A data-integrated method for analyzing stochastic biochemical networks

Michael W. Chevalier and Hana El-Samad

Citation: The Journal of Chemical Physics **135**, 214110 (2011); doi: 10.1063/1.3664126
View online: http://dx.doi.org/10.1063/1.3664126
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/135/21?ver=pdfcov
Published by the AIP Publishing

A data-integrated method for analyzing stochastic biochemical networks

Michael W. Chevalier[a] and Hana El-Samad[b]
Department of Biochemistry and Biophysics, California Institute for Quantitative Biosciences, University of California San Francisco, 1700, 4th Street, San Francisco, California 94143-2542, USA

(Received 4 April 2011; accepted 2 November 2011; published online 7 December 2011)

Variability and fluctuations among genetically identical cells under uniform experimental conditions stem from the stochastic nature of biochemical reactions. Understanding network function for endogenous biological systems or designing robust synthetic genetic circuits requires accounting for and analyzing this variability. Stochasticity in biological networks is usually represented using a continuous-time discrete-state Markov formalism, where the chemical master equation (CME) and its kinetic Monte Carlo equivalent, the stochastic simulation algorithm (SSA), are used. These two representations are computationally intractable for many realistic biological problems. Fitting parameters in the context of these stochastic models is particularly challenging and has not been accomplished for any but very simple systems. In this work, we propose that moment equations derived from the CME, when treated appropriately in terms of higher order moment contributions, represent a computationally efficient framework for estimating the kinetic rate constants of stochastic network models and subsequent analysis of their dynamics. To do so, we present a practical data-derived moment closure method for these equations. In contrast to previous work, this method does not rely on any assumptions about the shape of the stochastic distributions or a functional relationship among their moments. We use this method to analyze a stochastic model of a biological oscillator and demonstrate its accuracy through excellent agreement with CME/SSA calculations. By coupling this moment-closure method with a parameter search procedure, we further demonstrate how a model's kinetic parameters can be iteratively determined in order to fit measured distribution data.
 [doi:10.1063/1.3664126]

I. BACKGROUND AND MOTIVATION

Variability is a fundamental issue that impacts many areas of biology. In populations of organisms and cells, phenotypic differences between individuals are thought to be caused by a combination of genetic and environmental factors. However, significant variability between individuals of the same genotype in highly similar environments has been observed using optical measurement methods which monitor fluorescent reporter expression in single cells using flow cytometry or fluorescence microscopy. Fluorescence measurements can be used to build histograms that quantify, for example, the fluctuations in a process or variability in expression levels across a population of cells.

Depending on the context, an organism or cell may suppress or exploit a given non-genetic source of variability. For instance, in *Drosophila* embryos, variability in the gradient of the bicoid protein results in undesirable developmental alterations, and studies suggest that the system critically depends on molecular devices that are dedicated to filtering noise.[1,2] In a similar vein, Raj *et al.*[3] show that inducing gene expression variability of the intestinal specification network in the nematode *C. elegans* during development can have indeterminate effects: some mutant embryos fail to develop intestinal cells, whereas others produce intestinal precursors. On the other hand, an organism or cell may use heterogeneity in expression levels productively to allow populations to "hedge their bets" with respect to future environmental shifts. This is thought to explain the ability of subpopulations of bacteria to resist antibiotics and promote latency of infection by viruses.[4,5] As a consequence, the thorough investigation of the principles underlying such non-genetic individuality is essential for understanding the differential susceptibility of cells and organisms to diseases, drugs, and pathogens.

Amongst isogenic individuals under the same environmental conditions, variability is believed to take root in the stochastic nature of biochemical reactions. For example, fluctuations in gene expression, a major contributor of cell-to-cell variability, can arise from the stochastic steps involved in transcription and translation.[6–8] Further sources of fluctuations include diffusion-reactions and dissociations, allosteric changes, and degradation of biological molecules. In general, for a given network, the occurrence of the underlying biochemical reactions depends on the channels of possible interactions (topological structure of the interaction network), the interaction affinity (reaction rates), as well as the number of molecules available for such interactions.[10] This inevitably results in fluctuations in the levels of the various molecular species and translates into cell-to-cell phenotypic differences in cellular behaviors. The impact of such fluctuations was identified in simple synthetic gene circuits, inducing substantial variability in the period of oscillators and stochastic transitions in synthetic toggle-switches.[9] These behaviors can

[a]Electronic mail: Michael.Chevalier@ucsf.edu.
[b]Electronic mail: Hana.El-Samad@ucsf.edu. URL: http://biochemistry.ucsf.edu/labs/elsamad/home/index.html.

only be captured and thoroughly understood in the context of stochastic computational models.

The current stochastic representation of biological systems adopts the formalism of the chemical master equation (CME),[11–13] an N dimensional differential-difference equation (N is the number of model species), or its accompanying numerical stochastic simulation algorithm (SSA).[14] However, for many realistic biological models, the CME is computationally intractable and the SSA is too numerically inefficient. To circumvent these difficulties, recent efforts have focused on generating approximate solutions for the CME. For example, an approach known as the finite state projection (FSP) approach rewrites the CME as a set of linear ordinary differential equations for the probabilities of all the states that can be assumed by the system.[15] Since the number of states can be extremely large or infinite, this method proposes a truncation of the number of states and a certificate of accuracy for how closely the truncated space approximation matches the true solution. The truncated solution can then be used to probe different properties of the model or fit its parameters. For example, the FSP was used to demonstrate how mean and variability data can be exploited to uniquely identify gene expression parameters in a model of the *lac* operon.[16] One caveat to this approach, however, is that scaling to larger networks seems to be challenging. Another approach relies on recasting the problem as a set of coupled differential equations for the mean and higher moments of the multivariate distribution described by the CME. However, for any system with at least one bimolecular reaction, there is an infinite set of moments to solve for.[18] Previous work has attempted to truncate the system of coupled moment equations by assuming *a priori* functional forms for approximate moment closure.[18–20] While this generates a finite system of equations that can be solved analytically or numerically, the validity of these assumptions might not be *a priori* warranted.

In this paper, we are interested in developing a scalable approach for exploiting experimental distribution data to approximate and then to parameterize stochastic models of gene regulatory circuits described by the CME. These models can then be used to computationally probe the interesting dynamical properties of these circuits beyond what is feasible with experimentation alone. To achieve this, we develop a practical data-based methodology for truncating moment equations in the stochastic formulation of biological network models. Our motivation stems from a simple concept: instead of assuming an *a priori* statistical distribution or functional form for moment closure, one can instead use the higher moments of distributions that are directly *measured* from the circuit one is attempting to model. More specifically, we propose a hybrid CME/moment approach that integrates information from the experimental data into appropriate terms within the moment equations. This is not an unreasonable approach given that distribution data are readily available, and that computational models are precisely built to fit and interpret such data. After the fit to data, these models can be probed further to generate predictions about unknown aspects of the underlying networks and gain a deeper understanding of their dynamics.

The data-driven moment closure approach we propose integrates the data in a robust and meaningful manner and scales in a computationally tractable way with the number of genes, both attractive qualities as improvements in the number of simultaneously observable biological variables increases. We apply the approach to a stochastic model of a genetic oscillator and demonstrate its equivalence to the CME. By coupling this moment closure method with standard parameter estimation techniques, we demonstrate its practical usefulness for model parametrization using population data.

II. STOCHASTIC MODELING OF BIOCHEMICAL REACTIONS

In this work, we adopt the traditional framework of a system of well-stirred chemical reactions with N molecular species. We use the state $X(t)$ to denote the vector whose integer elements $X_i(t)$ are the number of molecules of the ith species at time t. If there are M elementary chemical reactions that can occur among these N species, then we associate with each reaction r_j ($j = 1, \ldots, M$) a non-negative *propensity function* defined such that $a_j(X(t))\,\tau + o(\tau)$ is the probability that reaction r_j will happen in the next small time interval $(t, t + \tau]$, as $\tau \to 0$. The polynomial form of the propensities $a_j(x)$ may be derived from fundamental principles under certain assumptions.[14] The occurrence of a reaction r_j leads to a change of $\nu_j \in \mathbf{Z}^N$ (the set of non-negative integers) for the state X. ν_j is, therefore, a stoichiometric vector that reflects the integer change in reactant species due to a reaction r_j.

This set of well-stirred chemical reactions can be represented by the joint probability density function $P(x, t)$ which describes the probability of the system being in state x at time t. The evolution of $P(x, t)$ is given by

$$\frac{\partial P(x,t)}{\partial t} = \sum_{j=1}^{M} [a_j(x - \nu_j)P(x - \nu_j, t) - a_j(x)P(x,t)], \tag{1}$$

Eq. (1) is the so-called CME.[11–13] The CME is also generally derived using the Markov property, by writing the Chapman-Kolmogorov equation, an identity that must be obeyed by the transition probability of any Markov process.

To analyze a stochastic model described by the CME above, numerical values need to be assigned to the kinetic rate constants featured in the propensity functions $a_j(x)$. The most natural methodology is to iteratively identify the parameters that match the species distributions generated by the stochastic computational model to those experimentally measured using microscopy or flow cytometry. However, in approaching this data fitting problem, one needs to solve the CME multiple times (either directly or through SSA simulations), which can be computationally infeasible when the models increase in size or complexity.[17]

Here, we propose that moment equations derived from the CME, when correctly resolved, can be efficiently used for model parametrization using experimental data. First, we note that if the propensity functions $a_j(x)$ present in the CME are at most quadratic (bimolecular binding reactions), that is, $a_j(x) \in \{c_0, c_1 x_i, c_2 x_i(x_i - 1)/2!, c_3 x_i x_k\}$ where c_0, c_1, c_2, c_3 are rate constants, then they can be expressed in the form of a quadratic Taylor series around the mean of the joint

probability density function,[18] $P(x, t)$. Specifically, if $z(t)$ is the solution of the macroscopic rate equations, then $a_j(x)$ can be precisely expressed as

$$a_j(x) = a_j(z(t)) + \sum_{i=1}^{N} \frac{\partial a_j(x)}{\partial x_i}|_{x=z(t)}[x_i - z_i(t)] + \sum_{i,i'}^{N} \frac{\partial^2 a_j(x)}{\partial x_i \partial x_{i'}}|_{x=z(t)} \frac{[x_i - z_i(t)][x_{i'} - z_{i'}(t)]}{2}. \quad (2)$$

In this case, the time dependent mean equations for the kth molecular species described by the CME is given by

$$\frac{\partial z_k(t)}{\partial t} = \sum_{j=1}^{K} \nu_{jk} \left(a_j(z(t)) + \sum_{i,i'}^{N} \frac{\partial^2 a_j(z(t))}{\partial x_i \partial x_{i'}}|_{x=z(t)} \frac{C_{ii'(t)}}{2} \right). \quad (3)$$

In terms of parameter identification, this differential equation for the mean can in principle be efficiently matched using standard parameter fitting algorithms to the mean of the experimentally measured distribution. However, notice here that Eq. (3) contains a covariance term (second centered moment). Therefore, solving this equation requires the computation of the covariance. Likewise, the time-dependent covariance equation contains the third central moment.[18] It is actually the case that when all reactions are at most bimolecular, the time-dependent Nth centered moment equations will contain an $N + 1$th moment terms. For the special case when all propensities are at most affine, the third central moment term is not present in the time-dependent covariance equation and the system closes at the first two central moments. To address the more general (and realistic) cases where moments are interdependent, approximate moment closure methods have been proposed. These methods fall within two categories: they either assume that the third and higher moments are zero[19] or make *a priori* assumption about the functional form or the nature of the closure of the distribution $P(x, t)$. In this fashion, a relationship between the $N + 1$ moments and the first N moments is derived, and the system of moment equations is closed at the Nth moment. For example, one such moment closure approximation known as separable derivative matching[20] approximates the $N + 1$th moment as a polynomial function of the first N moments. This approach matches time derivatives between the approximate closed system and the exact non-closed system at the initial time t_0 and the given initial conditions. This allows the exponents (which remain constant over the simulation) in the polynomial function to be uniquely determined, and the solution turns out to be consistent with the underlying distribution $P(x, t)$ being log-normal.[20] For those biological systems which do not exhibit log-normal behavior, the separable polynomial function maybe too restrictive.

III. DATA-DRIVEN MOMENT CLOSURE

Examining the equation for the mean evolution of a stochastic system described by the CME (Eq. (3)) reveals that when this mean depends on a bimolecular reaction, then only covariances of the reaction's two input species are needed. If single-cell pairwise measurements of the molecular numbers for the species featured in the bimolecular reactions are available as a function of time, then these time-dependent covariances could be numerically estimated from the data. When the parameters of the model are known, the covariances could be plugged into Eq. (3) to close it and the mean equation could be used to study other features of the system such as stability and dynamical behavior. When the parameters of the system are not known, the same procedure coupled to an iterative procedure for comparing the values generated by Eq. (3) to those present in the data could be used to estimate the parameters. Importantly, this could all be accomplished without imposing any assumptions on the originating distribution.

As mentioned above, the approach requires that the input species to the bimolecular reaction be pairwise measurable or at least that the ones that are not directly measurable be estimable with reasonable precision from the available measurements. In practice, simultaneous measurements of proteins using different fluorescent labels are increasingly accessible.[21] mRNA measurements are less common although methods based on Fluorescence *In situ* Hybridization (FISH) (Ref. 22) are rapidly improving this limitation. Gene states (the different configurations resulting from transcription factors or repressors binding and unbinding from that gene's promoter region) are on the other hand not directly measurable in any accurate way. That is, while we can measure simultaneous single-cell levels of transcription factors, measurement technology cannot currently resolve the configuration of transcription factors bound to the gene's promoter region at any given time. This limitation can be circumvented, if we can infer the distribution of gene states in a single cell based on the measured transcription factor profiles in that cell. We demonstrate below that this is feasible if the interactions between the transcription factors and gene promoters occur on a faster time scale than the change in the transcription factor level, an assumption that is often biologically warranted.

A. Closure of protein and mRNA mean equations

In Eq. (3), any bi-molecular reaction that affects the mean will have a contributing covariance term $C_{ii'}(t)$, where the species i and i' represent the two molecular species involved in the bi-molecular reaction. If these species are all proteins, then the covariance term can be directly computed from the data. We define $s_{ii'}(k, t_m)$ as the paired set of measured molecular counts for species i and i' in cell k at sample time t_m. Therefore, $s^i_{ii'}(k, t_m)$ refers to the molecular count measurement for species i, while $s^{i'}_{ii'}(k, t_m)$ is the measurement for species i'. The measured mean of species i is then

$$\mu_i(t_m) = \frac{1}{K} \sum_{k=1}^{K} s^i_{ii'}(k, t_m) \quad (4)$$

and the same for species i'. The covariance term $C_{ii'}(t_m)$ is similarly given by

$$C_{ii'}(t_m) = \frac{1}{K} \sum_{k=1}^{K} \left[s^i_{ii'}(k, t_m) - \mu_i(t_m) \right] \times \left[s^{i'}_{ii'}(k, t_m) - \mu_{i'}(t_m) \right]. \quad (5)$$

For this paper, we will assume first order transcription and mRNA decay. As a result, the mRNA mean equations have affine propensity functions with no covariances present. The time-dependent transcription rates for the mRNA equations are input from the gene-state calculations discussed below in Sec. III B. While these assumptions on mRNA dynamics are a limitation, it should be alleviated in the future as mRNA FISH measurements become more accessible.

The data-derived means and covariances are calculated at the data sample times t_m. However, solving Eq. (3) will require knowledge of $\mu_i(t)$ and $C_{ii'}(t)$ for $t_m \leq t \leq t_{m+1}$. Here, we use a simple linear interpolation

$$\mu_i(t) \approx \frac{(t-t_m)\mu_i(t_{m+1}) + (t_{m+1}-t)\mu_i(t_m)}{t_{m+1}-t_m} \tag{6}$$

and

$$C_{ii'}(t) \approx \frac{(t-t_m)C_{ii'}(t_{m+1}) + (t_{m+1}-t)C_{ii'}(t_m)}{t_{m+1}-t_m}. \tag{7}$$

Covariance terms involving proteins where one of the input species is a gene state are calculated through the transcription factor measurement/gene-state model approach discussed below in Sec. III B.

B. Estimation of gene-state means and gene-state/transcription factor covariances from transcription factor data

As mentioned above, gene states are not currently experimentally observable. However, we can exploit the separation of timescales between transcription dynamics and those of binding/unbinding of transcription factors to gene promoters to estimate gene states from transcription factor data since a transcription factor is a protein whose level can often be measured.

First, we define a gene module g_i as a set of genes with shared transcription factors for which we have simultaneous measurements. The transcription factors bind and unbind to their corresponding gene promoters. A biological network may have numerous gene modules. Figure 1 shows two examples of genes with shared transcription factors. For these examples, let us assume that we have only two-color (pairwise) simultaneous transcription factor measurements. For notational clarity, we define G_a as the gene whose expression results in protein A. In Figure 1(a), genes G_a and G_r share a single common transcription factor A and constitute a gene module. Figure 1(b) illustrates a set of three genes G_x, G_y, G_z that have four overlapping transcription factors A, B, C, D. Specifying the appropriate gene modules in this case is more complicated given the availability of only pairwise transcription factor measurements. For instance, given single-cell simultaneous measurements of A and B, we would need to solve the states of gene G_x while including the sequestration effects of B and A binding to gene G_y and gene G_z, respectively. Additional modules would be setup to solve for the states of gene G_y and gene G_z in a similar fashion. However, in general, one could imagine more complicated gene modules and transcription factor binding rules where M simultaneous transcription factor measurements are needed to solve the steady-state gene module states. While such scenarios pose no additional computational complications to our approach, they might be limiting in terms of the number of simultaneously measurable fluorescent reporters.

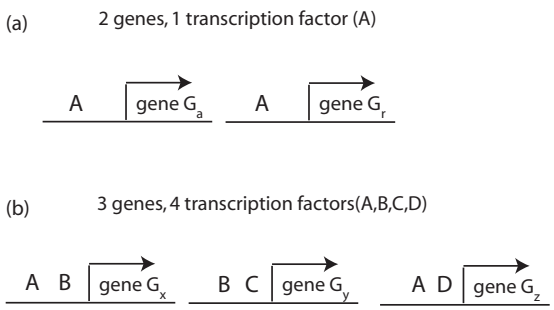

FIG. 1. Examples of gene modules with common transcription factors.

For a given gene module g_i, we define $s_{g_i}(k,t)$ as the array containing the molecular counts in cell k at time t of simultaneously measured transcription factors that affect g_i. Therefore, $s^T_{g_i}(k,t)$ refers to the molecular count measurement for the transcription factor T that affects the gene module. If Y is the subset of state variables describing the gene states, and $w_j(y)$, $j = 1, \ldots M_{g_i}$ are the propensity functions describing the binding and unbinding reactions that drive transitions between the gene states, then the time-evolution of the probability density function of gene states, given the single-cell time measurement $s_{g_i}(k,t)$ can be described by the following chemical master equation:

$$\frac{\partial P(y,t|s_{g_i}(k,t))}{\partial t} = \sum_{j=1}^{M_{g_i}} [w_j(y-v_j)P(y-v_j,t|s_{g_i}(k,t)) - w_j(y)P(y,t|s_{g_i}(k,t))]. \tag{8}$$

For each gene module, the states of the CME are all the possible bound and unbound configurations of the genes from all of the transcription factors involved. Usually, binding and unbinding reactions occur on a much faster time scale than the transcription factor dynamics. A commonly held assumption (one that is at the basis of the emergence of Michaelis-Menten and Hill Function[23]) is that states achieve a quasi-steady state relative to the transcription factor level dynamics. As a result,

$$\sum_{j=1}^{M_{g_i}} [w_j(y-v_j)P_{qs}(y-v_j|s_{g_i}(k,t)) - w_j(y)P_{qs}(y|s_{g_i}(k,t))] \cong 0. \tag{9}$$

A given single-cell flow-cytometry measurement samples a cell once at sample time t_m and then discards it. Here, $s_{g_i}(k,t_m)$ refers to the kth sampled single-cell flow-cytometry measurement at sample time t_m for the gene module g_i. The quasi-stationary chemical master equation for the gene states, given $s_{g_i}(k,t_m)$, is then given by

$$\sum_{j=1}^{M_{g_i}} [w_j(y-v_j)P_{qs}(y-v_j|s_{g_i}(k,t_m)) - w_j(y)P_{qs}(y|s_{g_i}(k,t_m))] = 0. \tag{10}$$

For every time t_m and cell k, this equation can now be solved to give $P_{qs}(y|s_{g_i}(k,t_m))$, the estimated probability of gene states in that cell given the measured level of the transcription factors impacting the gene. This, in turn, can be used to compute the mean of gene state i in cell k at time t_m as

$$\mu_i(k,t_m) = \sum_y y_i P_{qs}(y|s_{g_i}(k,t_m)) \tag{11}$$

and the mean of gene state i across the population at time t_m as

$$\mu_i(t_m) = \frac{1}{K}\sum_{k=1}^{K} \mu_i(k,t_m). \tag{12}$$

Here again, we are assuming measurements of transcription factor levels in K cells.

As before, the mean expression of every transcription factor T at time t_m is

$$\mu_T(t_m) = \frac{1}{K}\sum_{k=1}^{K} s_{g_i}^T(k,t_m). \tag{13}$$

The covariance of any transcription factor T and gene state i is then given by

$$C_{iT}(t_m) = \frac{1}{K}\sum_{k=1}^{K}[\mu_i(k,t_m) - \mu_i(t_m)]\left[s_{g_i}^T(k,t_m) - \mu_T(t_m)\right]. \tag{14}$$

These quantities can be input into the appropriate mRNA and protein equations as described in Sec. III A, which can then be solved. The integrated flow-chart for the overall moment closure framework is illustrated in Figure 2.

C. Solving for unknown parameters

If the parameters are known, then the data-driven moment closure approach will allow for the mean equations of the system to be numerically solved. In most cases, however, many of the parameters are unknown. It is usually the case that data are fit to models in order to determine these parameters, and the correctly parameterized model is then used to probe in-depth various aspects of the system's behavior. The moment closure methodology developed above could be systematically used in parameter identification. Specifically, for a given set of parameter values, the calculated means from Eq. (3) could be compared with the measured means to determine error values, and an iterative procedure initiated whereby the parameter space could be sampled to find regions in parameter space that satisfy the data within a given error tolerance.

Alternatively, examining Eq. (3), if the means, the appropriate covariances, and time-derivatives of the means are measured accurately over time for all variables in the model, then the unknown parameter set could be calculated directly (see Appendix A). Indeed similar observations for systems with affine propensities have previously been made.[24] However, it is in general the case that not all species in the model, especially the gene states, will be measurable. Furthermore, accurate time derivatives of the mean may not be achievable due to experimental noise in the data and/or sampling frequency limitations. Therefore, in the example below, we adopt a numerical iterative data-driven moment closure approach as a more practical method for searching the parameter space.

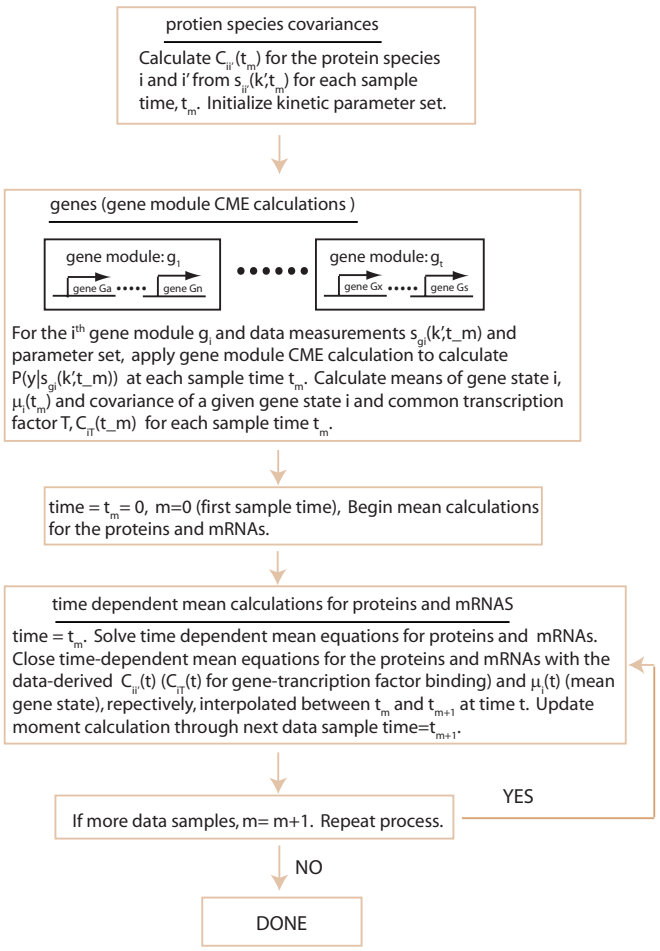

FIG. 2. Basic algorithm for the data-driven moment closure technique.

IV. NUMERICAL EXAMPLE: BIOLOGICAL OSCILLATOR

Delineating the functional principles of biological oscillators has been an issue of intense investigation for systems and synthetic biology. Oscillator design with nested positive and negative feedback loops has been actively pursued for synthetic biology because of their perceived robustness properties.[25] In this section, we analyze an *in silico* model for such a genetic oscillator circuit thought to approximate the essential functionality of a circadian oscillator.[26] The system is described by the following biochemical reactions:

$$D_a + A \underset{\theta_a}{\overset{\gamma_a A}{\rightleftharpoons}} D_a', \tag{15a}$$

$$D_r + A \underset{\theta_r}{\overset{\gamma_r A}{\rightleftharpoons}} D_r', \tag{15b}$$

$$\emptyset \xrightarrow{\alpha_a' D_a' + \alpha_a D_a} M_a \xrightarrow{\delta_{ma} M_a} \emptyset, \tag{15c}$$

$$\emptyset \xrightarrow{\alpha_r' D_r' + \alpha_r D_r} M_r \xrightarrow{\delta_{mr} M_r} \emptyset, \tag{15d}$$

$$\emptyset \xrightarrow{\beta_a M_a} A \xrightarrow{\delta_a A} \emptyset, \tag{15e}$$

$$\emptyset \xrightarrow{\beta_r M_r} R \xrightarrow{\delta_r R} \emptyset, \tag{15f}$$

$$A + R \xrightarrow{\alpha_c' AR} C \xrightarrow{\delta_a C} R. \tag{15g}$$

Briefly, transcription factor A positively regulates both its own gene G_a (with two gene states D_a and D_a') and the repressor protein gene, G_r (with two gene states D_r and D_r'). This and the inhibiting effect of the repressor protein R on A result in nested positive and negative feedback loops. Depending on the parameter values, the system can exhibit different behaviors. For some parameter sets, the system exhibits deterministic oscillations. In contrast, for other parameter sets, it exhibits noise-induced oscillations. In this numerical study, we focus on the more computationally challenging case where the system exhibits noise induced oscillations. For the parameter set which produces this behavior (Table I), A is produced at a faster rate than R, but has a shorter half-life. As R is produced, it negatively regulates A by combining with A to form the complex C. Eventually, the population of A is reduced to near zero, thereby down regulating the transcription of A and R. As the population of R decays toward zero, as long as the reactions are stochastic, the resulting fluctuations in the concentration of A and R enable the process to repeat (oscillate).

For these values of the parameters, the repressor protein R converges to a stable steady state (solid line in Figure 3(a)) according to the macroscopic reaction rate equations. However, using the SSA (Ref. 14) to account for the stochasticity in the reactions results in oscillations in R (dashed line in Figure 3(a)). To assess the applicability of the data-based moment closure technique, we generated 10 000 SSA runs (a typical number of cells measured in any flow cytometry run) that produce sample values of A, R, and C. As expected, averaging these SSA runs produces an average trajectory with damped oscillations due to stochastic fluctuations in the period of the individual oscillations. Assuming that gene states are quasi-stationary (i.e., Eq. (10) applies), we used A, R, and C and their pairwise combinations to estimate gene states and covariances needed to update the mean equations. The quasi-stationary CME calculation for this gene module is derived in Appendix B (the gene module from Figure 1(b) is also discussed).

Values for mean R and mRNA R resulting from the SSA and the data-driven moment closure calculations are plotted in Figures 3(b) and 3(c), showing excellent agreement between the two methods. The necessity to account *correctly* for the higher order moments is underscored by the fact that solving for the first two central moments and setting the third central moment to zero as prescribed in Ref. 19 clearly gives an

TABLE I. Parameters for oscillator example.

Parameter :	α_A	α_A'	α_R	α_R'	β_A	β_R	δ_{MA}	δ_{MR}
Value :	50	500	0.01	50	50	5	10	0.5
Parameter :	δ_A	δ_R	γ_A	γ_R	γ_C	θ_A	θ_R	
Value :	1	0.05	1	1	2	50	100	

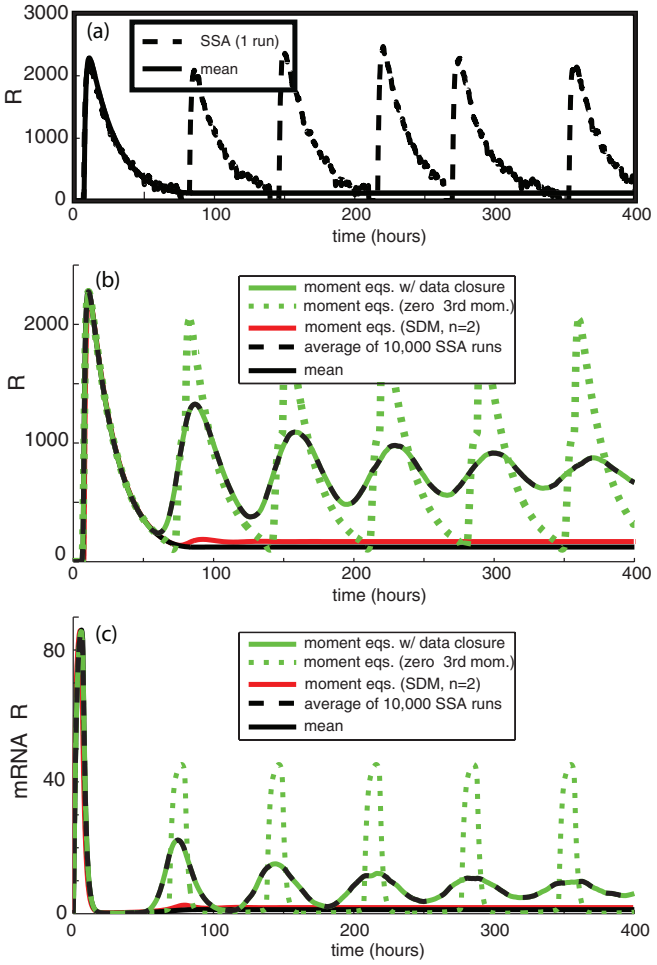

FIG. 3. Simulation results for the oscillator example (a) deterministic mean trajectory of the repressor R versus a sample stochastic trajectory generated by the SSA. (b) Mean trajectories of the repressor R computed from the average of 10 000 SSA runs (black, dashed), moment closure with data (green, solid), moment closure with zero third moment (green, dashed), separable derivative matching moment closure at n = 2 (red, solid), and deterministic mean trajectory (black, solid). (c) Mean trajectories of the mRNA for repressor R computed as in (b).

erroneous answer (Figures 3(b) and 3(c), moment equations (zero third moment)). In fact, ignoring the contribution of the third central moment results in sustained oscillations even in the average trajectory. Furthermore, accounting for the third (un-centered) moment using the separable derivative matching technique[20] (n = 2, i.e., third moment closure approximation (log-normal assumption)) also yields inaccurate results by showing average trajectories with overly damped oscillations (Figures 3(b) and 3(c), moment equations (SDM, n = 2)). This is likely due to the true distribution deviating from log-normal. Obviously, the error in the SDM is expected to decrease by increasing the moment order n. However, the rate of conversion to the correct solution is not known *a priori*. In contrast, our method circumvents this moment truncation error and only requires a quasi-stationarity assumption which could ideally be checked to hold before any calculations are attempted.

To further dissect the contributions of the covariance terms in Eq. (3), we addressed three cases. In Case 1, we

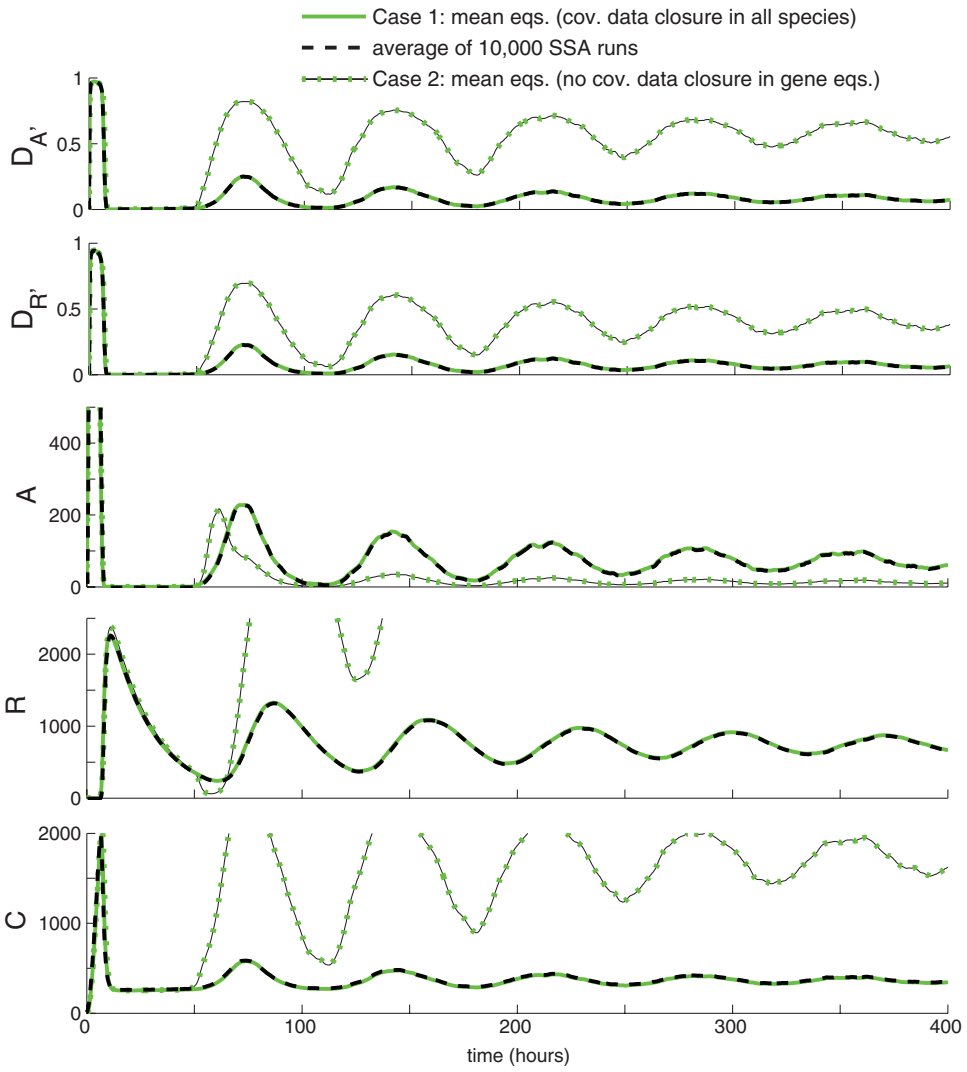

FIG. 4. Mean of the oscillator model variables plotted for 10 000 SSA runs. The results are also shown for the case where data-driven model closure was done with appropriate covariance calculations using data for all species, and the case where closure was done using covariance data for all species except for gene states whose covariance was set to zero.

assume the availability of direct measurements of gene states and solve Eq. (3) for all species (including the gene states) with data-derived covariance closure. Although genes states are not available for direct measurement with current technology, this case illustrates a best case scenario for the closure methodology and is instrumental in verifying that proper mean calculations are obtained when data-driven closure is available for all species. In Case 2, we solve Eq. (3) without covariances (set to zero) for the genes, but with correct covariances for the proteins/mRNAs in order to assess the contribution of gene-state fluctuations to the solution. In contrast, in Case 3, we solve Eq. (3) with covariances for the gene states but without covariances (set to zero) for the proteins/mRNAs. Theoretically, Case 1 should be identical to the SSA results since solving (3) with proper covariance closure is exact for any species in the model. Indeed this is shown to be the case in Figure 4 where the results from Case 1 and the results from the average of 10 000 SSA runs agree perfectly. In Case 2 where the gene covariances are ignored, erroneous time evolution of the model species result (Figure 4). The results for Case 3 (plotted in Figure 5) also diverge from the exact results of Case 1 and the SSA. The largest error is in the time response of the protein complex *C*. However, the total error for Case 3 is smaller than that for Case 2, likely because the gene states for Case 3 are calculated correctly. In contrast, in Case 2, the gene states have large errors in the absence of correct covariance terms, and this error is amplified as it propagates through the protein calculations. These different cases further illustrate the nontrivial contributions of higher order moments, and the magnitude of errors that can accumulate, if moment equations for all model species are not considered carefully.

V. PARAMETER IDENTIFICATION USING DATA-DRIVEN MOMENT CLOSURE

Next, we demonstrate the use of an iterative search procedure combined with the data-driven moment closure

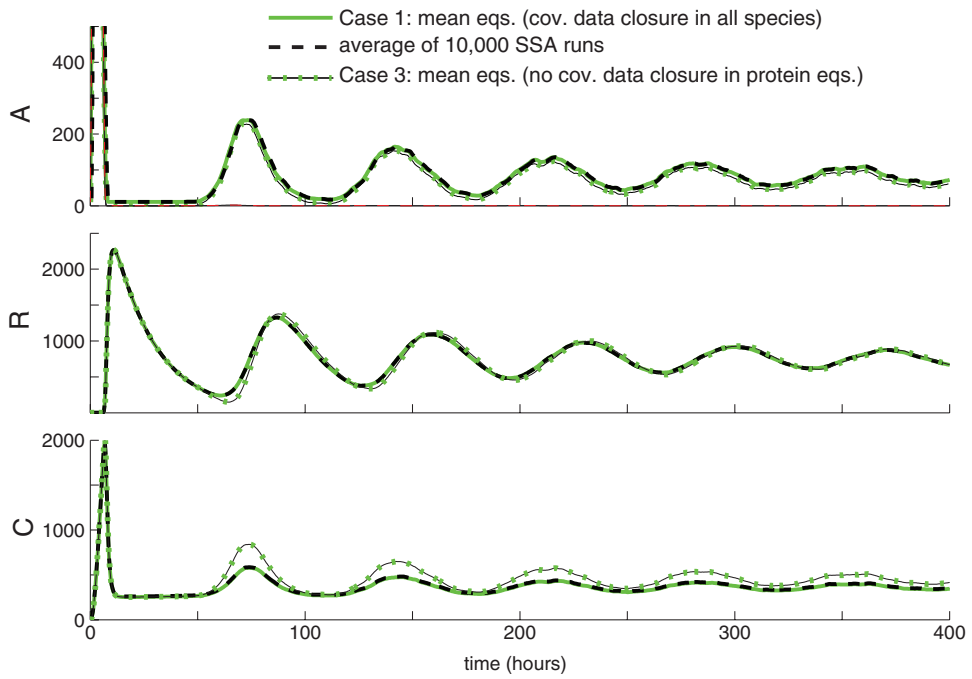

FIG. 5. Mean of the oscillator model variables plotted for 10 000 SSA runs. The results are also shown for the case where data-driven model closure was done with appropriate covariance calculations using data for all species, and the case where closure was done using covariance data for all species except for protein species whose covariance was set to zero.

to find parameter sets that fit the oscillator model to data. As in any iterative procedure for parameter fitting, an initial estimate for the parameters should be provided. To do so, we used a least-squares approach (Appendix A). Using the SSA data, we took 21 evenly spaced samples during the time period extending from 3.8 to 80 h. We also collected data at $\pm$.4 h about each sample time point. Theses data were used to approximate the necessary time derivatives of the mean using a finite difference approach. In order to compute the gene-state mean and covariance values from transcription factor (A) data, we assumed that the gene module rate constants θ_A, θ_R, γ_A, γ_R are known (with values given in Table I). The gene-state mean and covariance, along with the measured means, time derivatives of the mean, and covariances of the other molecular species were then input into the time-dependent mRNA and protein mean equations. We then used these equations in combination with a least squares procedure to obtain an initial parameter set estimate for the remaining parameters (resulting parameters are shown in Table II). This parameter set was of mixed quality with respect to fit to the data (Figures 6(a) and 6(b)): the mRNA results (the R mRNA shown as an example) agree well with the SSA, while the proteins do not (repressor protein R shown). This is not unexpected given the fairly coarse samples used to approximate the time derivatives. However, the parameter set computed in this way constitutes a good starting "seed" for an iterative algorithm for parameter refinement. With this parameter set (Table II), we initiated an iterative steepest descent search algorithm coupled with data-driven moment closure to find a parameter set that improves the fitting error. Since in the moment closure scheme, the mRNA rate constants do not appear in the time-dependent protein equations for R, A, and C, the values of the mRNA parameters do not need to be further fitted. Therefore, we only applied the iterative search to the parameters in the protein mean equations. Interestingly, the resulting parameters (Table III, model results in Figure 6(b)) generate an excellent fit to the data while being different from the original set in Table I. This is a feature of model under-determination where it is a local volume of the parameter space, rather than a single parameter set, that can recapitulate a model's fit to data within some given error tolerance ϵ. To validate this insight, we sampled directions in parameter space that connect parameters in Table III to the parameters in Table I and found them all to be viable. Since model under-determination is a common feature of biological models, our data-driven moment closure approach could be an efficient tool to characterize such local data-equivalent volumes of parameter space and to analyze their structure as a way to gain insight into parameter dependencies.

TABLE II. Parameters inferred through least squares inversion of data and data-driven calculations. (Parameters in bold were not solved for.)

Parameter :	α_A	α_A'	α_R	α_R'	β_A	β_R	δ_{MA}	δ_{MR}
Value :	62.1	622	0.0049	48.9	51.4	4.77	12.4	0.488
Parameter :	δ_A	δ_R	γ_A	γ_R	γ_C	θ_A	θ_R	
Value :	1.12	0.076	**1**	**1**	1.74	**50**	**100**	

VI. SUMMARY AND FUTURE WORK

In this work, we have presented a novel data integrated hybrid CME/moment method for modeling and analysis of stochastic biological networks. This approach systematically

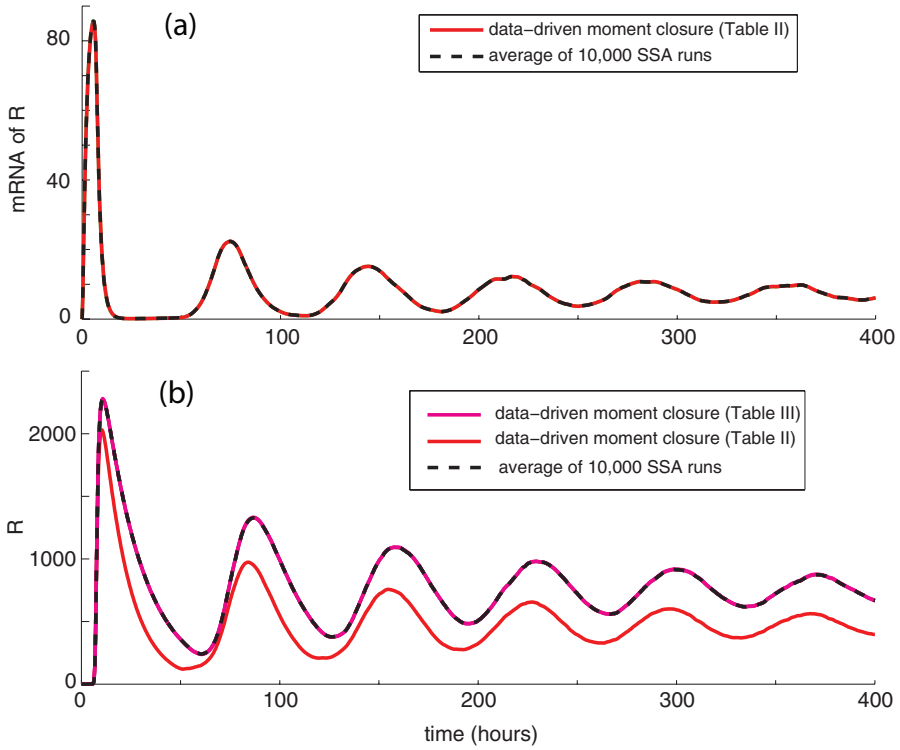

FIG. 6. Mean of R mRNA (a) and protein R (b) for the SSA (black, dotted) with initial parameters (Table I), parameters derived using least squares inversion (Table II, red) and the iterative parameter search (Table III, pink). For the last two cases, simulations results are reported for the mean equations using moment closure.

exploits cell-to-cell variability information without making any assumptions about the underlying distributions of the molecular species or the magnitude of their higher order moments. Integral to the methodology is the careful assessment of gene states, which are often unmeasurable. To circumvent this difficulty, we developed a simple data-driven approach to estimate the mean gene-states and gene-state/transcription factor covariances using measurable transcription factor data. We anticipate that these calculations can be easily extended to estimate different protein states that might not be easily measurable such as post-translational modifications, or proteins present as part of a molecular complex.

As argued through the oscillator example, the data-driven model closure method is particularly suited for use in parameter identification. An immediate area of application is the characterization of the rate constants in synthetic cellular circuits, since the structure of the network is already known by design. In this context, a practical approach would be to first characterize the gene modules and determine their parameters in independent experiments. In this way, by solving for subsets of the circuits parameter space, one can solve for the parameters more efficiently. This can be accomplished by measuring the input-output function of a gene module as its transcription factors are varying. This could be done by exerting experimental control over the concentration of the transcription factors (by placing them under the control of an inducible promoter), and measuring the stationary distribution of the promoter's output. We emphasize that both sufficient time sampling and cell numbers per sample are required to both accurately capture the system's time scales and measured means and covariances. After the kinetic parameters for gene modules are characterized, one can then obtain measurements from the fully implemented circuit and use the iterative search procedure with the data-driven moment closure discussed in this paper to characterize the remaining parameters. This step-by-step rational approach of characterizing the parts (genes) and then calculating the fully realized circuit sits at the heart of the engineer's approach to synthetic biology. Here, it has the additional benefit of being an integral part of a streamlined approach for analyzing stochastic synthetic circuits while properly accounting for the variability contained in single-cell data.

TABLE III. Parameters found through iterative search using data-driven moment closure within initial parameters taken from Table II. (Parameters in bold were not solved for.)

Parameter :	α_A	α_A'	α_R	α_R'	β_A	β_R	δ_{MA}	δ_{MR}
Value :	**62.1**	**622**	**0.0049**	**48.9**	55.3	5.02	**12.4**	**0.488**
Parameter :	δ_A	δ_R	γ_A	γ_R	γ_C	θ_A	θ_R	
Value :	1.11	0.05	**1**	**1**	1.67	**50**	**100**	

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Grant Nos. CCF-0943385 and NCI U54 CA143836 to H.E. We would like to thank members of the El-Samad lab,

especially Jacob Stewart-Ornstein and Charles Biddle-Snead for critical reading of the paper.

APPENDIX A: INFERRING THE RATE CONSTANTS FROM COMPLETE DATA

If means, covariances, and time-derivatives of the means are available, then the problem of inverting the rate constants is linear. If the system has at most bimolecular reactions, then the jth reaction propensity takes the form $a_j(x) = c_j f_j(x)$ where c_j is a constant and $f_j(x) \epsilon \{1, x_i, x_i(x_i - 1)/2!, x_i x_k\}$ then Eq. (3) can be written as

$$\frac{\partial z_k(t)}{\partial t} = \sum_{j=1}^{M} c_j \nu_{jk} \left(f_j(z(t)) + \sum_{i,i'}^{N} \frac{\partial^2 f_j(z(t))}{\partial x_i \partial x_{i'}} |_{x=z(t)} \frac{C_{ii'(t)}}{2} \right) \tag{A1}$$

which is linear in c_j. Let us also assume that we have S distinct sample times at which the means, their-time derivatives, and covariances are measured. Let c be the $M \times 1$ vector of kinetic parameters. Given N species and M reactions, let b be the $N \cdot S$ vector whose elements are $b_i = \partial z_k(t(s))/\partial t$, where $i = (s - 1)*N + k$ is the index for the measurement of the time-dependent mean equation for the kth species at the time sample s. Similarly, the elements of the $(N \cdot S) \times M$ matrix A can be defined as $A_{ij} = \nu_{jk}(f_j(z(t(s))) + \sum_{i,i'}^{N} \frac{\partial^2 f_j(z(t_s))}{\partial x_i \partial x_{i'}} |_{x=z(t)} \frac{C_{ii'(t)}}{2})$. Using this notation, the parameters c_i can be determined as the solution of the matrix problem of the form $Ac = b$. If $N \cdot S \geq M$ and A is full rank, then c may be solved for using convex optimization approaches, including least squares. The mathematical structure of this general system is similar to the work of Munsky and Khammash[24] who addressed systems of moment equations with affine propensities.

However, as discussed above, not all species in the model, such as the gene states, will be measurable. Also, due to experimental noise and sampling capabilities, accurate time derivative measurements of the mean may not be achievable.

APPENDIX B: EXAMPLE GENE CALCULATION

Here, we present an example calculation for the gene module from the numerical example of the genetic oscillator (also represented in Figure 1(a)). Consider a system with two genes, gene G_a and gene G_r, where gene G_a can be in state D_a or $D_a^{'}$ and gene G_r can be in state D_r or $D_r^{'}$. The state transitions are defined by the biomolecular reactions (15a) and (15b). For this gene module, we describe and present below the explicit stationary master equation. Since the states of the two genes are coupled through a common transcription factor, A, and the gene module we consider contains gene G_a and gene G_r, we construct a CME that contains both genes. We define the states of the CME as $[D_a, D_a^{'}, D_r, D_r^{'}]$, for which there are four states: [1, 0, 1, 0], [0, 1, 1, 0][1, 0, 0, 1][0, 1, 0, 1]. At some sample time, t_m, for the kth single-cell transcription factor measurement, A_k, the probability of being in state $i = 1, 2, 3, 4$ is simply P_{i_k}. The stationary master equation is

$$\begin{pmatrix} E & \theta_a & \theta_r & 0 \\ \gamma_a A_k & F & 0 & \theta_r \\ \gamma_a A_k & 0 & G & \theta_a \\ 0 & \gamma_r(A_k - 1) & \gamma_a(A_k - 1) & H \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} P_{1_k} \\ P_{2_k} \\ P_{3_k} \\ P_{4_k} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \tag{B1}$$

where $E = -\gamma_a A_k - \gamma_r A_k$, $F = -\theta_a - \gamma_r(A_k - 1)$, $G = -\theta_r - \gamma_a(A_k - 1)$, and $H = -\theta_a - \theta_r$. The first four rows represent the CME state transition matrix and the last row enforces that the total probability adds up to 1. Note for a given measurement, not all states may be accessible. For example, if $A_k = 0$, then P_{1_k} is the only state. If $A_k = 1$, then only $P_{1_k}, P_{2_k}, P_{3_k}$ are accessible. The CME state transition matrix must be constructed accordingly to ensure invertibility.

We now calculate a few example quantities. The mean value for D_a for the kth measurement is

$$\mu_{D_a}(k, t_m) = 1^* P_{1_k} + 0^* P_{2_k} + 1^* P_{3_k} + 0^* P_{4_k} \tag{B2}$$

and

$$\mu_{D_a}(t_m) = \frac{1}{K} \sum_{k=1}^{K} \mu_{D_a}(k, t_m) \tag{B3}$$

and

$$\mu_{A}(t_m) = \frac{1}{K} \sum_{k=1}^{K} A_{k'}. \tag{B4}$$

The covariance of D_a and A is given by

$$C_{D_a A}(t_m) = \frac{1}{K} \sum_{k=1}^{K} [\mu_{D_a}(k, t_m) - \mu_{D_a}(t_m)][A_k - \mu_A(t_m)]. \tag{B5}$$

The above means and covariances calculated from this would then be input into the protein and mRNA equations for the numerical example of the genetic oscillator.

We briefly outline how to setup the calculation for the states of gene G_x from the example in Figure 1(b). Given single-cell simultaneous measurements of A and B, we need to solve the states of gene G_x while including the sequestration effects of B and A binding to gene G_y and gene G_z. Gene G_x can be in one of the four states. Gene G_y can be either unbound or bound by transcription factor B, and gene G_z can be either bound or unbound by transcription factor A. Therefore, the stationary CME to solve for gene states would have a total of 16 states for all possible combinations of the gene states. Setting up the CME for this system would follow the same procedure as was done for the example in Figure 1(a). Similar gene modules for each of the other two genes (gene G_y and gene G_z) could be setup in the same manner.

[1]T. Gregor, E. Wieschaus, A. McGregor, W. Bialek, and D. Tank, Cell **130**, 141 (2007).
[2]T. Gregor, D. Tank, E. Wieschaus, and W. Bialek, Cell **130**, 153 (2007).
[3]A. Raj, S. Rifkin, E. Andersen, and A. van Oudenaarden, Nature (London) **463**, 18 (2010).
[4]N. Balaban, J. Merrin, R. Chait, L. Kowalik, and S. Leibler, Science **305**, 1622 (2004).
[5]M. Kaern, T. Elston, W. Blake, and J. Collins, Nat. Rev. Genet. **6**, 451 (2005).
[6]H. McAdams and A. Arkin, Proc. Natl. Acad. Sci. U.S.A. **94**, 814 (1997).
[7]T. Kepler and T. Elston, J. Biophys. **81**, 3116 (2001).
[8]P. Swain, M. Elowitz, and E. Siggia, Proc. Natl. Acad. Sci. U.S.A. **99**, 12795 (2002).
[9]D. A. Drubin, J. C. Way, and P. A. Silver, Genes Dev. **21**, 254 (2007).
[10]J. Paulsson, Phys. Life Rev. **2**, 157 (2005).
[11]D. McQuarrie, J. Appl. Probab. **4**, 413 (1967).
[12]D. Gillespie, Physica A. **408**, 404 (1992).
[13]N. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1992).
[14]D. Gillespie, J. Comput. Phys. **22**, 403 (1976).
[15]B. Munsky and M. Khammash, J. Chem. Phys. **124**, 044104 (2006).
[16]B. Munsky, B. Trinh, and M. Khammash, Mol. Syst. Biol. **5**, 318 (2009).
[17]D. Gillespie, J. Chem. Phys. **113**, 297 (2000).
[18]S. Engblom, Appl. Math. Comput. **180**, 498 (2006).
[19]C. Gomez-Uribe and G. Verghese, J. Chem. Phys. **126**, 24109 (2007).
[20]A. Singh and J. Hespanha, Bull. Math. Biol. **69**, 1909 (2007).
[21]K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan, Science **308**, 523 (2005).
[22]A. Raj, P. van den Boogard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi, Nat. Methods **5**(10), 877 (2008).
[23]Y. Cao, D. T. Gillespie, and L. R. Petzold, J. Chem. Phys. **122**, 014116 (2005).
[24]B. Munsky and M. Khammash, in *Proceedings of the 47th Conference on Decision and Control*, Cancun, Mexico, 2008.
[25]J. Stricker, S. Cookson, M. Bennett, W. Mather, L. Tsimring, and J. Hasty, Nature (London) **456**(7221), 516 (2008).
[26]J. Vilar, H. Kueh, N. Barkai, and S. Leibler, Proc. Natl. Acad. Sci. U.S.A. **99**, 5988 (2002).