete—at best, they provide the numbers of molecules for a subset of the system species at discrete time points. Because of the intractability of the incomplete data likelihood for stochastic biochemical systems (Wilkinson, 2006), the direct calculation of maximum likelihood parameter estimates (MLEs) is only possible given complete data. In light of this, we resort to an iterative, simulation-based approach for computing MLEs using the MCEM algorithm. Previously, we developed MCEM² (MCEM with modified cross-entropy method), a computationally efficient approach for estimating parameters of stochastic biochemical systems given incomplete data (Daigle *et al.*, 2012). We describe this approach in more detail in [Supplementary Section S1](#), and below we present the modifications needed to accommodate the models of transcriptional bursting discussed in the previous section.

Given our time-dependent transcriptional model reduction, it is relatively straightforward to construct closed-form update formulas that enable MCEM² to infer maximum likelihood estimates for c_1 and the geometrically distributed burst size parameter c_3 (\bar{c}_1 and \bar{c}_3 , respectively). Because of the occurrence of c_2 in an exponent, it is not possible to derive closed-form update formulas for this parameter, so we fix its value in the following derivation and select most probable values for c_2 separately via model selection. Specifically, we use MCEM² to identify which value of c_2 provides the lowest Akaike information criterion (AIC), a model selection score that quantifies both data likelihood and model complexity (see [Supplementary Section S3](#) for details).

We begin by representing each of the system's M reaction propensity functions (where M is the total number of reactions in the system) as an explicitly state- and time-dependent function $a_j(t) = a_j(X(t), t)$ defined as follows:

$$a_j(X(t), t) \equiv c_{1j} b_j(X(t)) \times \omega_j(t)^{c_{2j}}, \quad (11)$$

where $X(t)$ represents the state of the system at time t , $b_j(X(t))$ is the system state-dependent portion of the propensity function and $\omega_j(t)$ is a reaction-specific timer that resets to zero each time the j th reaction fires. We note that c_{1j} , c_{2j} and $\omega_j(t)$ are designated with arbitrary subscripts only for the purposes of the below derivation; in practice, each promoter is associated with only one timer $\omega(t)$ and at most two values each of c_1 and c_2 . For system reactions with mass action propensity functions (e.g. $a_j(t) = a_j(X(t)) = \theta_j b_j(X(t))$), the representation of Equation (11) is achieved by relabeling θ_j as c_{1j} and setting $c_{2j} = 0$. For the remaining system reactions that exhibit time-dependent propensities as outlined in the above model formulation, $b_j(X(t)) = 1$ and c_{2j} is nonzero.

With all reaction propensities in the form of Equation (11), we can use the results of Anderson (2007) to express the distribution function of the time interval τ until the next reaction firing, given the system at the current time t , as:

$$F(\tau, t) = 1 - \exp \left(- \sum_{j=1}^M \int_t^{t+\tau} c_{1j} b_j(X(t)) \times \omega_j(s)^{c_{2j}} ds \right), \quad (12)$$

where $X(t)$ is constant in the integral because no reactions take place within the time interval $[t, t + \tau)$. After differentiating Equation (12) with respect to τ , integrating and simplifying (given

$\omega_j(t + \tau) = \omega_j(t) + \tau$), we obtain the corresponding density function:

$$f(\tau, t) = \left(\sum_{j=1}^M c_{1j} b_j(X(t)) \times (\omega_j(t) + \tau)^{c_{2j}} \right) \times \exp \left(- \sum_{j=1}^M c_{1j} b_j(X(t)) \times \frac{(\omega_j(t) + \tau)^{\bar{c}_{2j}} - \omega_j(t)^{\bar{c}_{2j}}}{\bar{c}_{2j}} \right), \quad (13)$$

where $\bar{c}_{2j} = c_{2j} + 1$. Given the time of the next reaction firing $t + \tau$, we can also express the probability mass function of the index of the next fired reaction j' as a simple categorical probability:

$$p(j', \tau, t) = \frac{c_{1j'} b_{j'}(X(t)) \times (\omega_{j'}(t) + \tau)^{c_{2j'}}}{\sum_{j=1}^M c_{1j} b_j(X(t)) \times (\omega_j(t) + \tau)^{c_{2j}}}. \quad (14)$$

Taken together, Equations (12–14) allow us to represent the likelihood of a fully observed stochastic trajectory as the following product:

$$\prod_{i=0}^{r_k-1} (f(\tau_i, t_i) \times p(j'_{i+1}, \tau_i, t_i)) \times (1 - F(\tau_{r_k}, t_{r_k})), \quad (15)$$

where i indexes the events including the start of each simulation ($i = 0$), the total number of overall reaction firings r_k and arrival at the final time ($i = r_k + 1$), τ_i is the time interval between the i th and $(i + 1)$ th events, t_i represents the time immediately after the i th event and j'_i is the index of the i th reaction to fire. By taking the logarithm of Equation (15), differentiating with respect to c_{1j} , solving for the unique root and averaging across K' simulated trajectories, we obtain a closed-form Monte Carlo update formula:

$$\hat{c}_{1j}^{(1)} = \hat{c}_{1j}^{(0)} \times \frac{\sum_{k=1}^{K'} r_{jk}}{\sum_{k=1}^{K'} \left(\sum_{i=0}^{r_k} \hat{c}_{1j}^{(0)} b_j(X_k(t_{ik})) \times \frac{(\omega_{jk}(t_{ik}) + \tau_{ik})^{\bar{c}_{2j}} - \omega_{jk}(t_{ik})^{\bar{c}_{2j}}}{\bar{c}_{2j}} \right)}, \quad (16)$$

where $\hat{c}_{1j}^{(0)}$ and $\hat{c}_{1j}^{(1)}$ represent the initial guess and first update, respectively, for parameter c_{1j} , k indexes the K' simulated trajectories and r_{jk} is the number of times the j th reaction fires. The equivalence of Equation (16) and the update formula for parameters from mass action reactions [Equation (1) in the [Supplementary Information](#)] can be seen by setting $c_{2j} = 0$ and relabeling $\hat{c}_{1j}^{(0)}, \hat{c}_{1j}^{(1)}$ as $\hat{\theta}_j^{(0)}, \hat{\theta}_j^{(1)}$. For those reactions with time-dependent propensities, Equation (16) provides a closed-form expression for inferring the maximum likelihood estimate of c_{1j} using MCEM².

For time-dependent models that incorporate a geometrically distributed transcriptional burst reaction, inference of parameter c_{3j} proceeds by first modifying Equation (15) as follows:

$$\prod_{i=0}^{r_k-1} (f(\tau_i, t_i) \times p(j'_{i+1}, \tau_i, t_i) \times g(\beta_{i+1}, j'_{i+1})) \times (1 - F(\tau_{r_k}, t_{r_k})), \quad (17)$$

where $g(\beta, j')$ represents the geometric probability mass function evaluated at a burst size of β ($(1 - c_{3j})^\beta c_{3j}$) if j' is a transcriptional burst reaction and 1 otherwise. Following the same procedure as for c_{1j} , we obtain a closed-form update formula for c_{3j} :

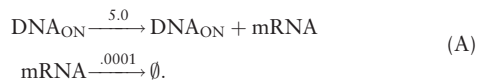
$$\hat{c}_{3j}^{(1)} = \hat{c}_{3j}^{(0)} \times \frac{\sum_{k=1}^{K'} r_{jk}}{\sum_{k=1}^{K'} (r_{jk} + \sum_{i=1}^{r_k} \mathbb{1}_{j'}(j'_{ik}) \times \beta_{ik})}, \quad (18)$$

where $\mathbb{1}_j(j')$ is an indicator function that takes a value of 1 if $j' = j$ (0 otherwise).

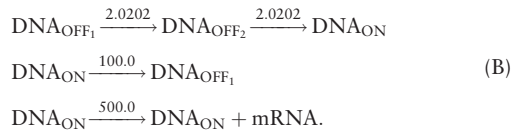
Taken together, the above derivations combined with MCEM² enable the construction of ‘bursty MCEM²’, a novel parameter estimation and model selection framework for inferring the number and configuration of promoter states from single-cell expression data. [Supplementary Section S2](#) provides additional details of the method.

2.3 Simulation study

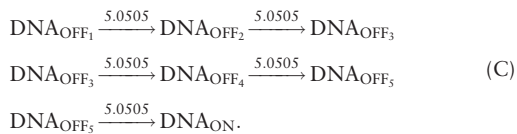
To illustrate the accuracy of our model reduction along with bursty MCEM²'s ability to select a plausible multi-state transcriptional model given observed data, we first performed a simulation study using three models with different numbers of promoter states. [Model \(A\)](#) represents a single-state promoter, which exhibits constant (non-bursty) production of mRNA:



In the above system, the transcription reaction produces an average of five mRNA molecules per unit of time, whereas mRNAs are degraded at a much lower rate. [Model \(B\)](#) contains a promoter with three states—two OFF states and a very short-lived ON state—which exhibits bursty mRNA production. It modifies [\(A\)](#) by replacing the first reaction with the following:



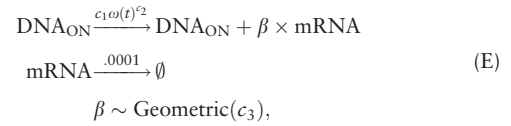
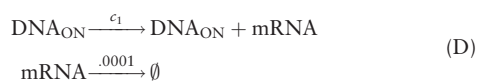
As before, the rates of [\(B\)](#) were chosen so that the promoter would produce five mRNA molecules per time unit on average. However, rather than exhibiting constant production, the promoter switches from OFF to ON once per time unit on average and quickly outputs a burst of mRNA molecules with mean burst size $500.0/100.0 = 5$ before returning to OFF. Finally, [model \(C\)](#) contains a promoter with six states that also exhibits bursty mRNA production. It modifies [\(B\)](#) by replacing the first two reactions with the following:



As in [model \(B\)](#), this promoter switches from OFF to ON once per time unit on average, and it outputs bursts of mRNA with mean size 5 before returning to OFF. The mRNA degradation rates of all three models are identical.

For each model, we first simulated a single trajectory over 100 time units and recorded the number of mRNA molecules at 400 equally spaced intervals. The starting conditions for each simulation were 0 mRNA molecules and 1 promoter in states DNA_{ON} , $\text{DNA}_{\text{OFF}_1}$ and $\text{DNA}_{\text{OFF}_1}$ for [models \(A\)](#), [\(B\)](#) and [\(C\)](#), respectively. [Supplementary Figure S1](#) ([Supplementary Section S5](#)) displays the three simulated trajectories.

Given each synthetic data trajectory, we used our time-dependent transcriptional model reduction with bursty MCEM² to infer the unknown parameters (c_1 , c_2 , c_3) from [models \(D\)](#) and [\(E\)](#):



where the mRNA degradation rate is given. The general version of [model \(E\)](#) is shown in [Figure 1a](#). We used bursty MCEM² to perform parameter inference and model selection on each data trajectory using [model \(D\)](#) and three versions of [model \(E\)](#) with different values of c_2 : (i) $c_2 = 0$, (ii) $c_2 = 0.4$ and (iii) $c_2 = 1.4$. These four model parameterizations correspond to promoters with one, two, three and six states, respectively. We set the initial guesses for parameters c_1 and c_3 to 1 and 0.5, respectively. Since the initial value of the reaction clock $\omega(0)$ was unobserved, we set it to 0 for all bursty MCEM² model simulations. [Table 1](#) lists the resulting MLEs (\hat{c}_1 , \hat{c}_3), log likelihoods, AIC values and relative likelihoods for each data-model combination.

As this table shows, bursty MCEM² selected the model with the correct number of states as most probable for each synthetic data trajectory. For data from the one- [\(A\)](#) and six-state [\(C\)](#) models, competing models did not achieve a relative likelihood within an order of magnitude of a 0.368 cutoff (relative likelihood difference of 2 AIC units from best fitting model); for data from the three-state [model \(B\)](#), the two-state version of [model \(E\)](#) provided a second-best fit that was still substantially worse (likelihood = 0.052) than the cutoff. When inferring parameters of the one-state [model \(D\)](#) using data from [model \(A\)](#), $\hat{c}_1 = 4.81$ is close to the true value of 5 and \hat{c}_1 's 68% confidence interval (4.59–5.03) includes the true value. Similarly, the inferred mean burst sizes for the most probable models fitting data from [models \(B\)](#) and [\(C\)](#) (5.37 and 4.20, respectively) are close to the true value of 5 and the 68% and 95% confidence intervals, respectively, of these two estimates ([4.73–6.09] and [3.28–5.32]) both include the true value. Finally, we note that results for [model \(D\)](#) were unavailable when fitting data from [models \(B\)](#) and [\(C\)](#), as the lack of transcriptional bursts in a one-state model precluded the simulation of trajectories matching data with bursts.

2.4 Glutaminase data inference

Next, we applied bursty MCEM² to actual time-lapse microscopy data from a reporter gene driven by a mammalian promoter. [Suter et al. \(2011\)](#) measured gene expression at eight endogenous mouse promoters upstream of luciferase reporter genes. Using these data, the authors estimated the values of transcriptional parameters by modeling each promoter as occupying two or three states. Our goal was to use our model reduction with bursty MCEM² to infer whether one of these promoters likely occupies more than three states. To this end, we extracted a single trajectory of luminescence data collected once every 5 min from the glutaminase promoter ([Fig. 1C](#) in [Suter et al., 2011](#)) and performed data smoothing and calibration to convert light intensity values to numbers of proteins (see [Supplementary Section S4](#) for details). [Figure 2](#) displays the glutaminase trajectory before and after preprocessing.

We then used bursty MCEM² to infer the unknown parameters (c_1 , c_2 , c_3 , k_{on} , k_{off} , k_{m}) from [models \(F\)](#), [\(G\)](#), [\(H\)](#) and [\(I\)](#) given the data from [Figure 2b](#):

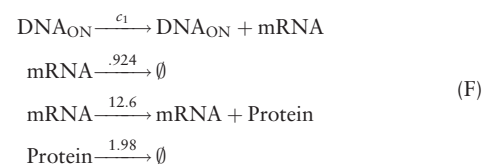


Table 1. Bursty MCEM² parameter inference and model selection results for simulation study

No. states data	No. states model	c_2	\tilde{c}_1	$\frac{1-\tilde{c}_3}{\tilde{c}_3}$	Log likelihood	AIC	Relative likelihood
1 (A)	1 (D)	NA	4.81	NA	-566.94	1135.87	1
	2 (E)	0	41.43	0.12	-569.95	1145.91	0.0066
	3 (E)	0.4	139.67	0.16	-570.50	1146.99	0.0038
	6 (E)	1.4	1990.31	0.26	-572.42	1150.85	0.00056
3 (B)	1 (D)	NA	NA	NA	NA	NA	NA
	2 (E)	0	0.97	5.04	-402.32	810.65	0.052
	3 (E)	0.4	1.08	5.37	-399.37	804.74	1
	6 (E)	1.4	1.48	5.18	-402.91	811.82	0.029
6 (C)	1 (D)	NA	NA	NA	NA	NA	NA
	2 (E)	0	1.19	3.60	-421.53	849.05	1.14×10^{-7}
	3 (E)	0.4	1.40	4.00	-412.58	831.16	0.00088
	6 (E)	1.4	1.91	4.20	-405.54	817.08	1

Corresponding model letter is displayed next to the number of states in the first and second columns. \tilde{c}_1 and \tilde{c}_3 represent MLEs for c_1 and c_3 , respectively. The expression $(1 - \tilde{c}_3)/\tilde{c}_3$ represents mean burst size for models with transcriptional burst reactions. True values for c_1 and $(1 - c_3)/c_3$ are both equal to 5. Relative likelihoods are computed separately for each data trajectory; results in bold represent most probable models. Values of 'NA' for c_2 and $(1 - \tilde{c}_3)/\tilde{c}_3$ due to model (D) lacking these parameters; remaining NA values from fitting data of models (B) and (C) due to inability of one-state model to generate trajectories matching data with bursts.

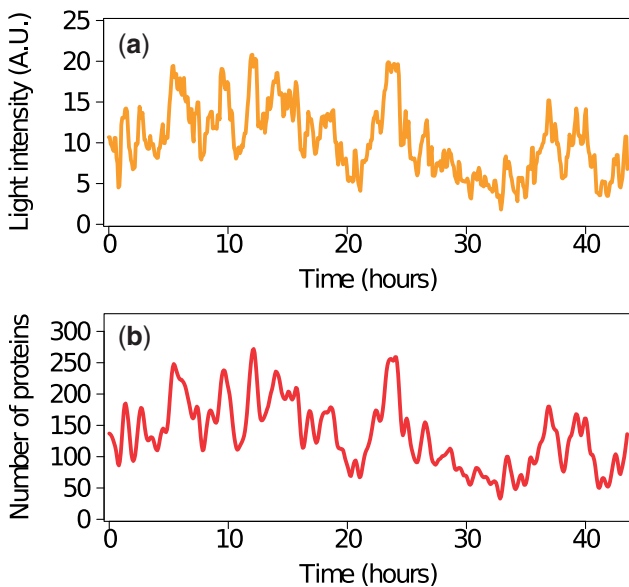
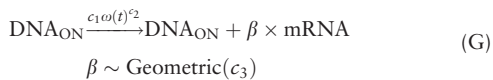


Fig. 2. Glutaminase promoter time-lapse microscopy data from Suter *et al.* (2011). (a) before and (b) after data smoothing and calibration. Data consist of 539 measurements sampled approximately once every 5 min for 43.5 h. AU, arbitrary units



where the latter three models also include the mRNA degradation, protein translation and protein degradation reactions of model (F) with given rate constants. Models (F) and (G) are

similar in structure to models (D) and (E), respectively, with protein translation and degradation reactions added. Models (H) and (I) represent promoters with multi-step ON to OFF and OFF to ON transitions, respectively. Figure 1b represents models (H) and (I) by selecting appropriate values for c_{11} , c_{21} , c_{12} and c_{22} . Like models (B) and (C), these two models exhibit burst-like transcription with the proper parameterization, and they do not assume instantaneous production of bursts. All four models share fixed, identical rates of mRNA degradation (derived from the 45-min glutaminase reporter mRNA half life experimentally determined by Suter *et al.*, 2011), protein degradation (derived from the 21 minute luciferase protein half life experimentally determined by Suter *et al.*, 2011) and protein translation (reported in Molina *et al.*, 2013). We performed model selection over all models, using fixed values of c_2 (when present) ranging from 0 to 11. Altogether, these models approximate the behavior of promoters containing 1–101 states. We set the initial guesses for all other unknown parameters to 1 (except for c_3 , which was set to .5). As before, we set the initial value of the reaction clock $\omega(0)$ to 0 for all bursty MCEM² model simulations. In addition, we set the unobserved initial state of the promoter for models (H) and (I) to DNA_{OFF}. Finally, for the unobserved initial number of mRNA molecules, we tried values from zero to 30 in increments of five. However, as 20 molecules allowed the simulation of trajectories with the largest observed data likelihood, we used this number for all bursty MCEM² model simulations. Supplementary Tables S1 and S2 (Supplementary Section S5) and Table 2 lists the resulting parameter estimates, log likelihoods, AIC values and relative likelihoods (calculated across all models) for versions of models (F), (G), (H) and (I).

As Supplementary Tables S1 and S2 show, no versions of models (F), (G) or (H) provide plausible fits to the data. From these results, we predict that the glutaminase promoter (i) occupies at least one OFF state, (ii) does not exhibit instantaneous bursting of mRNA and (iii) does not undergo a multi-step ON to OFF transition with a single OFF state. Instead, the results in Table 2 suggest that model (I) exhibiting a multi-step OFF to ON transition with a single ON state best fits the data, and we predict that the glutaminase promoter occupies between 10 and 50 OFF states (model versions with relative likelihood $\geq .368$) in the process of transitioning to a single ON state. Once in the ON state, our results predict a mean

Table 2. Bursty MCEM² parameter inference and model selection results for **model (I)** using glutaminase data

No. states	c_2	\bar{c}_1	\bar{k}_{off}	\bar{k}_m	Log likelihood	AIC	Relative likelihood
2	0	2.14	5.52	75.52	-1738.01	3484.02	0.000047
3	0.40	2.78	4.26	72.64	-1734.57	3477.15	0.0015
4	0.75	3.50	3.96	72.62	-1732.58	3473.16	0.011
5	1.10	4.33	3.67	71.15	-1731.10	3470.19	0.047
6	1.40	5.26	3.68	71.89	-1730.33	3468.67	0.10
7	1.65	6.30	3.60	70.85	-1729.75	3467.49	0.18
9	2.10	8.66	3.69	72.10	-1729.08	3466.17	0.35
11	2.50	11.44	3.62	71.12	-1728.60	3465.19	0.58
16	3.40	19.36	3.52	70.23	-1728.06	3464.13	0.98
21	4.10	28.40	3.35	68.14	-1728.06	3464.11	0.99
26	4.75	40.28	3.30	67.62	-1728.05	3464.09	1
51	7.40	161.84	2.84	61.65	-1728.77	3465.53	0.49
101	11.00	1019.21	2.49	56.74	-1729.18	3466.36	0.32

\bar{c}_1 , \bar{k}_{off} and \bar{k}_m represent MLEs for c_1 , k_{off} and k_m , respectively. Relative likelihoods are computed across all versions of **models (F)**, **(G)**, **(H)** and **(I)**; results in bold font represent models that best fit the data (relative likelihood ≥ 0.368).

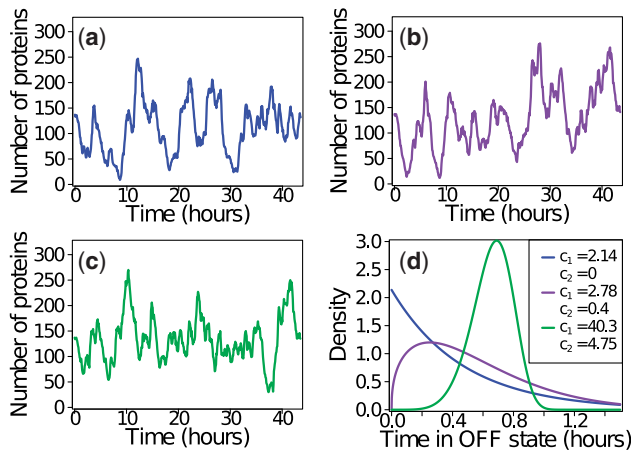


Fig. 3. Differences between versions of **model (I)** with varying numbers of OFF states. Simulated trajectories using the MLEs from **Table 2** for (a) two-state, (b) three-state and (c) most probable 26-state model versions. (d) Comparison of OFF state dwell time distributions between model versions [obtained by differentiating the Weibull distribution function (6)]. Two- and three-state promoter models exhibit more variable OFF times than 26-state promoter model. Values of c_1 and c_2 used for each model version are shown

transcriptional burst size of ~ 20 ($\bar{k}_m/\bar{k}_{\text{off}}$) before returning to the first OFF state.

These predictions are in stark contrast to the two- and three-state promoter models used in *Suter et al. (2011)* to model the glutaminase promoter. To illustrate the differences between these models, **Figure 3** displays simulated trajectories and OFF state dwell time distributions from versions of **model (I)** with 2, 3 and most probable 26 states (and the corresponding MLEs from **Table 2**). As **Figure 3d** shows, the 26-state version of **model (I)** exhibits much less variable OFF state dwell times than the two- or three-state versions. This is reflected in the tendency of the trajectory in **Figure 3c** to stay above a small number of proteins (~ 25), while the trajectories in **Figure 3a** and **b** occasionally decrease to near zero. The glutaminase data trajectory in **Figure 2b** also consistently remains above 25 proteins, providing visual evidence for the superior fit of a 26-state promoter model to the data.

3 Discussion

In this work, we present a novel model reduction for multi-state promoters along with an efficient computational technique for inferring the number and configuration of promoter states from single-cell gene expression data. Specifically, we first developed a time-dependent reaction propensity function for transcriptional bursting that generates Weibull-distributed ON/OFF promoter state dwell times. Using this function, we demonstrated an ability to closely approximate the behavior of promoters undergoing multi-step transitions between OFF and ON states. Next, we created ‘bursty MCEM²’, which, when combined with our model reduction, allows for computationally efficient parameter inference and model selection across a wide range of promoter architectures.

As an example of the computational savings conferred, we note that modeling a promoter transitioning between nine OFF states and one ON state requires specification of 10 switching parameters and simulation of 10 chemical reactions per transcriptional cycle. In contrast, use of our time-dependent propensity function introduces only three parameters and requires the simulation of only two reactions per cycle. These computational and parametric savings increase linearly with the number of promoter states modeled, enabling efficient analysis of arbitrarily complex promoter architectures with bursty MCEM².

Through a simulation study, we demonstrated the ability of bursty MCEM² to correctly identify the number of promoter states used to generate single-cell time-series trajectories. In addition, we showed that our method accurately infers the values of kinetic parameters within the chosen promoter model. We then applied bursty MCEM² to experimental time-lapse microscopy data from a reporter gene driven by the endogenous mouse glutaminase promoter (*Suter et al., 2011*). Our results suggest that, rather than occupying two or three states as previously described, the glutaminase promoter likely traverses through 10 or more OFF states before transcribing mRNA from an ON state. We hypothesize that a distinct advantage of a promoter architecture with >10 states lies in its potential noise buffering properties. As the number of states increases, promoter OFF times become more deterministic, leading to more consistent rates of mRNA production. As shown in **Figure 3c**, this leads to less variable protein levels over time (particularly at lower levels), which can confer more robust behavior in response to stochastic perturbations. Because of the intrinsic noise present in the transcriptional and translational machinery, we note that the confidence interval accompanying our estimated number of glutaminase promoter OFF states is somewhat large (10–50). The addition of measurements at later time points would certainly help reduce this uncertainty. However, if protein measurements were replaced with direct quantification of mRNA, we would expect a much more precise estimate, since translational noise would no longer affect our analysis. Results from the simulation study support this, where we obtained more precise estimates of numbers of promoter states by performing inference on mRNA data.

When discussing particular numbers of promoter states within our model reduction, it is important to note that these values represent ‘effective’ numbers of states rather than distinct biochemical configurations. In particular, since the correspondence between the parameters of the Weibull distribution and the number of promoter states (4) is most accurate when the switching rates are equal, the effective number of states is likely an underestimation of the true number. This follows from the observation that as the switching rates depart from equality, the slowest promoter transitions become rate-limiting and thus mask the presence of faster transitions. Given that

one of our goals in this work was to demonstrate that mammalian promoters occupy more than two or three states, this discrepancy only strengthens the conclusions drawn from our results.

In conclusion, we anticipate that our novel approach to modeling promoter fluctuations together with bursty MCEM² provides powerful tools for characterizing transcriptional bursting across genes under different environmental conditions. Future work will focus on discovering general transcriptional regulatory principles by applying these methods to single-cell expression data from a wide range of promoters.

Funding

A.S. was supported by the National Science Foundation Grant [DMS-1312926], University of Delaware Research Foundation (UDRF) and Oak Ridge Associated Universities (ORAU). B.J.D. and L.R.P. were supported by NIH RO1-EB014877, DOE DE-SC0008975 and the Institute for Collaborative Biotechnologies through grant [W911NF-09-0001] from the U.S. Army Research Office. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. We also acknowledge computing support from the UCSB Center for Scientific Computing from the CNSI, MRL: an NSF MRSEC (DMR-1121053) and NSF CNS-0960316.

Conflict of Interest: none declared.

References

- Anderson,D.F. (2007) A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *J. Chem. Phys.*, **127**, 214107.
- Cai,L. *et al.* (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature*, **440**, 358–362.
- Choi,P.J. *et al.* (2008) A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, **322**, 442–446.
- Daigle,B.J. Jr. *et al.* (2012) Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC Bioinformatics*, **13**, 68.
- Dar,R.D. *et al.* (2012) Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl Acad. Sci. USA*, **109**, 17454–17459.
- Evans,M. *et al.* (2000) *Statistical Distributions*. 3rd edn. Wiley, New York, NY.
- Gillespie,D.T. (2007) Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, **58**, 35–55.
- Golding,I. *et al.* (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.
- Harper,C.V. *et al.* (2011) Dynamic analysis of stochastic transcription cycles. *PLoS Biol.*, **9**, e1000607.
- Lee,T.I. and Young,R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237–1251.
- Malhotra,M. and Reibman,A. (1993) Selecting and implementing phase approximations for semi-Markov models. *Commun. Statist-Stochastic Models*, **9**, 473–506.
- Molina,N. *et al.* (2013) Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proc. Natl Acad. Sci. USA*, **110**, 20563–20568.
- Munsky,B. *et al.* (2009) Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.*, **5**, 318.
- Raj,A. *et al.* (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.*, **4**, e309.
- Raser,J.M. and O'Shea,E.K. (2005) Noise in gene expression: origins, consequences, and control. *Science*, **309**, 2010–2013.
- Sanchez,A. *et al.* (2013) Stochastic models of transcription: From single molecules to single cells. *Methods*, **62**, 13–25.
- Singh,A. (2014) Transient changes in intercellular protein variability identify sources of noise in gene expression. *Biophys. J.*, **107**, 2214–2220.
- Singh,A. *et al.* (2010) Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. *Biophys. J.*, **98**, L32–L34.
- So,L.-H. *et al.* (2011) General properties of transcriptional time series in *Escherichia coli*. *Nat. Genet.*, **43**, 554–60.
- Suter,D.M. *et al.* (2011) Mammalian genes are transcribed with widely different bursting kinetics. *Science*, **332**, 472–4.
- Toni,T. *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.
- Wilkinson,D.J. (2006) *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC Mathematical and Computational Biology Series. Taylor and Francis Group, Boca Raton, FL.
- Yu,J. *et al.* (2006) Probing gene expression in live cells, one protein molecule at a time. *Science*, **311**, 1600–1603.
- Yuh,C.H. *et al.* (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–902.
- Zhang,J. *et al.* (2012) Analytical distribution and tunability of noise in a model of promoter progress. *Biophys. J.*, **102**, 1247–57.