Special function methods for bursty models of transcription

Gennady Gorin[1] and Lior Pachter[2,*]

[1]*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, USA*

[2]*Division of Biology and Biological Engineering & Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California 91125, USA*

(Received 4 April 2020; accepted 10 August 2020; published 31 August 2020)

We explore a Markov model used in the analysis of gene expression, involving the bursty production of pre-mRNA, its conversion to mature mRNA, and its consequent degradation. We demonstrate that the integration used to compute the solution of the stochastic system can be approximated by the evaluation of special functions. Furthermore, the form of the special function solution generalizes to a broader class of burst distributions. In light of the broader goal of biophysical parameter inference from transcriptomics data, we apply the method to simulated data, demonstrating effective control of precision and runtime. Finally, we propose and validate a non-Bayesian approach for parameter estimation based on the characteristic function of the target joint distribution of pre-mRNA and mRNA.

DOI: 10.1103/PhysRevE.102.022409

I. BACKGROUND

Recent improvements in transcriptomics and fluorescence microscopy methods have enabled the rapid and accurate quantification of mRNA on a transcriptome-wide scale with single-molecule precision [1–6]. Simultaneous advances in biophysical and statistical modeling have enabled the effective discrimination of gene expression models and the determination of physical parameters from these data. The estimation of underlying parameters relies on the ability to compute the distribution of molecules for a proposed set of parameters. The chemical master equation (CME) is the standard modeling framework for low-copy single-molecule kinetics, treating such systems with Markov chains traversing state spaces of integer molecule counts [7–9]. However, solutions are available only for a relatively small set of models [8,10]. Furthermore, the existence of a closed-form solution does not guarantee its computational tractability.

Currently popular approaches to solving the CME can be roughly divided into three categories: simulation, matrix, and analytical methods. Simulation methods, such as the Gillespie stochastic simulation algorithm [11,12], are easily implemented and parallelized; the sample statistics of numerous realizations asymptotically approach the statistics of the underlying process, although the speed of approach varies. Matrix methods, such as finite state projection [13] or multifinite buffers [14], rely on matrix exponentiation or eigenvalue calculation to directly solve a truncation of the infinite-dimensional CME system; however, barring convenient symmetries, these methods require a characteristic running time of roughly $O(n^3)$, where n is the state space size. Finally, analytical methods directly solve the underlying system of ordinary differential equations (ODEs), e.g. using a generating function representation [8] or a convenient basis [15] and can be run in $O(n)$ time.

Due to lower computational complexity, these analytical methods are highly relevant to the determination of biophysical parameters from high-dimensional, multimodal data, such as those available by modern transcriptomics and proteomics methods. Recent findings suggest that the use of joint data can provide substantial improvements to model and parameter estimation [16], motivating the development of more efficient solvers for the CME. Current chemistries can quantify spliced and unspliced mRNA molecules [3,17], as well as surface proteins [18,19]. The following multimodal models have analytical CME solutions, as well as drawbacks limiting their direct application to biological data:

(1) Combination of Poissonian solutions [15,20]: cannot be applied to proteomics, and does not explicitly model multistate genes.

(2) Constitutive mRNA and protein production [21]: exact solution, but applies poorly to eukaryotic systems due to prevalence of multistate genes.

(3) Telegraph mRNA and protein production [22,23]: perturbative solution that relies on timescale separation between mRNA and protein lifetimes, and inapplicable to a large fraction of eukaryotic genes.

(4) Multistate gene solutions with a single product [24–26]: exact solution, but does not provide information regarding downstream gene products. Current sequencing methods cannot be easily integrated with DNA accessibility testing.

(5) Bursty mRNA production and isomerization [27]: exact solution, but relies on numerical integration and uses a fairly simple burst model.

A recent method, RNA velocity [17], uses joint distributions of spliced and unspliced mRNA to perform short-time

*Corresponding author: lpachter@caltech.edu

extrapolation on the cell landscape and has been extended to a more detailed treatment using stochastic biophysics [28]. However, the foundation of these methods rests on a Poissonian description of monomolecular systems [15] under a *deterministic* modulation of transcriptional initiation. This description is underdispersed with respect to the conventional approach, which models the transcriptional states as discrete and subject to *stochastic* switching [29]. In its current state, the theoretical foundation of RNA velocity is not feasible to reconcile with a model of transcriptional initiation consisting of discrete binding events.

A common simplification of the conventional approach describes transcription as a rare event with a Poisson process yielding stochastic "bursts" of mRNA. The bursty model describes a large fraction of mammalian genes [30–34] and serves as the implicit foundation of the negative binomial model for scRNA-seq counts [34]. Due to the low abundance of biomolecules [35] and the effectiveness of the discrete gene state model [36], we suggest that unique molecular identifier (UMI)-based scRNA-seq data [37] is most naturally modeled using the CME [10].

A publication by Singh and Bokes [27] describes a model for the bursty production of nascent mRNA and its conversion to mature transcripts; in effect, this system corresponds to a cell population with a single deterministic set of parameters. The computation of probability densities for this model relies on numerical integration. An analytical result is desired for the rapid evaluation of likelihoods, as well as for qualitative insights into the mathematical structure of the solution. To approach this problem, we propose a semianalytical method for the evaluation of joint distributions under this model. Furthermore, we apply this method to parameter estimation, and discuss its applications to a set of burst size distributions that have not been previously solved to the best of our knowledge.

II. METHODS

We follow previous literature [27] in implementing a Markov model for production, isomerization, and degradation of mRNA [Fig. 1(a)]. A single gene locus undergoes transcriptional bursting at a rate of k_i, producing B nascent mRNA transcripts (pre-mRNA) per burst, with $P(B=\rho)=\alpha_\rho$. The nascent transcripts are isomerized to mature mRNA. B is a random variable; if the underlying gene expression follows a two-state telegraph model with short bursts of finite size, B is drawn from a geometric distribution [29]. The reactions are modeled as a Poisson processes with constant rates, which enables their representation using a homogeneous continuous-time Markov chain (CTMC). $P(n,m,t)$, the law of this CTMC model, yields the probability of finding n nascent and m mature molecules at time t.

The full set of CME ODEs is as follows:

$$\begin{aligned}\frac{dP(n,m,t)}{dt} &= k_i\left[\sum_{\rho=0}^{n}\alpha_\rho P(n-\rho,m,t)-P(n,m,t)\right]\\ &+\beta((n+1)P(n+1,m-1,t)-nP(n,m,t))\\ &+\gamma((m+1)P(n,m+1,t)-mP(n,m,t)).\end{aligned}\tag{1}$$

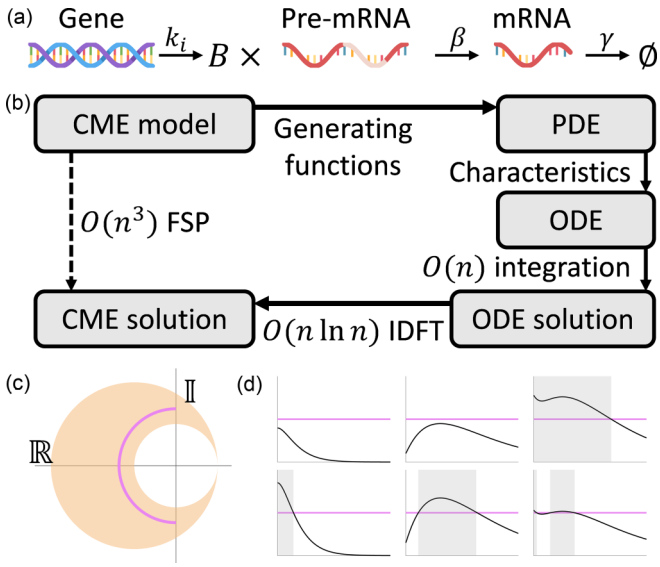

FIG. 1. (a) Schema of modeled physiology (k_i: burst frequency; B: burst size drawn from discrete distribution on $\mathbb{N}$; β: pre-mRNA splicing rate; γ: mRNA degradation rate). (b) Outline of the solution procedure. (c) Taylor and Laurent approximation criterion (orange: approximations common region of convergence; purple: threshold value of $|U|$). (d) Sample shapes of $|U|$ and their partitions (black curve: $|U|$; purple: threshold value of $|U|$; gray: Laurent approximation regions).

Using the probability-generating functions (PGF) $G(x,y,t)=\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}x^n y^m P(n,m,t)$ and $F(x)=\sum_{\rho=0}^{\infty}\alpha_\rho x^\rho$, the CME recurrence relation may be cast into the form of a single partial differential equation (PDE),

$$\frac{\partial G}{\partial t}=k(F(x)-1)G+\beta(y-x)\frac{\partial G}{\partial x}+\gamma(1-y)\frac{\partial G}{\partial y},\tag{2}$$

subject to the initial condition $G(x,y,0)=\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}x^n y^m P(n,m,0)$ and the normalization condition $G(1,1,t)=1$. Introducing the transformations $x=1+u$, $y=1+v$, and $G=e^{\phi}$ results in the following PDE:

$$\frac{\partial\phi}{\partial t}=k(M(u)-1)+\beta(v-u)\frac{\partial\phi}{\partial u}+\gamma v\frac{\partial\phi}{\partial v},\tag{3}$$

such that $M(u)=F(1+u)$. The solution of the PDE at time t is expressed by the following integral:

$$\phi(u,v,t)=k_i\int_0^t[M(U(s))-1]\,ds+\phi(U(t),V(t),0).\tag{4}$$

Per the method of characteristics, $V(s)=ve^{-\gamma s}$, $U(s)=vfe^{-\gamma s}+(u-vf)e^{-\beta s}$ whenever $\gamma\neq\beta$ and $e^{-\gamma s}(u+\gamma vs)$ otherwise, where $f\equiv\frac{\beta}{\beta-\gamma}$. Finally, the PGF G is recovered by exponentiating ϕ. We follow the approach of Bokes [21,27] in evaluating the PGF for x,y around the complex unit circle, interpreting these values as the two-dimensional discrete Fourier transform, or characteristic function values, of the original probability distribution, and converting them to the discrete domain by application of the inverse discrete Fourier transform (IDFT). This method has time complexity $O(\mathcal{N}\log\mathcal{N})$, where $\mathcal{N}$ is the state space size, such that $\mathcal{N}=\max n\times\max m$ of interest [Fig. 1(b)]. For systems with

TABLE I. Integrals of U^i for various approximations and levels of degeneration.

		$\gamma = \beta$	$\gamma \neq \beta$
U		$e^{-\gamma s}(u+\gamma v s)$	$U(s) = vfe^{-\gamma s} + (u-vf)e^{-\beta s}$
Taylor	$v=0$	$-\frac{u^i}{i\gamma}e^{-i\gamma s}$	$-\frac{u^i}{i\beta}e^{-i\beta s}$
	$v\neq 0$	$-\frac{1}{\gamma i}(\frac{v}{i})^i i!\sum_{j=0}^{i}\frac{i^j}{j!}[(\frac{u}{v}+\gamma s)^j e^{-i\gamma s}]$	$\frac{1}{\beta-\gamma}(u-vf)^i$ $\times\sum_{j=0}^{i}\binom{i}{j}(\frac{vf}{u-vf})^j\frac{1}{j-if}e^{[j(\beta-\gamma)-i\beta]s}$
Laurent	$v=0$	$\frac{u^{-i}}{i\gamma}e^{i\gamma s}$	$\frac{u^{-i}}{i\beta}e^{i\beta s}$
	$v\neq 0$	$-\frac{e^{-iu/v}}{\gamma v}(u+\gamma v s)^{1-i}E_i[-\frac{i}{v}(u+\gamma v s)]$	$\frac{(vf-u)^\rho (vf)^{-i-\rho}}{(\gamma-\beta)\rho(\frac{u-vf}{vf}e^{-(\beta-\gamma)s})^\rho}$ $\times{}_2F_1(i,-\rho;-\rho+1;\frac{u-vf}{vf}e^{-(\beta-\gamma)s})$

relatively low copy numbers up to ≈100, where CME modeling is necessary, $N \sim 100\times 100$, requiring on the order of 10 000 evaluations of the integral $\int_0^t [M(U(s))-1]\,ds$.

The model with a geometric burst size distribution of mean b requires the evaluation of $\int_0^t \frac{bU}{1-bU}\,ds$. This integral does not have a closed-form solution and must be treated using repeated numerical quadrature. However, an approximation to the integral can be computed by decomposing the integrand into an integrable power series. Any expression in the form of $\frac{X}{1-X}$ is amenable to an expansion in powers of $X = bU$. In the region $|X|>1$, the Laurent expansion $-\sum_{i=0}^{\infty}X^{-i}$ is available. The intuitive choice of the complementary Taylor expansion $\sum_{i=0}^{\infty}X^i$, which is valid for $|X|<1$, is inappropriate for integration across the boundary $|X|=1$: the approximation diverges and the integral of the expansion ceases to be identical to the original integral. Instead, we leverage the form of U and note that $\mathrm{Re}(U)<0$ for all nontrivial choices of u, v. Therefore, we utilize the Taylor expansion about -1, which is valid for $|X+1|<2$; the form of the series is $-\sum_{i=0}^{\infty}2^{-i-1}(1+X)^i - 1/2$. As shown in the illustration of their shared domain of convergence [Fig. 1(c)], it is possible to select the appropriate approximation based solely on a threshold for the real-valued $|U|$, which simplifies the computation.

Thus, X is decomposed into multiple approximation domains $\{S_j\}$, such that $|X|$ evaluated at the boundary ∂S_j is α, the threshold choice, and successive domains alternate in having $|X|$ strictly greater or less than α [Fig. 1(d)]. As discussed in the Supplemental Material [38], the form of U guarantees that $|\{S_j\}|\leqslant 4$; at most two Laurent and two Taylor approximations are necessary.

Examination of the expansions shows that both can be expressed as $\sum_i \Omega_{j,i}U^i$. If $|U(s)|\geqslant\alpha\forall s\in S_j$, the Laurent approximation is appropriate, and $\Omega_{j,i}=-b^{-i}$. For a Laurent order of approximation N_L, $i\in\{0,-1,-2,\ldots,-N_L\}$. Conversely, if $|U(s)|\leqslant\alpha\forall s\in S_j$, the Taylor approximation is appropriate. For a Taylor order of approximation N_T, binomial expansion of $(1+X)^i$ yields $\Omega_{j,i}=\sum_{k=i}^{N_T}b^i 2^{-k-1}\binom{k}{i}$. The resulting approximation $\sum_i\Omega_{j,i}U^i$ has $i\in\{1,2,\ldots,N_T\}$.

Finally, the full integrand $\frac{bU}{1-bU}$ is approximately $\sum_j\sum_i\Omega_{j,i}U^i$. Therefore, the sought integral $\int_0^t\frac{bU}{1-bU}\,ds$ can be computed using the truncated power series $\sum_j\sum_i\Omega_{j,i}\int_{S_j}U^i\,ds$, where each expansion is integrated only over its appropriate domain of convergence S_j. The details of computation are provided in the Supplemental Material, and the integrals $\int U^i\,ds$ are given in Table I. Numerical routines to evaluate the exponential integral and the Gaussian hypergeometric function are readily available; however, they are not necessarily optimized for speed. We discuss the approximation schema used to make them practical for large-scale computation in the Supplemental Material.

Furthermore, the same approach can be used for other burst distributions. We consider a degenerate distribution (a gene locus that produces b transcripts per burst), a uniform distribution (a gene locus equally probably to produce any number of transcripts between a and b) [47,48], and a shifted geometric distribution (a gene locus guaranteed to produce at least one transcript per burst, e.g., due to the inhibitor being removed by an advancing RNA polymerase). We find that the approximate solutions to these systems can also be expressed in the form $\sum_j\sum_i\Omega_{j,i}\int_{S_j}U^i\,ds$, as shown in Table II. Equivalently, as long as numerical routines are available to compute $\int U^i\,ds$ for $i\in\mathbb{Z}$, a broad array of burst distributions can be computed simply by determining the appropriate integration limits (domains where the expansions converge) and computing the coefficients $\Omega_{j,i}$.

III. RESULTS AND DISCUSSION

We have presented an approximation for the CME solution of bursty pre-mRNA production and its conversion to mature mRNA. We explored several burst distributions discussed in previous studies and explored an extension to a polymerase-inhibitor interaction model. The CME solutions can be found via the computation of $\sum_i\Omega_i\int U^i\,ds$ for the finite-support distributions and $\sum_j\sum_i\Omega_{j,i}\int_{S_j}U^i\,ds$ for the infinite-support distributions. The analytical solutions of $\int U^i\,ds$ are given in Table I, whereas the combinatorial weights for specific burst distributions are given in Table II.

The series form of the solution enables the modulation of approximation order for computational facility [Fig. 2(a)]. The control of method precision and runtime motivates the development of adaptive methods that determine a broad parameter domain using a low-fidelity approximation, then refine it using higher-order or quadrature-based methods.

The purpose of the current investigation is the development of a unified framework for the computation of CME solutions

TABLE II. Integrands, expansion coefficients, summation indices, and expansion domain thresholds associated with approximating the CME solutions for four burst distributions.

	Burst distribution			
	b step	Uniform	Geometric	Shifted geometric
$\frac{1}{k_i}\frac{\partial\phi}{\partial s}$	$(1-U)^b-1$	$\frac{1}{n}\sum_{i=a}^{b}(1+U)^i-1$	$\frac{bU}{1-bU}$	$\frac{bU}{1+(1-b)U}$
$\Omega_{j,i}$	$\binom{b}{i}$	$\frac{1}{n}\left[\binom{b+1}{i+1}-\binom{a}{i+1}\right]$	$b^i\sum_{k=i}^{N_T}\frac{1}{2^{k+1}}\binom{k}{i}$ $-b^{-i}$	$b(b-1)^{i-1}\sum_{k=i}^{N_T}\frac{1}{2^{k+1}}\binom{k}{i}$ $-b(b-1)^{-i-1}$
i	$1,\ldots,b$	$1,\ldots,b$	$1,\ldots,N_T$ $0,-1,\ldots,-N_L$	$1,\ldots,N_T$ $0,-1,\ldots,-N_L$
U	$\mathbb{C}$	$\mathbb{C}$	$\lvert U\rvert<\frac{1+\sqrt{3}}{2b}$ $\lvert U\rvert>\frac{1+\sqrt{3}}{2b}$	$\lvert U\rvert<\frac{1+\sqrt{3}}{2(b-1)}$ $\lvert U\rvert>\frac{1+\sqrt{3}}{2(b-1)}$

for a variety of burst models, as well as the determination of analytical solutions for the approximations. To maintain generality, we do not emphasize a particular implementation of the underlying special functions, but presuppose the availability of efficient implementations of the incomplete gamma and Gaussian hypergeometric function. Nevertheless, as a proof of concept, we develop a case study to benchmark the performance of the degenerate case $\beta=\gamma$. Although this case is limited, it has direct applications to the modeling of multistep splicing: a wide variety of introns are spliced at similar rates, which suggests that modeling the first two splice events as occurring at the same rate is physiologically reasonable [49]. This interpretation corresponds to treating degradation as simply another splice event and considering the distribution of only the first two gene products in a sequential splicing process. Furthermore, we discuss several considerations for implementation and evaluation of the special functions (see the Supplemental Material [38]).

In light of the motivating broader goal of parameter estimation, we use the algorithm to compute likelihood (Kullback-Leibler divergence) landscapes for joint simulated data [11] with a geometric burst size distribution and $b=19$, $k_i=2.5$, $\beta=\gamma=1$ [Fig. 2(b)]. The landscapes produced by the approximation method (shown for $N_L=N_T=7$) closely follow those produced via numerical integration, and we posit that the approximation does not degrade the ability to construct parameter estimates. We validate this result for 24 synthetic data sets with parameters in the shown two-dimensional space and $N_T=7, N_L=1$, achieving similar landscapes and best-fit parameter sets defined by the fifth percentile of the Kullback-Leibler divergence (Supplemental Material: Concordance between quadrature and special function solutions [38]). Repeating this analysis for a range of approximation orders allows benchmarking the method.

Over the entire domain shown in Fig. 2(b), the quality of approximation can be easily controlled by modulating the Taylor approximation order [Fig. 2(c)]. Surprisingly, the order of the Laurent approximation appears to be of minimal relevance to the overall precision. We hypothesize that low-order Laurent approximations are primarily responsible for tail oscillations [Fig. 2(a)], which provide a small contribution to overall divergence due to the low number of data points in that region. Furthermore, the effective reproduction of likelihood landscapes using a first-order Laurent expansion suggests that the result generalizes throughout the parameter space (Supplemental Material: Concordance between quadrature and special function solutions [38]).

The runtime is largely a function of the Laurent approximation order [Fig. 2(d)], due to its explicit reliance on the computation of special functions. We particularly note that the commercial adaptive quadrature method used for benchmarking [50] provides poor control of runtime. We characterize the timing and error behavior in further detail for a particular order of approximation using the parameter estimation procedure with $N_T=7, N_L=1$. The validation procedure yields a mean runtime decrease of 39%. This result is potentially valuable for computationally intensive, large-scale inference over the transcriptome, and suggests that further investigation and optimization can lower the computational costs further without substantial degradation of inference capabilities.

The procedure for regenerating the discrete distributions from generating functions presents certain problems for inference. As shown in Fig. 2(a), the result of the IDFT is not guaranteed to be a probability distribution; the IDFT enforces $\sum_k \pi_k=1$ but does not enforce $\pi_k\geqslant 0\forall k$. However, the properties of the Markov chain ostensibly guarantee that $\pi_k\geqslant 0$, with the inequality becoming strict at equilibrium. For the computation of divergence, we treat this problem in an *ad hoc* manner, by setting $\pi_k\leqslant 0$ to a small float near machine epsilon. A natural, and potentially valuable, extension of this method is the development of transformations using non-negative, non-Fourier basis functions.

An alternative approach is available and yields faster performance at the expense of interpretability in the Bayesian framework. Instead of computing Kullback-Leibler divergence in the probability domain, it is possible to compute a measure of distance between the characteristic functions of proposed and observed distributions or even their corresponding cumulants (logarithms). This approach provides two advantages. First, the roundoff and computational expense of repeated exponentiation and logarithm operations is eliminated. Second, the overall computational complexity of an inference procedure that uses the Fourier transform method and samples $\mathcal{M}$ candidate parameters is $O(\mathcal{M}\mathcal{N}\log\mathcal{N})$. Performing the entire analysis in the Fourier domain requires only a single Fourier transform to determine the empirical

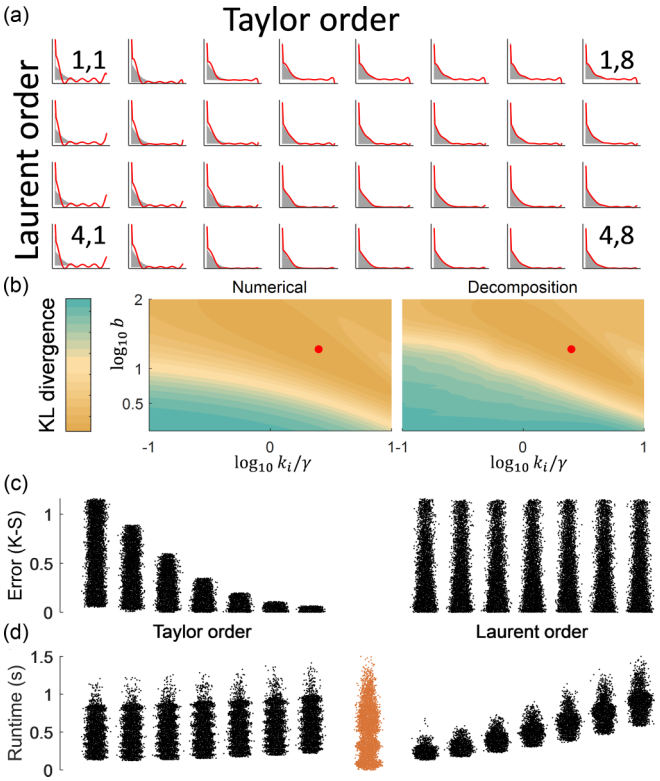

FIG. 2. (a) Comparison of marginal mature mRNA copy-number distributions for a range of approximation orders (#, # tuple and subplot location: Laurent and Taylor approximation order; gray: histogram from 10^5 Gillespie simulations; red line: distribution calculated from approximation). (b) Likelihood landscape for a set of simulated steady-state data with $\gamma = \beta$, calculated over 50×50 trial parameter combinations (numerical: quadrature-based computation; decomposition: expansion-based computation; red point: ground truth). (c) Kolmogorov-Smirnov error between quadrature- and expansion-based joint distributions for parameter sets in (b), calculated for combinations of Taylor and Laurent orders $\in \{1, \ldots, 7\} \times \{1, \ldots, 7\}$ (black point: single parameter set; uniform jitter added). (d) Joint distribution calculation times, determined over the domain in (b) and approximation orders in (c) (black point: single parameter set computed using expansions; orange point: single parameter set computed using numerical quadrature; uniform jitter added).

characteristic function, reducing the computational complexity to $O(\mathcal{N}\log\mathcal{N} + \mathcal{M}\mathcal{N})$, equivalent to $O(\mathcal{M}\mathcal{N})$ in the practical limit of large $\mathcal{M}$. This approach has been explored for use in goodness-of-fit testing and model selection [51,52], but only rarely for parameter inference [53]. However, we anticipate that the computational advantages may outweigh the incompatibility with Bayesian inference, similarly to the recent interest in using nonparametric Kolmogorov and Wasserstein distances for parameter inference [54–56]. Further, we note that optimization of the characteristic function uses information about the entire distribution, potentially overcoming identifiability issues observed using other computationally inexpensive non-Bayesian approaches, such as the method of moments [16]. Since characteristic function methods have primarily been used for analysis of (continuous) stable distributions [53,57,58], their performance for inference from discrete-valued random variable observations has not, to our knowledge, been systematically explored, signifying a substantial lacuna. Thus, this approach is a natural next step for optimizing inference from large data sets.

We implement this method and compare its performance to the Bayesian approach (Supplemental Material: Concordance between likelihood and chf distance landscapes [38]). As expected, the landscapes are qualitatively different, but sufficient to perform parameter estimation. In fact, the observed monotonicity of characteristic function distance landscapes is more amenable to the use of gradient-based global optimization methods. Since the state space sizes considered here are relatively small, the observed improvement in runtime is marginal. However, the performance of the approach does suggest that it may be a useful alternative to Bayesian methods for more general inference problems. In particular, Bayesian estimation does not rigorously apply to situations where the support of the proposed probability distribution does not contain the full support of the data, such as where proposals are generated by simulation [59]. We believe further theoretical characterization of this method is necessary to establish its performance characteristics.

Our discussion of parameter estimation only touched upon inference from steady-state data, which is relevant for fixed-cell experiments that produce information about molecule distributions without a natural time coordinate, such as those available via scRNA-seq [60] and smFISH [61,62]. However, experimental methods with temporal information are available [29,63–66]. Given live-cell data, where cell identities are tracked across time, it is straightforward to extend this method to compute the probability of transitioning from an initial state to any other state, and thus compute the full likelihood of a time series (Supplemental Material: Addenda [38]). Repeating this process for all observed cells, assuming their trajectories are independent, and summing the log-likelihoods of their time series yields a joint likelihood for the observations of the entire experiment [67–69]. Furthermore, given fixed-cell data, where only the *population-level statistics* are tracked across time, it is likewise straightforward to compute the probability of transitioning from one copy-number distribution to another, and use it for likelihood computation [70] (Supplemental Material: Addenda [38]). Conversely, even from a presumed steady-state data set, inferred information about parameters enables the extrapolation of a cell's state to an arbitrary time horizon simply by evaluating the solution with the observed molecule counts as the initial condition, similarly to the RNA velocity procedure [17].

Several mechanistic extensions are available with minimal computational overhead. Technical challenges in single-cell transcriptomics, such as sparsity of sampling in sequencing [71] and noise in fluorescence microscopy [72], have resulted in alternative competing explanations for qualitative features of observed biomolecule distributions, such as heavy-tailed laws [25,27] and apparent dropouts [73–76]. We anticipate that intrinsic degeneracies, as well as aleatory effects, in mapping from a model parameter space to an observable space preclude the unambiguous identification of underlying biophysical schema: the presence of parameter equivalence classes, even in inference of simple models, is well characterized [59,77–79]. Nevertheless, we also anticipate that the development of analytical solutions, as well as numerical

solvers, for a diversity of transcriptional mechanisms, sampling behaviors, and multimodal observables will aid in making inference sufficiently robust for design and extrapolation. For example, as a natural extension, it is straightforward to calculate the laws for observed pre-mRNA and mRNA copy numbers by computing the distributions under an arbitrary sampling schema. This approach enables the natural integration of experimental noise in the same framework as the underlying transcriptional and molecular stochasticity, enabling the simultaneous inference of experimental and physiological parameters.

In summary, we have demonstrated an alternative approach to the computation of semianalytical solutions to the CME and developed a fully analytical form for the marginal distributions. The computation of solutions can be extended to several burst distributions that have not been previously treated in this context. The approximation method yields a set of Taylor and Laurent expansions; validation demonstrates that the computation cost of likelihood computation may be decreased using a low-fidelity Laurent expansion, without substantial distortion of the resulting likelihood landscapes. To facilitate inference even further, we have implemented and validated an inference procedure based directly on the characteristic function. These results have direct applications to improved parameter estimation for transcriptional models, particularly the description of sequential splicing, and provide a foundation for the theoretical study of special function-based approaches to more complex systems that do not currently possess analytical solutions.

The algorithm for the $\beta = \gamma$ system is available at `https://github.com/pachterlab/GP_2020`, along with MATLAB codes to reproduce Fig. 2 and the supplemental figures.

ACKNOWLEDGMENTS

The DNA, pre-mRNA, and mature mRNA used in Fig. 1(a) are derivatives of the DNA Twemoji by Twitter, Inc., used under CC-BY 4.0. The routine for computing the Taylor approximation coefficient $\Omega_{j,i}$ uses a function by Ben Barrowes [80], translated from the FORTRAN original by Zhang and Jin [81]. The routine for computing the Taylor series approximation to the exponential integral $E_1(z)$ is a heavily modified version of a function by Ben Barrowes [80], translated from the FORTRAN original by Zhang and Jin [81]. The subplots in supplemental Figs. 2–4 [38] were aligned using a function by Pekka Kumpulainen [82]. G.G. and L.P. were partially funded by NIH U19MH114830.

[1] H. Xu, S. O. Skinner, A. M. Sokac, and I. Golding, Stochastic Kinetics of Nascent RNA, Phys. Rev. Lett. **117**, 128101 (2016).

[2] S. Lee, A. Y. Zhang, S. Su, A. P. Ng, and A. Z. Holik, M.-L. Asselin-Labat, M. E. Ritchie, and C. W. Law, Covering all your bases: Incorporating intron signal from RNA-seq data Bioinformatics (2018).

[3] S. Shah, Y. Takei, W. Zhou, E. Lubeck, J. Yun, C.-H. L. Eng, N. Koulena, C. Cronin, C. Karp, E. J. Liaw *et al.*, Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH, Cell **174**, 363 (2018).

[4] C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, and L. Cai, Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+, Nature (London) **568**, 235 (2019).

[5] F. Erhard, M. A. P. Baptista, T. Krammer, T. Hennig, M. Lange, P. Arampatzi, C. S. Jürges, F. J. Theis, A.-E. Saliba, and L. Dölken, scSLAM-seq reveals core features of transcription dynamics in single cells, Nature (London) **571**, 419 (2019).

[6] E. M. Wissink, A. Vihervaara, N. D. Tippens, and J. T. Lis, Nascent RNA analyses: Tracking transcription and its regulation, Nat. Rev. Gen. **20**, 705 (2019).

[7] R. Phillips, *PhysicalBiology of the Cell*, 2nd ed. (Garland Science, New York, 2013).

[8] C. Gardiner, *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*, 3rd ed. (Springer, Berlin, 2004).

[9] D. A. McQuarrie, Stochastic approach to chemical kinetics, J. Appl. Prob. **4**, 413 (1967).

[10] P. Érdi and G. Lente, *Stochastic Chemical Kinetics: Theory and (Mostly) Systems Biological Applications*, Springer Series in Synergetics 109 (Springer, New York, 2014).

[11] D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, J. Comput. Phys. **22**, 403 (1976).

[12] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions, J. Phys. Chem. **81**, 2340 (1977).

[13] B. Munsky and M. Khammash, The finite state projection algorithm for the solution of the chemical master equation, J. Chem. Phys. **124**, 044104 (2006).

[14] Y. Cao, A. Terebus, and J. Liang, Accurate chemical master equation solution using multi-finite buffers, Multiscale Model. Sim. **14**, 923 (2016).

[15] T. Jahnke and W. Huisinga, Solving the chemical master equation for monomolecular reaction systems analytically, J. Math. Biol. **54**, 1 (2006).

[16] B. Munsky, G. Li, Zachary R. Fox, D. P. Shepherd, and G. Neuert, Distribution shapes govern the discovery of predictive models for gene regulation, Proc. Nat. Acad. Sci. USA **115**, 7533 (2018).

[17] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A. Furlan *et al.*, RNA velocity of single cells, Nature (London) **560**, 494 (2018).

[18] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert, Simultaneous epitope and transcriptome measurement in single cells, Nat. Methods **14**, 865 (2017).

[19] V. M. Peterson, K. X. Zhang, N. Kumar, J. Wong, L. Li, D. C. Wilson, R. Moore, T. K. McClanahan, S. Sadekova, and J. A Klappenbach, Multiplexed quantification of proteins and transcripts in single cells, Nat. Biotechnol. **35**, 936 (2017).

[20] J. J. Vastola, Solving the chemical master equation for monomolecular reaction systems analytically: A Doi-Peliti path integral view, arXiv:1911.00978 [q-bio] (2019).

[21] P. Bokes, J. R. King, A. T. A. Wood, and M. Loose, Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression, J. Math. Biol. **64**, 829 (2012).

[22] V. Shahrezaei and P. S. Swain, Analytical distributions for stochastic gene expression, Proc. Nat. Acad. Sci. USA **105**, 17256 (2008).
[23] F. Veerman, C. Marr, and N. Popović, Time-dependent propagators for stochastic models of gene expression: An analytical method, J. Math. Biol. **77**, 261 (2018).
[24] T. Zhou and J. Zhang, Analytical results for a multistate gene model, SIAM J. Appl. Math. **72**, 789 (2012).
[25] L. Ham, Rowan D. Brackston, and M. P. H. Stumpf, Extrinsic noise and heavy-tailed laws in gene expression, Phys. Rev. Lett. **124**, 108101 (2020).
[26] L. Ham, D. Schnoerr, R. D. Brackston, and M. P. H. Stumpf, Exactly solvable models of stochastic gene expression, J. Chem. Phys. **152**, 144106 (2020).
[27] A. Singh and P. Bokes, Consequences of mRNA transport on stochastic variability in protein levels, Biophys. J. **103**, 1087 (2012).
[28] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis, Generalizing RNA velocity to transient cell states through dynamical modeling, Nat. Biotechnology (2020).
[29] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Real-time kinetics of gene activity in individual bacteria, Cell **123**, 1025 (2005).
[30] Keren B. Halpern, S. Tanami, S. Landen, M. Chapal, L. Szlak, A. Hutzler, A. Nizhberg, and S. Itzkovitz, Bursty gene expression in the intact mammalian liver, Mol. Cell **58**, 147 (2015).
[31] R. Golan-Lavi, C. Giacomelli, G. Fuks, A. Zeisel, J. Sonntag, S. Sinha, W. Köstler, S. Wiemann, U. Korf, Y. Yarden, and E. Domany, Coordinated pulses of mRNA and of protein translation or degradation produce EGF-induced protein bursts, Cell Rep. **18**, 3129 (2017).
[32] R. D. Dar, B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum, C. D. Cox, M. L. Simpson, and L. S. Weinberger, Transcriptional burst frequency and burst size are equally modulated across the human genome, Proc. Nat. Acad. Sci. USA **109**, 17454 (2012).
[33] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, Mammalian genes are transcribed with widely different bursting kinetics, Science **332**, 472 (2011).
[34] L. Amrhein, K. Harsha, and C. Fuchs, A mechanistic model for the negative binomial distribution of single-cell mRNA counts, Bioinformatics (2019).
[35] B. Munsky, G. Neuert, and A. van Oudenaarden, Using gene expression noise to understand gene regulation, Science **336**, 183 (2012).
[36] J. Peccoud and B. Ycard, Markovian modeling of gene product synthesis, Theor. Pop. Biol. **48**, 222 (1995).
[37] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu *et al.*, Massively parallel digital transcriptional profiling of single cells, Nat. Commun. **8**, 14049 (2017).
[38] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.102.022409 for a detailed derivation of expansions and integrals, which includes Refs. [39–46].
[39] M. Abramowitz and I. Stegun, editors, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th ed. (United States National Bureau of Standards, Washington, DC, 1970).
[40] Y. L. Luke, *Mathematical Functions and Their Approximations* (Academic Press, New York, 1975).
[41] Y. L. Luke, *The Special Functions and Their Approximations*, Vol. II (Academic Press, London, 1969).
[42] T. C. Scott and R. B. Mann, General relativity and quantum mechanics: Towards a generalization of the Lambert W function, arXiv:math-ph/0607011 (2006).
[43] A. Maignan and T. C. Scott, Fleshing out the generalized Lambert W function, ACM Commun. Comput. Algebra **50**, 45 (2016).
[44] T. C. Scott, G. Fee, and J. Grotendorst, Asymptotic series of generalized Lambert W function, ACM Commun. Comput. Algebra **47**, 75 (2014).
[45] P. Castle, Taylor series for generalized Lambert W functions, arXiv:1801.09904 [math] (2018).
[46] M. Carrasco and R. Kotchoni, Efficient estimation using the characteristic function, Econometric Theory **33**, 479 (2017).
[47] S. Be'er, M. Heller-Algazi, and M. Assaf, Effect of reaction-step-size noise on the switching dynamics of stochastic populations, Phys. Rev. E **93**, 052117 (2016).
[48] H. Kuwahara, S. T. Arold, and X. Gao, Beyond initiation-limited translational bursting: The effects of burst size distributions on the stability of gene expression, Integr. Biol. **7**, 1622 (2015).
[49] J. Singh and R. A. Padgett, Rates of in situ transcription and splicing in large human genes, Nat. Struct. Mol. Biol. **16**, 1128 (2009).
[50] MATLAB R2019b (2019), https://www.mathworks.com/products/new_products/release2019b.html .
[51] S. Lee, Simos G. Meintanis, and M. Jo, Inferential procedures based on the integrated empirical characteristic function, AStA Adv. Stat. Analys. **103**, 357 (2019).
[52] M. D. Jiménez-Gamero, A. Batsidis, and M. V. Alba-Fernández, Fourier methods for model selection, Ann. Inst. Stat. Math. **68**, 105 (2016).
[53] M. Bee and L. Trapin, A characteristic function-based approach to approximate maximum likelihood estimation, Commun. Stat. Theory Meth. **47**, 3138 (2018).
[54] M. Sommerfeld and A. Munk, Inference for empirical Wasserstein distances on finite spaces, J. R. Stat. Soc.: Series B **80**, 219 (2018).
[55] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert, On parameter estimation with the Wasserstein distance, arXiv:1701.05146 [math, stat] (2019).
[56] L. Györfi, I. Vajda, and E. van der Meulen, Minimum Kolmogorov distance estimates of parameters and parametrized distributions, Metrika **43**, 237 (1996).
[57] J. Yu, Empirical characteristic function estimation and its applications, Econometric Rev. **23**, 93 (2004).
[58] A. Feuerverger and P. McDunnough, On the efficiency of empirical characteristic function procedures, J. R. Stat. Soc.: Series B **43**, 20 (1981).
[59] G. Gorin, M. Wang, I. Golding, and H. Xu, Stochastic simulation and statistical inference platform for visualization and estimation of transcriptional kinetics, PLoS ONE **15**, e0230736 (2020).
[60] A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, and S. A. Teichmann, The technology and biology of single-cell RNA sequencing, Mol. Cell **58**, 610 (2015).
[61] A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer, Visualization of single RNA transcripts in situ, Science **280**, 585 (1998).

[62] A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi, Imaging individual mRNA molecules using multiple singly labeled probes, Nat. Methods **5**, 877 (2008).

[63] H. G. Garcia, M. Tikhonov, A. Lin, and T. Gregor, Quantitative imaging of transcription in living *Drosophila* embryos links polymerase activity to patterning, Curr. Biol. **23**, 2140 (2013).

[64] E. A. Specht, E. Braselmann, and A. E. Palmer, A critical and comparative review of fluorescent tools for live-cell imaging, Ann. Rev. Phys. **79**, 93 (2017).

[65] J. A. Schofield, E. E. Duffy, L. Kiefer, M. C. Sullivan, and M. D. Simon, TimeLapse-seq: Adding a temporal dimension to RNA sequencing through nucleoside recoding, Nat. Methods **15**, 221 (2018).

[66] V. A. Herzog, B. Reichholf, T. Neumann, P. Rescheneder, P. Bhat, T. R. Burkard, W. Wlotzka, A. von Haeseler, J. Zuber, and S. L. Ameres, Thiol-linked alkylation of RNA to assess expression dynamics, Nat. Methods **14**, 1198 (2017).

[67] B. J. Daigle, M. K. Roh, L. R. Petzold, and J. Niemi, Accelerated maximum likelihood parameter estimation for stochastic biochemical systems, BMC Bioinformatics **13**, 68 (2012).

[68] A. M. Corrigan, E. Tunnacliffe, D. Cannon, and J. R. Chubb, A continuum model of transcriptional bursting, eLife, 5, e13051 (2016).

[69] A. Golightly and D. J. Wilkinson, Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo, Interface Focus **1**, 807 (2011).

[70] S. K. Poovathingal and R. Gunawan, Global parameter estimation methods for stochastic biochemical systems, BMC Bioinformatics **11**, 414 (2010).

[71] D. Grün, L. Kester, and A. van Oudenaarden, Validation of noise models for single-cell transcriptomics, Nat. Methods **11**, 637 (2014).

[72] I. Sgouralis and S. Pressé, An introduction to infinite HMMs for single-molecule data analysis, Biophys. J. **112**, 2021 (2017).

[73] P. Qiu, Embracing the dropouts in single-cell RNA-seq data, Bioinformatics (2018).

[74] V. Svensson, Droplet scRNA-seq is not zero-inflated, Nat. Biotechnology **38**, 147 (2020).

[75] T. Andrews and M. Hemberg, False signals induced by single-cell imputation, F1000Research **7**, 1740 (2019).

[76] W. V. Li and J. J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data, Nat. Commun. **9**, 997 (2018).

[77] C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein, Fundamental limits on dynamic inference from single-cell snapshots, Proc. Nat. Acad. Sci. USA **115**, E2467 (2018).

[78] L. Weber, W. Raymond, and B. Munsky, Identification of gene regulation models from single-cell data, Phys. Biol. **15**, 055001 (2018).

[79] E. Cinquemani, Identifiability and reconstruction of biochemical reaction networks from population snapshot data, Processes **6**, 136 (2018).

[80] B. Barrowes, Computation of special functions, https://www.mathworks.com/matlabcentral/fileexchange/6218-computation-of-special-functions.

[81] S. Zhang and J. Jin, *Computation of Special Functions* (Wiley, New York, 1996).

[82] P. Kumpulainen, tight_subplot (Nh, Nw, gap, marg_h, marg_w), https://www.mathworks.com/matlabcentral/fileexchange/27991-tight_subplot-nh-nw-gap-marg_h-marg_w.