Published in final edited form as:
Quant Biol. 2014 March ; 2(1): 1–29. doi:10.1007/s40484-014-0025-7.

Modeling stochastic noise in gene regulatory systems

Arwen Meister, **Chao Du**, **Ye Henry Li**, and **Wing Hung Wong**[*]
Computational Biology Lab, Bio-X Program, Stanford University, Stanford, CA 94305, USA

Abstract

The Master equation is considered the gold standard for modeling the stochastic mechanisms of gene regulation in molecular detail, but it is too complex to solve exactly in most cases, so approximation and simulation methods are essential. However, there is still a lack of consensus about the best way to carry these out. To help clarify the situation, we review Master equation models of gene regulation, theoretical approximations based on an expansion method due to N.G. van Kampen and R. Kubo, and simulation algorithms due to D.T. Gillespie and P. Langevin. Expansion of the Master equation shows that for systems with a single stable steady-state, the stochastic model reduces to a deterministic model in a first-order approximation. Additional theory, also due to van Kampen, describes the asymptotic behavior of multistable systems. To support and illustrate the theory and provide further insight into the complex behavior of multistable systems, we perform a detailed simulation study comparing the various approximation and simulation methods applied to synthetic gene regulatory systems with various qualitative characteristics. The simulation studies show that for large stochastic systems with a single steady-state, deterministic models are quite accurate, since the probability distribution of the solution has a single peak tracking the deterministic trajectory whose variance is inversely proportional to the system size. In multistable stochastic systems, large fluctuations can cause individual trajectories to escape from the domain of attraction of one steady-state and be attracted to another, so the system eventually reaches a multimodal probability distribution in which all stable steady-states are represented proportional to their relative stability. However, since the escape time scales exponentially with system size, this process can take a very long time in large systems.

Keywords

gene regulation; stochastic modeling; simulation; Master equation; Gillespie algorithm; Langevin equation

INTRODUCTION

Like any physical quantity, gene expression level measurements are subject to noise. In fact, the experimental techniques for measuring biological quantities tend to be much noisier than measurements in other scientific disciplines. The extrinsic noise arising from measurement

[*] Correspondence: whwong@stanford.edu.

CONFLICT OF INTEREST
The authors Arwen Meister, Chao Du, Ye Henry Li and Wing Hung Wong declare that they have no conflict of interests.

error can be mitigated by averaging many experimental replicates. However, in addition to straightforward extrinsic noise, gene expression is also characterized by intrinsic noise arising from the fundamental stochasticity of the underlying processes, which cannot be simply averaged away [1,2].

Experimental studies using single-cell biotechnologies have revealed that the biological mechanisms of gene regulation, including promoter activation and deactivation, transcription, translation, and degradation, are inherently stochastic [3–10]. Stochasticity can sometimes dramatically affect the behavior of gene regulatory networks [8,11,12] as stochasticity leads to different phase diagrams and can cause instability [13], and small molecular numbers can seriously impact system behavior [14]. These observations have important implications in synthetic biology, including the engineering of switches, feedback loops, and oscillatory systems [15–18].

A number of stochastic models have been applied to the gene regulation problem, since it is clear that in many cases, additive noise (independent of expression level) is not sufficient [19]. Since gene regulation depends on a series of chemical reactions (including the binding of TFs and RNAP to the promoter, transcription, translation, and degradation), it can be modeled with chemical equations [20,21]. A number of ad hoc approaches have also shown promise, including Poisson models (although expression regulation can lead to non-constant rates, disrupting the Poisson character), fluctuation noise analysis in small systems [4,22–26], and structural inference on large networks based on noise correlation [27,28]. The highest quantitative resolution is obtained from more sophisticated analysis based on the Master equation [29], including approximation methods [25,30–33], and more accurate modeling via simulation or theoretical deduction [14,19,34–37].

While the Master equation is generally accepted as the gold standard for modeling the processes of gene regulation in molecular detail, it is too complex to solve exactly except in simple cases. Approximations are needed to make it useful, but researchers have still not reached a clear consensus about the proper way to carry them out. In this paper, we attempt to shed light on these issues by discussing theoretical approximations of the Master equation based on an expansion method due to N. G. van Kampen and R. Kubo, and simulation algorithms due to Gillespie and Langevin. The van Kampen expansion shows that the stochastic model reduces to a deterministic model in a first-order approximation, provided the system has a single stable steady-state. We also discuss additional theory due to van Kampen for modeling the complex behavior of stochastic systems with multiple stable steady-states. To illustrate the methods and provide further insight into the behavior of multistable system, we perform a detailed simulation study comparing the various approximation and simulation methods applied to synthetic gene regulatory systems with a variety of qualitative characteristics.

DYNAMICAL SYSTEM MODELS OF GENE REGULATION

Deterministic dynamical system models of gene regulation lay the groundwork for stochastic modeling efforts. Master equation models are discretized stochastic generalizations of dynamical system models, so we start by introducing these basic models to provide context and clarify the basic modeling assumptions. For both dynamical system models and their stochastic counter-parts, the form of the RNA polymerase binding probability function is a key modeling choice, and researchers have considered many different possibilities. Linear functions yield the simplest models, but lead to dynamical systems (or stochastic systems) with only one steady-state, while nonlinear choices yield much more complex systems with multiple steady-states, like many of the most interesting biological systems. Lyapunov theory characterizes the stability of steady-states of a dynamical system and hence the system's long term behavior; multi-stability will become an even more interesting and critical issue when stochasticity comes into play. In this section, we discuss dynamical system models, linear and nonlinear choices of binding probability function, and steady-states and their stability, in order to set the stage for stochastic models of gene regulation.

Dynamical system model

Before formulating the Master equation model for gene regulation, we introduce its natural precursor, a deterministic dynamical system model. Although dynamical system models do not account for intrinsic noise, they capture cells' ability to assume different characters as they transition through their lifecycles and respond to stimuli, and their quantitative form means that they can be used to predict systems' future behavior. They also lend themselves to inference algorithms that allow the structure of novel gene regulatory systems to be learned from data [38]. As we will see in a later section, the stochastic model reduces to the deterministic model in a first-order approximation.

In the standard dynamical system model of gene regulation, the levels of RNA and protein evolve according to a system of differential equations. The basic assumption is that each species of RNA is transcribed at a rate proportional to the probability of RNA polymerase (RNAP) binding to the gene promoter (as a function of the expression levels of the transcription factor (TF) proteins that regulate it) and degrades at a rate proportional to its current level, while the corresponding protein is translated at a rate proportional to the current RNA level and also degrades at a rate proportional to its own level.

Hence, the dynamical system model is given by

$$\begin{array}{c} \frac{\mathrm{d}x_i}{\mathrm{d}t} = \tau_i f_i(y) - \gamma_i^{\mathrm{r}} x_i \\ \frac{\mathrm{d}y_i}{\mathrm{d}t} = \rho_i x_i - \gamma_i^{\mathrm{P}} y_i \end{array} \quad (1)$$

where $x \in \mathbb{R}^n$ represents RNA concentrations and $y \in \mathbb{R}^n$ represents protein concentrations corresponding to a set of n genes, $f(y)$ is the probability that RNAP is bound to the promoter as a function of the concentrations of regulatory proteins, τ_i is the transcription rate when RNAP is bound to promoter, ρ_i is the translation rate, and $\gamma_i^{\mathrm{r}}, \gamma_i^{\mathrm{P}}$ are the RNA and protein degradation rates, respectively.

In some situations, it is necessary or more appropriate to ignore the distinction between RNA and protein and use a model of the form:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}=\tau_i f_i(x)-\gamma_i x_i \quad (2)$$

involving only the RNA concentrations *x*, which serve as a surrogate for the protein concentrations *y*. We will make similar modeling assumptions when we apply the Master equation to the gene regulation problem.

RNAP binding probability models

In Equation (1), the term *f*(*y*) represents the probability of RNAP binding to the gene promoter as a function of the protein concentrations *y*. There are many possible approaches to the modeling of the RNAP binding probability, yielding different functional forms of *f*.

The simplest approach is to use a linear function *f*(*x*) = *Ax*, where $x \in \mathbb{Z}^n$, $f : \mathbb{R}^n \to \mathbb{R}^n$, and *A* is an *n*×*n* matrix. Such a system can have only one steady-state, as we will discuss in more detail later on. Each coefficient a_{ij} represents the linear regulatory effect of gene *j* on gene *i*, and it is clear how perturbing the system and observing its response would allow for straightforward inference of these coefficients. If we are only interested in one specific steady-state of a biological system, this may be a good choice, and has led to success in many cases [39–42]. However, the usefulness of the linear model is severely limited by the fact that it cannot capture the ability of gene regulatory systems to maintain multiple stable steady-states, one of the key features that often motivate their study. Hence, we now turn our attention to nonlinear models.

Michaelis-Menten kinetics and the Hill equation are classical nonlinear model for activation or repression by a single factor, based on thermodynamic theory. Michaelis-Menten kinetics [43] can be applied to gene regulation by a single transcription factor by modeling transcription as the enzymatic reaction series

$$\begin{aligned} X+D_0 &\overset{k_{\pm1}}{\longleftrightarrow} D_1 \\ D_1+P &\xrightarrow{k_t} D_1+P+Y \end{aligned}$$

where *X* is an activator, D_0 is an unbound promoter, D_1 is an activator-bound promoter, *P* is an RNA polymerase, and *Y* is an RNA transcript. The corresponding kinetic equations are:

$$\frac{\mathrm{d}D_1}{\mathrm{d}t}=k_1 D_0 X-k_{-1}D_1 \quad (3)$$

$$\frac{\mathrm{d}Y}{\mathrm{d}t}=k_t P D_1 \quad (4)$$

Let us assume that the reversible TF-promoter binding and RNAP-promoter binding reactions occur much faster than gene transcription, so that the quasi-steady-state assumption

$$\frac{\mathrm{d}D_1}{\mathrm{d}t}=0$$

approximately holds. That is, the bound and unbound promoter states maintain equilibrium concentrations. Then we can rearrange to obtain

$$\begin{gathered}\tfrac{\mathrm{d}Y}{\mathrm{d}t}=k_t P\tfrac{D_T X}{K_1+X},\\ \text{where } K_1=\tfrac{k_{-1}}{k_1},\ D_T=D_0+D_1\end{gathered}$$

A model for gene repression can be derived in a similar manner by replacing Equation (4) with

$$D_0+P \xrightarrow{k_t} D_0+P+Y$$

(since now transcription only occurs for the unbound promoter), leading to

$$\begin{gathered}\tfrac{\mathrm{d}Y}{\mathrm{d}t}=k_t P\tfrac{D_T K_1}{K_1+X},\\ \text{where } K_1=\tfrac{k_{-1}}{k_1},\ D_T=D_0+D_1\end{gathered}$$

The Hill equation [44] is another classical model with a similar form, which models cooperative binding. For activation by a single transcription factor, for example, it has the form

$$\begin{gathered}\tfrac{\mathrm{d}Y}{\mathrm{d}t}=\tfrac{Z}{1+Z},\\ \text{where } Z=\left(\tfrac{X}{K_1}\right)^n,\end{gathered}$$

where n is the Hill coefficient.

More sophisticated thermodynamic models can account for multiple regulatory mechanisms [45,46]. The Bintu et al formulation is most general thermodynamic model in common use, capable of capturing the full spectrum of network interactions [47,48]. The Bintu RNAP binding probability function takes the general form

$$f_i(y)=\frac{b_{i_0}+\sum_{j=1}^{m} b_{ij}\Pi_{k\in S_{ij}} y_k}{1+\sum_{j=1}^{m} c_{ij}\Pi_{k\in S_{ij}} y_k} \qquad (5)$$

where S_{ij} lists the gene products that interact to form a regulatory complex, and b_{ij}, c_{ij} are nonnegative coefficients satisfying $c_{ij} \geq b_{ij} \geq 0$, which depend on the binding energies of regulator complexes to the promoter. b_{i_0} and c_{i_0} correspond to the case when the promoter is not bound by any regulator ($\Pi_{k \in S_{i0}} y_k = 0$), and the coefficients are normalized so that c_{i0} = 1.

The form of f_i can model many different types of regulation in quantitative detail. Terms that appear in the denominator only are repressors, and the degree of repression depends on the magnitude of the coefficient, while terms that appear in the numerator and denominator may act as either activators or repressors depending on the relative magnitudes of the coefficients and the current gene expression levels.

Despite its detail and generality, not even the Bintu et al model captures the full complexity of the situation. One of the key assumptions implicit in the Bintu model is that RNAP levels are approximately constant on the time scale of interest. A second assumption is that nonspecific binding energy is equal for all non-promoter locations the RNAP could occupy. Finally, the model omits several reversible intermediate reactions such as binding and unbinding of RNAP and TFs. Since these reactions typically occur very quickly relative to the transcription time scale, we can reasonably assume that the quantities involved are in thermodynamic steady-state. Hence we can apply the quasi-steady-state assumption to eliminate the reversible reactions from the model. (The same argument applies to the stochastic formulation of the next section, as Rao and Arkin demonstrate how to use the quasi-steady-state assumption to eliminate intermediate species from a multivariate Master equation [49].) Nevertheless, eliminating short lived intermediate species is an approximation.

Steady-states and stability

The ability of gene regulatory networks to maintain multiple distinct steady-states is one of their most crucial properties and a major motivation for their study. Since steady-states and stability will be central to our discussion, we need to establish the necessary mathematical groundwork. Fortunately, the classical theory for general dynamical systems due to A. Lyapunov (1857–1918) perfectly serves our purposes. We will only outline the key definitions and theorems here; for a complete discussion, see a text like Walker's "Dynamical systems and evolution equations" [50].

Consider a general nonlinear dynamical system of the form

$$\dot{x}\,(t)=f(x(t)) \qquad (6)$$

where $x(t) \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}^n$, and f is continuous. Assume that this system has an equilibrium point x_e, i.e., $f(x_e) = 0$. If f is linear or affine, $f(x_e) = 0$ has exactly one solution so the system has exactly one steady-state; if f is nonlinear, the system may have zero, one, or multiple steady-states. (We must therefore use nonlinear functions to model gene regulatory systems if we wish to capture their ability to maintain multiple stable steady-states.) Let $\phi(t;\overline{x})$ denote the unique solution trajectory $x(t)$ corresponding to $x(0)=\overline{x}$. Lyapunov defined *stability* and a stronger condition, asymptotic stability, as follows:

Definition—The equilibrium x_e is said to be stable if for all $\varepsilon > 0$ there exists $\delta > 0$, such that

$$\overline{x} \in B(x_e, \delta) \Rightarrow \phi(t;\overline{x}) \in B(x_e, \varepsilon), \\ \text{for all}\, t \geq 0,$$

(where $B(x, \varepsilon)$ denotes the open ball of radius ε centered at x).

It is said to be asymptotically stable if it is stable, and for all $\varepsilon > 0$ there exists $\delta > 0$, such that

$$\overline{x} \in B(x_e, \delta) \Rightarrow \lim_{t\to\infty} \phi(t;\overline{x})=x_e$$

In essence, stability means that there exists a neighborhood of the steady-state such that trajectories that start inside that neighborhood remain there for all time. Asymptotical stability means that in addition to this, nearby trajectories are attracted to the steady-state and eventually get infinitely close to it. The next theorem provides the essential stability criterion.

Theorem—Assume $f \in C^1(\mathbb{R}^n)$, and set $A=\frac{\partial f}{\partial x}(x_e)$ (the Jacobian matrix at x_e). If there exists a symmetric positive definite matrix P such that $A^TP = PA \prec 0$, then x_e is asymptotically stable [50].

THE MASTER EQUATION

The basic approach to treating the stochasticity of gene regulation is to model gene expression level as a Markov process, whose future state depends probabilistically only on the current state. This is the most appropriate description for most processes in physics and chemistry [29], and this case is no exception: the mechanisms of transcription, translation, and degradation mean that the probability of each of these events depends only on the current quantity of each of the species involved in these processes, including RNAP, transcription factors, and ribosomes (each of which is the product of one or more genes, and can therefore be accounted for in our formulation). The Master equation is the natural model for gene regulation under the Markov assumption.

In this section we introduce the Master equation, which follows directly from the Markov property [29]. We will briefly discuss the general form of the Master equation, then turn our attention to the special case of birth-and-death processes, which provide an excellent model for gene regulation. We will need the multivariate form of the Master equation to treat systems with multiple genes. In the next subsection, we will apply the basic general theory outlined in this subsection to the gene regulation problem.

Markov processes

A Markov process is a stochastic process such that for any $t_1 < t_2 < \cdots < t_n$,

$$\mathbb{P}(y_n, t_n | y_1, t_1; \dots ; y_{n-1}, t_{n-1}) = \mathbb{P}(y_n, t_n | y_{n-1}, t_{n-1})$$

Hence a Markov process is completely determined by the functions $P(y_1, t_1)$ and the transition probabilities $P(y_2, t_2|y_1, t_1)$. The Master equation holds for any Markov process: Appendix A contains a complete derivation of the Master equation as an equivalent form of the Chapman-Kolmogorov equation, which is a direct consequence of the Markov property (adapted from Chapters IV and X of van Kampen's *Stochastic Processes in Physics and Chemistry* [29]). For a general Markov process Y, the Master equation reads

$$\frac{\partial P(y,t)}{\partial t} = \int \{W(y|y')P(y',t) - W(y'|y)P(y,t)\} \mathrm{d}y' \quad (7)$$

where $W(y' \mid y) = 0$ is the transition probability per unit time from y to y'. If the range of Y is a discrete set of states labeled by n, then Eq. (7) reduces to

$$\frac{\mathrm{d}p_n(t)}{\mathrm{d}t} = \sum_{n'} (W_{n,n'} p_{n'}(t) - W_{n',n} p_n(t)) \quad (8)$$

Birth-and-death processes—Birth-and-death (or one-step) processes are a special class of Markov processes whose range consists of integers n and whose transition matrix permits only jumps between adjacent sites:

$$W_{n,n'} = r_{n'} \delta_{n,n'-1} + g_{n'} \delta_{n,n'+1}$$

(Note that this does not mean that it is impossible for the system to make two jumps within one time step Δt, but only that the probability is $O(\Delta t^2)$.) Hence the Master equation reduces to

$$\dot{p}_n = r_{n+1} p_{n+1} + g_{n-1} p_{n-1} - (r_n + g_n) p_n \quad (9)$$

The birth and death rates, g_n, r_n, respectively, can be arbitrary functions of n, even nonlinear ones. If only non-negative integers are allowed, then for $n = 0$ we must replace $\dot{p}$ with

$$\dot{p}_0 = r_1 p_1 - g_0 p_0$$

or alternatively we may define $r_0 = g_{-1} = 0$.

One important example of a one-step process with constant transition rates is the Poisson process: $r_n = 0$, $g_n = q$, $p_n(0) = \delta_{n,0}$, i.e.,

$$\dot{p}_0 = q(p_{n-1} - p_n)$$

It is random walk over the integers taking steps to the right only, but at random times. The negative Poisson process (taking steps to the left) is a good model for protein degradation, as we will see shortly.

Multivariable birth-and-death processes—The generalization of the Master equation for birth-and-death processes to multiple variables is straightforward. Consider an n-dimensional birth-and-death process $\mathbf{X}(t) \in \mathbb{Z}^n$ with birth and death rates **g; r** : $\mathbb{Z}^n \to \mathbb{R}^n$, respectively. That is, $g_j(\mathbf{k})$, $r_j(\mathbf{k})$ denote the birth and death rates, respectively, of the jth species when $\mathbf{X} = \mathbf{k} \in \mathbb{Z}^n$. The Master equation governing this process is given by

$$\frac{\mathrm{d}P(\mathbf{k},t)}{\mathrm{d}t} = \sum_{j=1}^{n} [g_j(\mathbf{E}_j^- \mathbf{k})P(\mathbf{E}_j^- \mathbf{k},t) + r_j(\mathbf{E}_j^+ \mathbf{k})P(\mathbf{E}_j^+ \mathbf{k},t) - (g_j(\mathbf{k}) + r_j(\mathbf{k}))P(\mathbf{k},t)]$$

where $\mathbf{E}_j^{\pm}\mathbf{k} = [k_1, \ldots, k_{j-1}, k_j \pm 1, k_{j+1}, \ldots, k_n]^T$.

The Master equation model for gene regulation

Now we are ready to apply the basic theory of the previous subsection to the gene regulation problem. We will show how to model gene regulation as a birth-and-death process, where birth corresponds to transcription and death to degradation, and derive the appropriate Master equation model. In the one gene case, an explicit steady-state solution is available. In order to account for multiple genes and the distinction between RNA and protein, we require a multivariate formulation.

Gene regulation as a birth-and-death process—We now wish to develop a stochastic model for gene regulation. We will start simply, considering a system with a single gene, and temporarily ignoring the distinction between RNA and protein. Let $X(t)$ be a discrete random variable representing the number of RNA transcripts present in the cell at time t. $X(t)$ has a time-dependent probability distribution given by $P(k, t) \equiv \mathbb{P}\{X(t) = k\}$. Analogous to the deterministic model of section **Dynamical system model**, we can model $X(t)$ as a birth-and-death process with birth rate $\tau F(k)$ and death rate γk, where F models the RNAP-promoter binding probability as a function of the current number of transcripts (in the single-gene case, we can only account for self-regulation). If there are initially k RNA transcripts, then over an infinitesimal timestep Δt either a degradation event may occur with probability $\gamma k \Delta t$, an RNAP-promoter binding event may occur followed by RNA transcription with probability $\tau F(k)\Delta t$, or neither may occur. (It is highly unlikely ($O(\Delta t^2)$) that two or more of these events occur within a single timestep.) Hence, as Figure 1 shows, the probability $P(k, t)$ increases by $P(k - 1)$ times the probability transcription plus $P(k,t)$

times the probability of degradation, and decreases by $P(k)$ times the probability of transcription plus the probability of degradation. The Master equation governing the evolution of $P(k,t)$ over time is therefore:

$$\frac{\mathrm{d}P(k,t)}{\mathrm{d}t}=\tau F(k-1)P(k-1,t)+\gamma(k+1)P(k+1,t)-(\tau F(k)+\gamma k)P(k,t). \quad (10)$$

Explicit steady-state solution for one gene systems—A general single-species birth-and-death process governed by the Master equation

$$\dot{p}_k=r_{k+1}p_{k+1}+g_{k-1}p_{k-1}-(r_k+g_k)p_k$$

has an explicit steady-state probability distribution given by

$$p_k^s=\frac{g_0g_1\cdots g_{k-1}}{r_1r_2\cdots r_k}p_0 \quad (11)$$

(van Kampen VI.3.8 [29]). The proof is by induction. Applied to a single-gene system, this becomes

$$p^s(k)=p^s(0)\frac{(\tau/\gamma)^k}{k!}\prod_{j=1}^{k-1}F(j). \quad (12)$$

This formula is very useful for studying one gene systems with minimal computation. For example, it can be used to directly compute the steady-state mean and variance of a single-gene system.

Multiple genes—In order to study stochasticity in gene regulation, we must extend our framework to include multiple-gene systems as well. In order to do this we can apply the Master equation for multivariate birth-and-death processes. Consider a system with n genes, and let $X(t) \in \mathbb{Z}^n$ be a discrete random vector, where $X_j(t)$ represents the number of RNA transcripts of gene j present in the cell at time t. X(t) has a time-dependent probability distribution given by $P(\mathbf{k},t) \equiv \mathbb{P}(\mathbf{X}(t) = \mathbf{k}) = \mathbb{P}\{X_j(t) = k_j, 1 \le j \le n\}$, for $\mathbf{k} \in \mathbb{Z}^n$. Similar to the one gene case, we can model $X(t)$ as a birth-and-death process with Master equation:

$$\frac{\mathrm{d}P(\mathbf{k},t)}{\mathrm{d}t}=\sum_{j=1}^{n}[\tau_jF_j(\mathbf{E}_j^-\mathbf{k})P(\mathbf{E}_j^-\mathbf{k},t)+\gamma_j(k_j+1)P(\mathbf{E}_j^+\mathbf{k},t)-(\tau_jF_j(\mathbf{k})+\gamma_jk_j)P(\mathbf{k},t)],$$

where

$$\mathbf{E}_j^{\pm}\mathbf{k}=[k_1,\cdots,k_{j-1},k_j\pm 1,k_{j+1},\cdots,k_n]^T. \quad (13)$$

RNA and protein—Initially, we simplified the discussion by ignoring protein translation and focusing only on the number of RNA transcripts of each gene. The same multivariate Master equation that allowed us to handle multiple genes also allows us to model the stochasticity of protein translation. If we introduce another discrete random vector $\mathbf{Y}(t) \in \mathbb{Z}^n$, where $Y_j(t)$ denotes the number of protein translates of gene j, and define $P(\mathbf{k}^r; \mathbf{k}^p; t) \equiv$ $\mathrm{P}(\mathbf{X}(t) = \mathbf{k}^r; \mathbf{Y}(t) = \mathbf{k}^p)$ the Master equation corresponding to the deterministic model (1) is

$$\begin{aligned}&\frac{\mathrm{d}P(\mathbf{k}^r,\mathbf{k}^p,t)}{\mathrm{d}t}\\ &=\sum_{j=1}^{n}\{\tau_j F_j(\mathbf{k}^p)P(\mathbf{E}_j^-\mathbf{k}^r,\mathbf{k}^p,t)+\gamma_j^r(k_j^r+1)P(\mathbf{E}_j^+\mathbf{k}^r,\mathbf{k}^p,t)\\ &\quad+p_jk_j^rP(\mathbf{k}^r,\mathbf{E}_j^-\mathbf{k}^p,t)+\gamma_j^p\left(k_j^p+1\right)P(\mathbf{k}^r,\mathbf{E}_j^+\mathbf{k}^p,t)\\ &\quad-\left(\tau_jF_j(\mathbf{k}^p)+\gamma_j^rk_j^r+p_jk_j^r+\gamma_j^pk_j^p\right)P(\mathbf{k}^r,\mathbf{k}^p,t)\}.\end{aligned}$$

If we apply the quasi-steady-assumption discussed in section **RNAP binding probability models** to the translation step, that is, we assume that protein levels are approximately proportional to RNA levels at all times, then this equation reduces to the form (13) [49]. Although this assumption may not always be biologically accurate, model (13) is the only practical option in many cases, since experimental technologies for measuring both mRNA and protein levels concurrently are not yet available.

Any modeling effort is necessarily a compromise between accuracy and tractability, and this case is no exception. Since the biological mechanisms of gene transcription are extraordinarily complex and not completely understood, our model relies on a number of simplifying assumptions, both biological and physical in nature. One of the most explicit is the assumption that the rates of degradation, translation, and transcription (when RNAP is bound) are constant for each gene. In reality, the rates are affected by many other processes including chromatin remodeling, translational regulation, and protein folding. As discussed in section **RNAP binding probability models**, modeling the RNAP binding function also involves several simplifications and assumptions.

EXPANSION AND SIMULATION METHODS

The Master equation cannot be solved explicitly except in the simplest cases. For a one gene system, we have an explicit formula for the steady-state distribution (Equation (12)), but no such formula exist for multiple genes. Therefore, in order to make further progress we will need approximations of the Master equation and efficient simulation methods. Fortunately, much of the work has already been done by physicists studying the Master equation. Beginning in the 1970s, N. G. van Kampen [29] and Ryogo Kubo [51] developed a systematic expansion method for approximating the Master equation at any level of detail. Gillespie created a stochastic simulation algorithm to generate statistically correct trajectories of the Master equation; another simulation method based on the Langevin equation is less accurate but more efficient. We will summarize their findings in this section and show how they can be applied to the gene regulation problem. In the next section we will perform simulation studies on simple synthetic gene regulatory systems to illustrate the application of these methods and understand their strengths and weaknesses.

The Gillespie algorithm

The Gillespie algorithm enables numerical simulation of statistically correct trajectories of a system governed by the Master equation. The iterative Monte Carlo procedure randomly chooses the next event that will occur and the intervening time interval, then updates the molecular numbers of each species and the trajectory time [52]. If the simulated system is in state $\mathbf{X}(t) \in \mathbb{R}^n$ at time t, the waiting time τ before its next jump is drawn from an exponential distribution, and the probability of jumping to state $\mathbf{X}^{(\mu)}$ is $w_\mu \propto W(\mathbf{X}^{(\mu)}|\mathbf{X})$ (the Master equation transition probability for $\mathbf{X} \to \mathbf{X}^{(\mu)}$ per unit time). The basic steps of the algorithm are

1. Initialize the molecular numbers of each species, $X_1, \ldots, X_n$, and set $t = 0$.
2. Randomly choose the next event to occur, and an exponential waiting time τ, by generating uniform random numbers r_1, r_2 from Unif(0,1), and setting

$$w_0=\sum_{\mu} w_\mu, \qquad \tau=\frac{1}{w_0}\log\frac{1}{r_1}, \qquad \mu:\sum_{v=1}^{\mu-1} w_v<w_0 r_2<\sum_{v=1}^{\mu} w_v.$$

3. Update the time and molecular numbers based on the chosen event and time

$$t \to t+\tau, \qquad \mathbf{X}(t) \to \mathbf{X}_\mu.$$

4. Repeat steps 2–3 until the simulation time reaches limit $t > T_{\text{sim}}$.

The Gillespie algorithm provides an exact simulation of the Master equation at a high computational cost, which increases rapidly with the number of species and the system size. While it is very attractive for small systems, alternative approaches are needed for gene regulatory systems with many genes and large systems sizes. In the next few sections, we will discuss theoretical approximations as well as an efficient but inexact simulation method based on the Langevin equation.

The van Kampen expansion

N. G. van Kampen provides a systematic approximation method involving an expansion in the powers of small parameter inversely related to the system size [29]. The Master equation can be approximated at any level of detail by truncating the expansion to omit the higher-order terms. Ryogo Kubo, a contemporary of van Kampen, arrived at an equivalent formulation by a slightly different approach [51]. We will follow van Kampen's development here since it is more transparent. For simplicity, we only describe the one-dimensional expansion, but the multivariate case is similar; van Kampen Chapter X.5 shows how to extend the theory to multiple variables [29].

In order to establish the relative scales of macroscopic and microscopic (jump) events, van Kampen introduces a system-size parameter Ω, such that for large Ω the fluctuations are relatively small. His approximation takes the form of an expansion in the powers of $\Omega^{-\frac{1}{2}}$. A critical assumption is that the transition probability function *W* has the form

$$W_\Omega(X+r|X) \equiv \Omega\Phi\left(\frac{X}{\Omega};r\right),$$

which means that the transition probabilities depend only on the macroscopic variable $x=\frac{X}{\Omega} \in \mathbb{R}$ and on the size of the jumps $r \in \mathbb{Z}$. For our application, we assume that this is the case, and that the jumps can only have magnitude 1:

$$\begin{gathered} W(X+1|X)=F(X) \equiv \Omega f\left(\frac{X}{\Omega}\right) \iff \Phi_0(x;+1)=f(x) \\ W(X-1|X)=\gamma X \iff \Phi_0(x;-1)=\gamma x. \end{gathered}$$

The expansion begins with the Ansatz that the probability distribution *P*(*X*, *t*) has a peak of order Ω tracking the macroscopic solution, with width of order $\Omega^{-\frac{1}{2}}$ corresponding to the fluctuations:

$$X(t)=\Omega\phi(t)+\Omega^{\frac{1}{2}}\xi. \quad (14)$$

The motivation for the Ansatz is the observation that the relative fluctuation effects in chemical systems tend to scale as the inverse square root of the system size [53]. It is justified a posteriori by the fact that *P*(*x*, *t*), expressed in terms of ξ, turns out to be independent of Ω to first approximation. As part of the expansion procedure, φ is chosen to track the peak, and turns out to be exactly the deterministic solution.

To compute the expansion, van Kampen redefines *P*(*X*, *t*) as a function Π of the new parameters φ, ξ via

$$P(X,t)=P\left(\Omega\phi(t)+\Omega^{\frac{1}{2}}\xi,t\right) \equiv \Pi(\xi,t),$$

rewrites the Master equation in terms of Ω, and proceeds to expand it in negative powers of Ω. To simplify the calculations, he defines the *jump moments*

$$\alpha_v(x)=\int r^v\Phi(x;r)\mathrm{d}r. \quad (15)$$

The first jump moment corresponds to the deterministic equation:

$$\frac{\mathrm{d}y}{\mathrm{d}t}=\alpha_1(y)=\frac{1}{\Omega}\int rW_\Omega(Y+r|Y)\mathrm{d}r.$$

For a birth-and-death process, this simplifies to

$$\alpha_1(y)=\frac{1}{\Omega}W_\Omega(Y+1|Y)-\frac{1}{\Omega}W_\Omega(Y-1|Y);$$

in our case

$$\alpha_1(y)=\tau f(y)-\gamma y,\ \alpha_2(y)=\tau f(y)+\gamma y. \quad (16)$$

For multiple genes, the first-order jump moments are just $\alpha_{1,i}(y) = \tau_i f_i(y) - \gamma y_i$ for $1 \leq i \leq n$, but the second-order jump moments, $\alpha_{2,i,j}$, $1 \leq i, j \leq n$, are more complex since they involve interactions: see van Kampen Chapter X.5 for further details.

The complete calculation (adapted from Chapter X of van Kampen) is provided in Appendix B. A crucial step in the expansion is the cancellation of terms of order $\Omega^{-\frac{1}{2}}$, which cannot belong to a proper expansion for large Ω. The cancellation is made possible by choosing $\varphi(t)$ (the macroscopic part of X) such that

$$\frac{\mathrm{d}\phi}{\mathrm{d}t}=\alpha_1(\phi)$$

That is, φ exactly satisfies the deterministic equation. The final result (to order Ω^{-1}) is that

$$\frac{\partial\Pi}{\partial t}=-\alpha'_1(\phi)\frac{\partial\xi\Pi}{\partial\xi}+\frac{1}{2}\alpha_2(\phi)\frac{\partial^2\Pi}{\partial\xi^2}+\frac{1}{2}\Omega^{-\frac{1}{2}}\left(\alpha'_2(\phi)\frac{\partial^2\xi\Pi}{\partial\xi^2}-\alpha''_1(\phi)\frac{\partial\xi^2\Pi}{\partial\xi}-\frac{1}{3!}\alpha_3(\phi)\frac{\partial^3\Pi}{\partial\xi^3}\right)+O(\Omega^{-1}) \quad (17)$$

with jump moments α_v defined by (15).

As we will discuss in greater detail later, the validity of the expansion relies on the assumption that the macroscopic equation $\frac{\mathrm{d}\phi}{\mathrm{d}t}=\alpha_1(\phi)$ has a single stable stationary state (satisfying $\alpha_1(\varphi) = 0$, $\alpha'_1(\varphi) \leq -\varepsilon<0$), which attracts all trajectories. If this is not the case, it is possible for a random fluctuation to send a stochastic trajectory out of the domain of attraction of the deterministic steady-state near which we would expect it to remain. For now, we will assume that the condition holds. Then the expansion is valid and can be truncated at the desired level of detail and translated back into the original variable via $X(t)=\Omega\phi(t)+\Omega^{\frac{1}{2}}\xi(t)$ to yield various approximation schemes.

The linear noise approximation—Restricting attention to the terms of order $\Omega^0 = 1$ in this expansion yields the linear noise approximation

$$\frac{\partial\Pi}{\partial t}=-\alpha'_1(\phi)\frac{\partial\xi\Pi}{\partial\xi}+\frac{1}{2}\alpha_2(\phi)\frac{\partial^2\Pi}{\partial\xi^2}+O\left(\Omega^{-\frac{1}{2}}\right). \quad (18)$$

This is a linear Fokker-Planck equation, and the solution turns out to be a Gaussian (see van Kampen VIII.6 [29]). Hence it is completely characterized by the first and second moments of ξ, which are of the most interest to us anyway. Multiplying Equation (18) by ξ and ξ^2, respectively, yields differential equations in the mean and variance of ξ (denoted $\langle\xi\rangle$, $\langle\langle\xi\rangle\rangle$, respectively):

$$\frac{\partial}{\partial t}\langle\xi\rangle=\alpha'_1(\phi)\langle\xi\rangle \quad (19)$$

$$\frac{\partial}{\partial t}\langle\langle\xi\rangle\rangle=2\alpha'_1(\phi)\langle\langle\xi\rangle\rangle+\alpha_2(\phi). \quad (20)$$

After solving for the mean and variance of ξ and solving the deterministic equation for φ, we can use the Ansatz (14) to find the mean and variance of X:

$$\langle X(t)\rangle=\Omega\phi(t)+\Omega^{\frac{1}{2}}\langle\xi(t)\rangle, \langle\langle X(t)\rangle\rangle=\Omega\langle\langle\xi\rangle\rangle.$$

The initial condition $P(X, 0) = \delta(X - X_0)$ implies $\varphi_0 = x_0$ and $\langle\xi\rangle_0 = \langle\langle\xi\rangle\rangle_0 = 0$, hence $\langle\xi\rangle t \equiv 0$. (Even if ξ has a nonzero initial distribution, if $\alpha'_1(\varphi) < -\varepsilon<0$ we still will have $\langle\xi\rangle \le e^{-\varepsilon t} \to 0$. Hence the mean of the solution to the Master equation with initial distribution δ_{x0} approximately satisfies the deterministic equation:

$$\frac{\partial}{\partial t}\langle x\rangle=\alpha_1(\langle x\rangle)+O(\Omega^{-1}). \quad (21)$$

This result extends to the multivariate case. If $\xi \in \mathbb{R}^n$, the mean becomes a vector in $\mathbb{R}^n$ and the variance becomes an n × n covariance matrix. In Equations (19) and (20), the functions α'_1, α_2 are replaced with a Jacobian matrix and a matrix of second-order jump moments, but the basic structure of the expansion is similar, and (21) still holds. The multivariate case is discussed further in van Kampen Chapter X.5, and by Komorowski et al. (with different notation) [54].

Connection to nonlinear deterministic model—Equation (21) provides the link between the stochastic Master equation and the nonlinear deterministic dynamical system model (1). It shows that the deterministic equation is an approximate model for the evolution of the mean expression of the stochastic process. That is,

$$\frac{\partial}{\partial t}\langle x\rangle \approx \alpha_1(\langle x\rangle)=\tau f(\langle x\rangle)-\gamma\langle x\rangle$$

with error on the order of a single molecule. Therefore, under a few reasonable assumptions about system size and steady-state stability, the population mean still approximately satisfies the nonlinear deterministic equation

$$\frac{\mathrm{d}y}{\mathrm{d}t}=\tau f(y)-\gamma y.$$

One of the attractions of this model is an associated algorithm for learning an unknown underlying gene regulatory network from experimental data [38,55]. The algorithm selects terms to include in the model and estimates their coefficients without requiring any prior knowledge of the regulators, based on gene expression data at perturbed steady-states. Thus, relatively easy-to-acquire steady-state gene expression data can be used to learn a model that quantitatively describes the complete dynamical behavior of the system.

The Fokker-Planck and Langevin equations

In section **The linear noise approximation**, we saw that the linear noise approximation gave rise to a linear Fokker-Planck equation. Fokker-Planck or (mathematically equivalent) Langevin equations predate the van Kampen expansion and are still often used as approximations of the Master equation or directly as models of Markov processes with small jumps (although this sometimes leads to difficulties that must be resolved by van Kampen's approach). In this section we will discuss these two types of equations and their applications on the gene regulation. Although the approximation is not entirely consistent due to the nonlinearity of the problem, the Langevin equation is the basis of an efficient simulation approach that enables large-scale simulations of multiple-gene systems.

The Fokker-Planck equation is a differential equation consisting of a "transport" and a "diffusion" term:

$$\frac{\partial P(y,t)}{\partial t}=-\frac{\partial}{\partial y}\alpha_1(y)P+\frac{1}{2}\frac{\partial^2}{\partial y^2}\alpha_2(y)P. \quad (22)$$

(For the multivariate version, see van Kampen VIII.6.1.) In the general form of the equation, α_1, α_2 are any real differentiable functions with $\alpha_2 > 0$, but in Planck's derivation of the equation as an approximation to the Master equation [56], they are exactly the first and second jump moments (15). Since the Fokker-Planck equation is always linear in *P*, we follow van Kampen in appropriating the term linear to mean that α_1 is linear and α_2 constant.

The Langevin equation is a stochastic differential equation (SDE) of the form

$$\mathrm{d}y=\alpha_1(y)\mathrm{d}t+\sqrt{\alpha_2(y)}\mathrm{d}W, \quad (23)$$

where *W*(*t*) is a Wiener process, or Brownian motion. Again, α_1, $\alpha_2 > 0$ may be any C^1 functions in general, but in the case of interest to us, they represent the jump moments (15). Equations (22) and (23) are mathematically equivalent using the Ito interpretation of (23) (see van Kampen IX.4 [29] for the proof).

These equations are very appealing for modeling physical processes since they are easy to derive and interpret. For both equations, a_1, a_2 (thought of for now as general functions, not as the jump moments) can be inferred without even knowing the underlying Master equation, using only the macroscopic law and fluctuations around the steady-state solution (known from statistical mechanics). The approach works very well in situations where the macroscopic law a_1 is linear [56–59]. However, confusion can arise when a_1 is nonlinear, since effects on the order of the fluctuations are invisible macroscopically [60]. One of the major motivations for van Kampen's systematic expansion was the need to resolve disagreements between authors who had developed different but equally plausible characterizations of the noise in nonlinear systems using this approach.

For systems with linear deterministic equations, the van Kampen approximation agrees exactly with the Fokker-Planck model, since the linear noise approximation yields a linear Fokker-Planck equation. However, discrepancies may arise for nonlinear systems, and we should consider the van Kampen theory definitive in such cases. The error in the nonlinear Fokker-Planck model is that it retains the full functional dependence on the nonlinear functions a_1, a_2 (in effect, keeping infinitely many terms of their Taylor expansions) while cutting off their third-order and higher derivatives in the expansion about the deterministic path $\varphi(t)$. In contrast, the truncated van Kampen expansion replaces a_1, a_2 by their Taylor polynomials at a level of detail consistent with the order of the approximation. The van Kampen expansion provides a completely consistent approximation of the Master equation to any desired order of accuracy, while the Fokker-Planck model is a slightly inconsistent second-order approximation only. Nevertheless, the discrepancy between the Fokker-Planck and van Kampen approximations is often not too serious (and a second-order approximation is typically good enough), so the models are still very useful in many cases.

The Langevin equation, in particular, lends itself to efficient simulation [37,53]. Simulation provides insight into the behavior of individual trajectories as well as moment information, and applies directly to multistable systems (while van Kampen requires alternative theory since the expansion method only applies to systems with one stable steady-state). However, simulation can be very expensive. The exact Gillespie algorithm and other direct simulation methods are only computationally feasible for very small systems. Fortunately, the Langevin simulation works well for large systems with many genes, since trajectories of the Langevin equation can be simulated by evolving a small system of stochastic differential equations, rather than accounting for every single reaction like the Gillespie algorithm. Hence the Langevin simulation is appropriate for large systems with complex qualitative structures.

With these risks and potential rewards in mind, we will show how to apply the Langevin approach to the gene regulation problem. In the next section we will compare the results of Langevin simulations with the more accurate predictions of van Kampen or the direct Master equation simulation where possible. Using the first- and second-jump moments (16) for our problem, the one-dimensional Langevin equation is

$$\mathrm{d}y=(f(y)-\gamma y)\mathrm{d}t+\sqrt{f(y)}\mathrm{d}W_1+\sqrt{\gamma y}\mathrm{d}W_2 \quad (24)$$

where W_1, W_2 are independent Wiener processes.

Systems with multiple stable steady-states

We have alluded several times to the fact that stochasticity can lead to unexpected results for systems with multiple stable steady-states. The basic reason is that random fluctuations can send stochastic trajectories out of the domain of attraction of one deterministic steady-state and into the domain of another. Van Kampen treats these issues in detail in Chapter XIII of his book [29]. In this section, we will summarize the points that are most relevant to our topic. In the next chapter, simulation studies will illustrate these points and provide additional insight.

For simplicity, consider a birth-and-death process with two distinct stable steady-states, $\varphi_a<\varphi_c$ and an unstable steady-state $\varphi_b(\varphi_a<\varphi_b<\varphi_c)$. By this we mean that the corresponding deterministic equation $\mathrm{d}\varphi/\mathrm{d}t=\alpha_1(\varphi)$ has the following properties:

$$\alpha_1(\phi_a)=\alpha_1(\phi_b)=\alpha_1(\phi_c)=0 \quad (25)$$

$$\alpha_1^{\prime}(\phi_a)<0,\ \alpha_1^{\prime}(\phi_b)>0,\ \alpha_1^{\prime}(\phi_c)<0. \quad (26)$$

A deterministic trajectory will eventually converge to the nearest stable steady-state, that is, trajectories with initial conditions $\leq\varphi_b$ will converge to φ_a, and those with initial conditions $\geq\varphi_b$ will converge to φ_c. (A trajectory with initial condition φ_b will remain there, but this is not physically meaningful even in the deterministic case since the slightest perturbation will send the trajectory toward φ_a or φ_b)

When we take stochasticity into account, it is also possible for a large fluctuation to send a trajectory out of the domain of attraction of φ_a and into that of φ_b. These large fluctuations are usually unlikely, so it may take a very long time before one occurs. For systems of macroscopic size, this escape time can be so long that the event may never be observed. In smaller systems, however, transitions between steady-state domains can be a fairly common occurrence.

For systems in which giant fluctuations are relatively rare, we can distinguish two time scales: a short time scale on which equilibrium is established within the domain of attraction of a particular steady-state, and a long time scale on which giant fluctuations occur (sending trajectories out of the domain of attraction of one steady-state and into another). The rate of occurrence of the giant fluctuations is roughly equal to the height of the steady-state distribution at the unstable point φ_b, which means that the escape time scales exponentially with the system size, Ω.

A system that starts out near the unstable point φ_b evolves in three basic stages. At first, each trajectory has a reasonable probability of moving toward either of the stable points φ_a or φ_c, so the distribution widens quickly, but fluctuations across φ_b are quite possible. In the next stage, the probability has split into two autonomous parts, and fluctuations across φ_b cease, since each trajectory has settled into the domain of attraction of either φ_a or φ_c. In the final stage, the probability has reached a final bimodal stochastic steady-state distribution peaked at φ_a and φ_c. There is still a chance that fluctuations will send trajectories from one regime to another, but the probabilities are balanced so as to maintain the distribution.

A system that starts out near the stable point φ_a evolves differently, but eventually reaches the same bimodal stochastic steady-state distribution peaked at φ_a and φ_c (i.e., stage three), although it takes much longer to do so. Giant fluctuations can release trajectories from the domain of attraction of φ_a, but these occur on the long time-scale, so the probability peak at φ_c builds up much more slowly. Of course, if giant fluctuations are not particularly rare (in small systems, for example), then the initial condition has little impact on the time required to reach the steady-state distribution.

The relationship between the escape times and the probability of the regimes in the stochastic steady-state distribution is simple. Define the probabilities π_a, π_c of a trajectory $\varphi(t)$ being in the domain of φ_a, φ_c, respectively, by

$$\pi_a=\sum_{-\infty}^{\phi_b} p_n(t),\ \pi_c=\sum_{\phi_b}^{\infty} p_n(t).$$

Let τ_{ac}, τ_{ca} represent the escape times, that is, $\frac{1}{\tau_{ac}}$ is the probability per unit time for a trajectory in the domain of φ_c to cross the boundary φ_b into the domain of φ_a. Then we have

$\dot{\pi}_a=-\dot{\pi}_c=-\frac{\pi_a}{\tau_{ca}}+\frac{\pi_c}{\tau_{ac}}$ [van Kanmpen XIII. 1.4]

At steady-state ($\dot{\pi}_a=\dot{\pi}_c=0$),

$$\frac{\pi_a^s}{\tau_{ca}}=\frac{\pi_c^s}{\tau_{ac}}.$$

We can identify the escape time τ_{ca} with the mean first-passage time from φ_a to φ_c. For the one dimensional process defined by Equation (9), the mean first-passage time from φ_a to φ_c is given by

$$\tau_{ca}=\sum_{k=a}^{c-1} \frac{1}{g_k p_k^s} \sum_{j=0}^{k} p_j^s, \quad (27)$$

where p^s is the stationary distribution (11), as shown in Appendix C. The escape rate is $O(p_b^s)$, the height of stationary distribution at the unstable point b, so the escape time scales exponentially with the system size [61].

The relative stability of the two stable steady-states, $\frac{\pi_a^s}{\pi_c^s}$, depends on the relative depths and widths of the two corresponding potential energy wells. To illustrate this, consider the Fokker-Planck equation modeling diffusion in a potential *U*:

$$\frac{\partial P(x,t)}{\partial t}=\frac{\partial}{\partial x}U^{'}(x)P+\theta\frac{\partial^2 P}{\partial x^2}. \quad (28)$$

Although this model is not even approximately appropriate for the gene regulation problem since the diffusion coeffcient is constant, while in the gene regulation problem it is a function of *x*, it helps clarify some important issues. To that end, suppose the derivative of *U* satisfies the bistability conditions (26), so that d*U* = d*x* and *U* have the shapes shown in Figure 2. d*U*/d*x* has zeros at the steady-states φ_a, φ_b, φ_c, and *U* has minima at the stable points φ_a, φ_c and a maximum at the unstable point φ_b.

The corresponding deterministic equation is $\dot{x}=-U'(x)$. The stationary distribution is given by

$$P^s(x)=Ce^{-U(x)/\theta},\ C^{-1}=\int e^{-U(x)/\theta}\mathrm{d}x,$$

and for small θ we can approximate

$$\begin{gathered}C^{-1}\approx e^{-U(a)/\theta}\sqrt{\frac{2\pi\theta}{U''(a)}}+e^{-U(c)/\theta}\sqrt{\frac{2\pi\theta}{U''(c)}}\\ \pi_a^s\approx\int_{-\infty}^{b}P^S(x)\mathrm{d}x=C\sqrt{\frac{2\pi\theta}{U''(a)}},\\ \pi_c^s\approx\int_{b}^{\infty}P^S(x)\mathrm{d}x=C\sqrt{\frac{2\pi\theta}{U''(c)}}\\ \frac{\pi_a^s}{\pi_c^s}\approx e^{-(U(a)-U(c))/\theta}\sqrt{\frac{U''(c)}{U''(a)}}\\ [\mathrm{van\ Kampen\ XIII.1.10-1.11}].\end{gathered}$$

Hence the relative stability of the two stable steady-states depends on both the depths of the potential energy wells (*U*(*a*) and *U*(*c*)) and their widths (*U*″(*a*) and *U*″(*c*)). In Figure 2, φ_c is more stable than φ_a, since its potential energy is lower and energy well is wider. The relative stability in this example is about $\frac{\pi_a}{\pi_c}=0.76$, meaning that at stochastic steady-state, about 43% of trajectories will be near φ_a and 57% will be near φ_c at a given time (as shown in Figure 2, right pane). Similarly, we can approximate the escape time (mean first-passage time) by

$$\begin{gathered}\tau_{ca}\approx\frac{2\pi}{\sqrt{U''(a)|U''(b)|}}e^{(U(b)-U(a))/\theta}.\\ [\mathrm{van\ Kampen\ XIII.2.2}]\end{gathered}$$

Hence the escape time depends on the height of the energy barrier *U*(*b*) and energy well *U*(*a*), and the widths *U*″ (*a*), *U*″ (*b*) of the well and barrier. Since the potential energy difference is *O*(Ω), we see again that the escape time scales exponentially with the system size. Diffusion in multiple dimensions is qualitatively similar; van Kampen discusses the two-dimension case in XIII.4 [29].

In order to extend some of these ideas to the gene regulation problem, at least approximately, we need a non-constant diffusion term in the Fokker-Planck equation. The Fokker-Planck approximation corresponding to the fully nonlinear Master equation used for gene regulation is given by

$$\frac{\partial P(x,t)}{\partial t}=-\frac{\partial}{\partial x}\alpha_1(x)P+\frac{1}{2}\frac{\partial^2}{\partial x^2}\alpha_2(x)P.$$

(As we noted earlier, although this does not technically constitute a consistent approximation, it works well in most cases.) The steady-state solution is given by

$$\begin{gathered}P^s(y)=\tfrac{C}{\alpha^2(y)}\exp\left(2\textstyle\int_0^y\tfrac{\alpha_1(t)}{\alpha_2(t)}\mathrm{d}t\right).\\ [\text{van Kanmpen VIII.1.4}]\end{gathered} \quad (29)$$

(In the multidimensional case, finding the steady-state solution is less straightforward, as discussed in van Kampen XI.4, but simulating the equivalent Langevin simulation would provide an approximate result.) We can define an "effective potential" by

$$U_{\text{effective}}=-2\textstyle\int_0^y\dfrac{\alpha_1(t)}{\alpha_2(t)}\mathrm{d}t-2\log(\alpha_2(y)) \quad (30)$$

and numerically evaluate π_a, π_c, and the relative stability, using

$$\pi_a^s\approx\textstyle\int_{-\infty}^b P^s(x)\mathrm{d}x, \pi_c^s\approx\int_b^\infty P^s(x)\mathrm{d}x.$$

In the one-dimensional case, we could use the exact steady-state solution of the Master equation (12) instead, although the Fokker-Planck stationary solution may be more convenient. In the multivariate case, we must use the Fokker-Planck/Langevin approach, since there is no general approach to finding stationary solutions of multivariate Master equations.

Summary

Nonlinear Master equation models capture the stochastic mechanisms of gene regulation in full molecular detail. The Master equation can rarely be solved explicitly for multiple gene systems, but theoretical approximations and simulation algorithms can give insight into these systems. The Gillespie algorithm allows us to numerically simulate exact trajectories of the Master equation, although the computational cost becomes prohibitive for large systems with many genes. The van Kampen expansion method allows us to rigorously approximate the Master equation at any level of detail we desire (the deterministic model (1) being the simplest), provided that the system has only one stable steady-state, and van Kampen provides alternative theory for analyzing systems with multiple stable steady-states. The Langevin equation (equivalent to the Fokker-Planck equation) is an inexact approximation to the Master equation and is the basis of a highly effcient simulation method that is well suited for large multiple-gene systems. In the next section, we will perform simulation studies on simple synthetic gene regulatory systems to illustrate the application of each of these methods and evaluate their performance. As one might expect, the behavior of systems with multiple stable steady-states is particularly interesting.

STOCHASTIC SIMULATION STUDIES

In this section, we study several small synthetic gene regulatory systems in order to gain insight into the effects of stochasticity on systems with different qualitative characteristics, and the suitability and accuracy of different approximation and simulation methods in various situations. The simulation studies will compare the true Master equation (when feasible), second-order van Kampen approximation, deterministic equation (linear-noise approximation), Gillespie simulation, and Langevin simulation, in order to understand the strengths and limitations of each.

One gene system with one stable steady-state

Consider a single self-repressing gene whose self-regulation is governed by the deterministic differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}t}=f(y)-\gamma y, f(y)=\frac{2\gamma}{1+y}, \gamma=0.1. \quad (31)$$

It has a single (non-negative) deterministic steady-state at $y = 1$, satisfying $f(y) - \gamma y = 0$. (The other solution, $y = -2$, is negative and therefore not physically meaningful, nor is it realizable by the system assuming a non-negative initial condition.) The corresponding Master equation is

$$\frac{\mathrm{d}P(k)}{\mathrm{d}t}=F(k-1)P(k-1)+\gamma(k+1)P(k+1)-(F(k)+\gamma k)P(k), \quad (32)$$

where $F(k) = \Omega f(k/\Omega)$. Numerical evolution of the Master equation by iteratively updating a vector of probabilities according to (32) is feasible in this case because the system is so simple. The Master equation also has the explicit steady-state solution:

$$P^s(k)=\frac{P^s(0)}{\gamma^k k!}\prod_{j=0}^{k-1}\frac{2\gamma\Omega}{\left(1+\frac{j}{\Omega}\right)}.$$

Figure 3 shows the stationary probability distributions for a range of values of Ω, revealing that as Ω increases, the distribution is increasingly sharply peaked at $y^s = 1$. That is, the mean approaches $y^s = 1$, and the variance goes to zero as Ω increases.

To second order, the van Kampen expansion gives

$$\begin{gathered}\frac{\mathrm{d}\phi}{\mathrm{d}t}=\alpha_1(\phi)=f(\phi)-\gamma\phi\\ \frac{\mathrm{d}\langle\xi\rangle}{\mathrm{d}t}=\alpha'{}_1(\phi)\langle\xi\rangle\tfrac{1}{2}\Omega^{-\frac{1}{2}}\alpha''{}_1(\phi)\langle\xi^2\rangle=(f'(\phi)-\gamma)\langle\xi\rangle+\tfrac{1}{2}\Omega^{-\frac{1}{2}}f''(\phi)\langle\xi^2\rangle\\ \frac{\mathrm{d}\langle\xi^2\rangle}{\mathrm{d}t}=2\alpha'{}_1(\phi)\langle\xi^2\rangle+\alpha_2(\phi)=2(f'(\phi)-\gamma)\langle\xi^2\rangle+(f(\phi)+\gamma\phi).\end{gathered}$$

We can solve for the steady-state values of φ, $\langle\xi\rangle$, and $\langle\xi^2\rangle$ by setting the left-hand-sides of all three equations to zero. The first equation is the deterministic evolution equation: we already know that its only non-negative solution is $\varphi^s=1$. Evaluating f and its derivatives at φ^s:

$$\begin{gathered}f(\phi^s)=0.2(1+\phi^s)^{-1}=0.1;f'(\phi^s)=-0.05;\\ f''(\phi^s)=0.05,\end{gathered}$$

and plugging into the last two equations yields

$$\phi^s=1,\ \langle\xi^2\rangle=\frac{2}{3},\ \langle\xi\rangle=\frac{1}{9}\Omega^{-\frac{1}{2}}.$$

Finally we obtain expressions for the steady-state mean and variance in terms of Ω:

$$\begin{gathered}\langle x^s\rangle=\phi^s+\Omega^{-\frac{1}{2}}\langle\xi^s\rangle=1+\tfrac{1}{9\Omega}\\ \langle\langle x^s\rangle\rangle=\Omega^{-1}\langle\langle\xi^s\rangle\rangle=\tfrac{2}{3\Omega}.\end{gathered}$$

The Langevin model for this system is given by the SDE

$$\mathrm{d}X=(F(X)-\gamma X)\mathrm{d}t+\sqrt{F(X)}\mathrm{d}W_1+\sqrt{\gamma X}\mathrm{d}W_2,$$

where $W_1(t)$, $W_2(t)$ are independent Wiener processes.

Figure 4 compares the exact Master equation, second-order van Kampen approximation, Gillespie simulation, and Langevin simulation for this system with initial condition $y^s = 1$ (the steady-state value) and three different values of Ω. As Ω increases, the agreement improves as the mean approaches the deterministic trajectory (that is, the steady-state value $y^s = 1$), and the variance decreases. The discrepancy between the stochastic mean and the deterministic trajectory and the variance are both $O(\Omega^{-1})$ (as predicted by the van Kampen expansion).

The Master equation governs the evolution of the probability distribution; Figure 5 shows the final probability distributions for each value of Ω. In each case, the initial probability is a delta-distribution centered at Ωy^s, and the probability spreads out over time to reach a steady-state distribution, which is extremely close to a Gaussian for $\Omega \gg 1$. For larger values of Ω, the final probability distribution remains sharply peaked around y^s.

Two gene system with one stable steady-state

Next we consider a two gene system, again with a single stable steady-state, governed by the deterministic differential equation

$$\frac{\mathrm{d}y_1}{\mathrm{d}t}=f_1(y)-\gamma y_1, f_1(y)=\frac{0.1+0.1y_2}{1+y_2} \quad (33)$$

$$\frac{\mathrm{d}y_2}{\mathrm{d}t}=f_2(y)-\gamma y_2, f_2(y)=\frac{0.4+0.1y_1y_2}{1+y_1y_2},\ \gamma=0.1. \quad (34)$$

It has a single deterministic steady-state at $y_1 = 1$, $y_2 = 2$. With two genes, directly evolving the Master equation is very expensive for moderately sized systems, as each probability distribution is now two-dimensional (in general, the computational cost of evolving the Master equation with system size Ω is $O(\Omega^n)$ per timestep), so we omit this method and focus on the van Kampen approximation and the Gillespie and Langevin simulations. In the Langevin simulation, we neglected the interaction terms in the second-order jump moments for simplicity. Figure 6 shows that the situation is qualitatively similar to the one gene case we just discussed. The approximation and simulation means differ from the deterministic trajectory by $O(\Omega^{-1})$, and the variance is also $O(\Omega^{-1})$. For $\Omega = 1$, the y_2 variance and mean discrepancy of the Langevin simulation and the van Kampen approximation are slightly lower than those of the exact Gillespie trajectories. This inaccuracy arises from zero-boundary effects and the non-Gaussianity of the probability distribution at small system sizes (on the order of a single molecule).

Constructing multistable systems

Gene regulatory systems with multiple stable steady-states are ubiquitous in nature as this property plays a key role in cellular lifecycles and responses to external stimuli. However, constructing synthetic systems with multiple stable steady-states with our chosen functional form (5) can be challenging. One approach, which Chickarmane et al. used to develop their ESC-inspired system [62], is to start with a well-understood biological network with multiple steady-states and use experimental data and knowledge of qualitative behavior to suggest the appropriate terms and parameter values. This can be an interesting and useful program, especially as the synthetic network may later be useful for gaining further insight into the behavior of the original biological network; however, there are very few biological networks well-understood enough to lend themselves to this type of modeling. Furthermore, it is limiting in the sense that it relies on existing known networks, and provides little insight into methods for generating original networks. Ideally, we would like to be able to create novel networks from scratch with specific properties of our own choosing. In this section we will discuss our efforts toward this end. Although we have not fully solved this problem by any means and would encourage further work in this direction, we have developed a heuristic algorithm that, together with some trial-and-error, allowed us to generate the two multistable synthetic gene networks we study shortly.

Suppose we wish to construct an *n*-gene system with *k* stable steady-states $e_1, \ldots, e_k$, of our choosing. That is, we want to find parameters b_{ij}, c_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$, where *m* is the number of terms in the model, so that

$$\begin{gathered} f_i(y)=\frac{b_{i_0}+\sum_{j=1}^{m} b_{ij}\Pi_{k\in S_{ij}} y_k}{1+\sum_{j=1}^{m} c_{ij}\Pi_{k\in S_{ij}} y_k} \\ \Rightarrow f_i(e_j)-\gamma e_{j,i}=0,\ 1\le i\le n,\ 1\le j\le m. \end{gathered}$$

Furthermore, $e_1, \ldots, e_k$ should be stable, so we require

$$\exists P_j \succ 0 \,\text{such that}\, J_f(e_j)^T P_j+P_j J_f(e_j) \prec 0,\ 1\le j\le k,$$

where $J_f(y)$ denotes the $n \times n$ Jacobian matrix of *f* at $y \in \mathbb{R}^n$.

Hence, we wish to find $b_i, c_i \in \mathbb{R}^m$ such that $f_i(e_j) = \gamma e_{j,i}$ while satisfying the Jacobian condition and the other constraints. That is, we want to solve the feasibility problem:

$$\begin{gathered} \text{find}\, b_i \in \mathbb{R}^m,\ c_i \in \mathbb{R}^m,\ P_j \in \mathbb{R}^{n\times n},\ 1\le i\le n,\ 1\le j\le k \\ \text{subject to}\, f_i(e_j)=\gamma e_{j,i} \\ 0\le b_i\le c_i,\ c_i(0)=1, \\ J_f(e_j)^T P_j+P_j J_f(e_j) \prec -\varepsilon, \\ \forall 1\le i\le n,\ 1\le j\le k. \end{gathered} \qquad (35)$$

If the problem is feasible, then b_i, c_i parametrize a system with the desired properties. Not all choices of the e_j necessarily lead to a feasible problem, so we may have to try several possibilities before we find a system with multiple stable steady-states.

The problem is nonconvex due to the rational form of *f* and the stability condition, so we can either use a nonconvex solver, or use heuristics and trial-and-error and solve with a convex solver. Specifically, we can use an iterative approach to enforce the stability constraint [63], and simply replace the denominator of each *f* with a constant value and add a constraint forcing the denominator to be equal to that constant. Of course, not all constant values lead to feasible problems, so if we use the heuristic approach, we must guess-and-check the denominator values as well as the steady-state locations.

One gene system with two stable steady-states

In this section, we study a one gene system with two stable steady-states (and one unstable steady-state) inspired by a synthetic system developed by Chao Du and refined using the algorithm of the previous section.

$$\frac{\mathrm{d}y}{\mathrm{d}t}=f(y)-\gamma y,\ f(x)=\frac{0.1+x+0.1x^4}{1+10x+0.5x^2+0.1x^4},\ \gamma=0.1 \quad (36)$$

gives rise to two stable steady-states: $e_1 \approx 1.0431$ and $e_2 \approx 7.9845$, and an unstable steady-state $e_3 \approx 4.0416$.

At the end of the last section, we discussed methods from Chapter XIII of van Kampen’s book for analyzing the equilibrium behavior of systems with multiple stable steady-states. These tools provide a great deal of insight into long-term system behavior with minimal computation, since they only require the stationary probability distribution (which can be computed directly in the single-gene case using (12), or approximated using the Fokker-Planck equation in general). These tools will allow us to predict some of the basic behavior of system (36) with very little effort. Simulations will confirm and complete the picture.

Let us first examine the most basic properties of the system with Ω = 1. As Figure 7 shows, the deterministic system is bistable. The deterministic function $\alpha_1(x) = f(x) - \gamma x$ has three zeros corresponding to the three deterministic steady-states. The derivative of the deterministic function is negative ($\mathrm{d}\alpha_1 = \mathrm{d}t<0$) at the stable steady-states, and positive at the unstable steady-state. The stationary distribution has a strong peak at the more stable steady-state, e_1, and a weaker one at the less-stable point e_2. The system is three times as likely to be in the domain of e_1 as in the domain of e_2. We can use the relative stability of the two stable points to estimate the steady-state mean: $\pi_1 e_1 + \pi_2 e_2 \approx 2.78$, which will be confirmed by our simulation study.

Next, let us examine the system with Ω = 10. Figure 8 shows the deterministic function, the effective potential, and the (approximate) stationary distribution, computed using the Fokker-Planck approach. The deterministic function and stationary distribution have the same qualitative properties as they did for Ω = 1, except that the e_1 peak of the stationary distribution is now even higher relative to e_2 ($\pi_1 \approx 97\%$; $\pi_2 \approx 3\%$) and the steady-state mean, 1.24, is therefore closer to e_1. The effective potential has minima at the stable steady-states, but the “energy” of the more stable state, e_1, is much lower.

Simulations reveal how the mean, variance, and probability distribution of the system actually evolve. Figures 9, 12 and 13 compare the exact Master equation, second-order van Kampen approximation, and Master equation and Langevin simulations for Ω = 1,10, and all but the exact Master equation for Ω = 100 (due to instability), respectively. Unlike for the one gene system described by Equation (31), the exact Master equation and both simulations deviate dramatically from both the van Kampen approximation and the deterministic trajectory, at least for Ω = 1,10. The reason for this is the bistability of the system. Especially when Ω is fairly small (hence the variance is relatively large) each stochastic trajectory starting from steady-state e_1 has a reasonably large probability of escaping from the domain of attraction of e_1 and being attracted to e_2, and vice versa. In the long run, the system settles to a bimodal steady-state distribution, in which both stable steady-states are represented proportional to their relative stability. Therefore, the steady-state mean regardless of the starting point converges to the roughly weighted average of the two deterministic stable steady-states predicted by the basic stability analysis described in Figure

7. The second-order van Kampen expansion centered at either of the two steady-states does not account for this blending effect and therefore underestimates both the variance and the deviation of the mean trajectory from the deterministic trajectory. In reality, the second-order expansion should never have been applied in this case since it is only valid for systems with a single stable steady-state, as van Kampen explains in Chapter X of his book [29].

Figure 10 shows how the probability distribution evolves from two different initial conditions, peaked at e_1 and e_2, respectively. Regardless of the starting point, the probability distributions eventually converge to identical steady-state distributions with a strong, sharp peak near e_1 and a weaker peak centered near e_2. When the initial condition is a peak at e_1, the probability spreads out over time and shifts some of its weight toward e_2, and vice-versa, although much more weight is shifted from e_2 to e_1 than the other direction.

The system behavior with Ω = 10 is qualitatively similar, as Figure 12 shows, but the bimodal steady-state probability distribution is even more sharply peaked at e_1 and the stochastic mean converges to an average closer to e_1, in agreement with the analysis of Figure 8.

The situation appears to be different for Ω = 100, as shown in Figure 13. In fact, the system seems to behave much more like a single stable steady-state system. In that the stochastic mean remains close to the initial steady-state, the van Kampen approximation agrees well with the simulation results, and the variance and difference between the mean and deterministic trajectory are both on the order of $O(\Omega^{-1})$. The explanation is that for very large systems, the probability of a jump between e_1 and e_2 is extremely small, so the escape time is much longer than the length of the simulation. Figure 11 confirms that the escape time scales exponentially with the system size, as discussed in the previous section and Appendix C. Therefore, for a large system like this one, the stochastic trajectories are highly unlikely to diverge from the deterministic steady-state where they originated for the duration of the simulation. If the simulation ran long enough, some trajectories would eventually escape from their initial domains of attraction, and the same blending of the two steady-states that we observed in the smaller systems would occur. The large system size means that the initial time period in which the two stable steady-states operate independently of each other takes up the entire simulation, however, so we never observe this blending.

Two gene system with two stable steady-states

We constructed a two gene system with two stable steady-states using the heuristic approach described in section **Constructing multistable systems**. We selected the two steady-states

$$e_1=\begin{bmatrix} 3 \\ 1 \end{bmatrix}, e_2=\begin{bmatrix} 2 \\ 4 \end{bmatrix}$$

and solved the optimization problem (35) for the coefficients b_i, c_i $c_i \in \mathbb{R}^m$ (where m is the number of terms in the model), for i = 1, 2, and Lyapunov matrices P_1, $P_2 \in \mathbb{R}^{n\times n}$, to enforce the steady-state condition ($f_i(e_j) = \gamma e_{j,i}$ for i = 1, 2, j = 1, 2), and the stability condition $(J_f(e_j)^T P_j + P_j J_f(e_j) \prec -\varepsilon$ at each steady-state $j = 1, 2$). The problem is non-convex due to the rational form of *f* and the stability condition. For convenience, we found a solution using a convex solver and a series of heuristics, then checked that the result was indeed a solution of (35), but we could also have used a general solver to solve the feasibility problem (35) directly.

The deterministic model for the two gene system with two stable steady-states is

$$\begin{aligned} \frac{\mathrm{d}y_i}{\mathrm{d}t}&=f_i(y)-\gamma y_i,\ \ f_i(y)=\frac{b^{(i)T}z(x)}{c^{(i)T}z(x)},\ \ i=1,2 \\ z(x)&=\begin{bmatrix} 1 & x_1 & x_2 & x_1x_2 & x_1^2 & x_2^2 \end{bmatrix}^T,\ \ \gamma=0.1, \end{aligned} \quad (37)$$

with b_i, c_i, $i = 1, 2$, given in Appendix D.

Figure 14 compares the second-order van Kampen approximation and the Gillespie and Langevin simulations for $\Omega = 10$ and $\Omega = 1000$ (where in the Langevin simulation, we again neglected the interaction terms in the second-order jump moments for simplicity). The qualitative behavior of this system is exactly the same as that of the bistable one gene system. For small system sizes, each stochastic trajectory has a reasonable probability of escaping from the domain of attraction of one stable steady-state and being attracted to the other, so in the long run, the system settles to a bimodal steady-state distribution. Hence, regardless of the initial condition, the steady-state mean converges to a weighted average of the two deterministic stable steady-states. The second-order van Kampen approximation centered at either of the two steady-states does not properly apply, and would seriously underestimate both the variance and the deviation of the mean trajectory from the deterministic trajectory. For very large systems, in contrast, the probability of a giant fluctuation between e_1 and e_2 is very small. Since the escape time scales exponentially with the system size, it can far exceed the length of the simulation for large systems. Therefore, the stochastic trajectories remain close to the deterministic steady-state where they originated for the duration of the simulation, and the van Kampen approximation is quite accurate within this timeframe.

Simulation summary

Our simulation studies support and illustrate the theory discussed in the last section by comparing the van Kampen expansion, Gillespie simulation, and Langevin simulation for systems with one or multiple stable steady-states, hence very different qualitative characteristics. For one gene systems, we can compare the performance of each approach to the exact trajectory of the Master equation. Our study of a one gene system with one stable steady-state shows that for system-size Ω, both the variance and the difference between the stochastic mean and deterministic trajectory are $O(\Omega^{-1})$, and the van Kampen expansion, Gillespie simulation and Langevin simulation are all in excellent agreement with Master equation, (except for slight inaccuracy in the van Kampen and Langevin approximations for very small systems). Furthermore, the deterministic and stochastic trajectories are almost identical for large systems. As the system size increases, the final probability distribution of the stochastic system becomes increasingly sharply peaked at the deterministic steady-state. The two gene system with one stable steady-state confirms these observations. The bistable systems exhibit much more complex behavior. Rather than staying near the initial deterministic steady-state, the Gillespie and Langevin simulations (and exact Master equation, for the one gene system) deviate dramatically from both the (improperly applied) van Kampen expansion and the deterministic trajectory, at least for small Ω. The explanation is that each stochastic trajectory has a reasonable probability of escaping from the domain of attraction of one stable steady-state and being attracted to another. In the long run, the system settles to a bimodal steady-state distribution, in which both stable steady-states are represented proportional to their relative stability, and the mean is the weighted average of the two stable steady-states (as predicted by the alternative van Kampen theory for multiple stable steady-states). However, for large bistable systems, the escape time can far exceed the length of the simulation, since escape time scales exponentially with system size. Therefore, the stochastic trajectories remain close to the deterministic steady-state where they originated for the duration of the simulation.

CONCLUSIONS

We can draw several important conclusions from the theory of the earlier sections and the results of these studies. The first is that for large systems with a single stable steady-state, the deterministic model is sufficient for most practical purposes, since the probability distribution of the stochastic solution consists of a peak tracking the deterministic solution with variance inversely proportional to the system size. Multistable systems display more complex behavior since large fluctuations can cause trajectories to jump from the domain of attraction of one steady-state into another. Eventually, multistable systems settle to multimodal steady-state probability distributions peaked at the deterministic steady-states, with peak strengths proportional to the relative stability of the steady-states. Moderate-sized or randomly initialized multistable systems reach their final multimodal distribution relatively quickly, but since the escape time scales exponentially with the system size, the steady-states of large multistable systems may operate independently of each other practically indefinitely.

These observations are particularly relevant to the deterministic model-based inference method we presented in a earlier publication [38]. Since the deterministic model is very accurate for large systems with a single steady-state, the inference method applies directly to the mean of gene expression measurements for this type of system. For multistable systems, we can find the deterministic steady-state expression levels needed for the algorithm by locating the expression peaks rather than averaging the measurements. For the large system sizes typical in gene expression studies, these peaks will be extremely close to the deterministic steady-states. It is also worth specifically relating the effects of stochasticity to gene perturbations, which are central to our inference algorithm and many other applications. A gene regulatory system immediately following a perturbation like gene knockdown is not in steady-state, so the expression distribution will be in flux for some period of time before reaching a final stochastic steady-state consistent with the perturbation. This steady-state is, in general, a multimodal distribution different from the system’s natural steady-state distribution due to the perturbation. The peaks of the distribution correspond to deterministic stable steady-states consistent with the fixed expression levels of the perturbed genes. If there is only one such deterministic steady-state, the final distribution will be unimodal; if there are multiple, the system will eventually explore them all. Generally, a perturbed system does not start out very close to a particular deterministic steady-state, so it has a reasonable probability of initial attraction to any possible state and the distribution quickly reaches its multimodal steady-state (on a very short time scale relative to the escape time, as discussed in section **Systems with multiple stable steady-states**). To collect data for the inference algorithm, the experimenter should apply each perturbation, wait for the system to settle to its stochastic steady-state distribution, and measure the expression peaks, which correspond to deterministic perturbed steady-states.

Stochastic effects become more dominant for small systems, where fluctuations have greater impact relative to the system as a whole. In particular, stochastic modeling can be critical for genes with very low expression numbers. In these cases, exact but expensive methods like the explicit Master equation solution for one gene systems or the Gillespie algorithm may be attractive. Our results indicate that the Langevin simulation is also reasonably accurate, especially for moderate-sized systems, at much lower computational cost than the Gillespie algorithm. For systems with one stable steady-state, the van Kampen expansion is excellent for approximating the Master equation at any level of detail desired, and alternative van Kampen theory can yield insight into the asymptotic behavior of multistable systems. We hope our discussion of gene regulation modeling via the Master equation and our analysis and demonstration of approximation and simulation methods will help future researchers treat stochasticity in gene regulation more confidently and effectively.

Acknowledgments

This work was supported by NIH (Grant No. HG 006018).

References

1. Swain PS, Elowitz MB, Siggia ED. Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci USA. 2002; 99:12795–12800. [PubMed: 12237400]
2. Paulsson J. Summing up the noise in gene networks. Nature. 2004; 427:415–418. [PubMed: 14749823]
3. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Sci Signal. 2002; 297:1183.
4. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. Regulation of noise in the expression of a single gene. Nat Genet. 2002; 31:69–73. [PubMed: 11967532]
5. Blake WJ, Kaern M, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. Nature. 2003; 422:633–637. [PubMed: 12687005]
6. Rao CV, Wolf DM, Arkin AP. Control, exploitation and tolerance of intracellular noise. Nature. 2002; 420:231–237. [PubMed: 12432408]
7. Kaern M, Elston TC, Blake WJ, Collins JJ. Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet. 2005; 6:451–464. [PubMed: 15883588]
8. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell. 2008; 135:216–226. [PubMed: 18957198]
9. Munsky B, Neuert G, van Oudenaarden A. Using gene expression noise to understand gene regulation. Science. 2012; 336:183–187. [PubMed: 22499939]
10. Hager GL, McNally JG, Misteli T. Transcription dynamics. Mol Cell. 2009; 35:741–753. [PubMed: 19782025]

11. Kittisopikul M, Süel GM. Biological role of noise encoded in a genetic network motif. Proc Natl Acad Sci USA. 2010; 107:13300–13305. [PubMed: 20616054]

12. Pedraza JM, van Oudenaarden A. Noise propagation in gene networks. Science. 2005; 307:1965–1969. [PubMed: 15790857]

13. Kepler TB, Elston TC. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophys J. 2001; 81:3116–3136. [PubMed: 11720979]

14. Ma R, Wang J, Hou Z, Liu H. Small-number effects: a third stable state in a genetic bistable toggle switch. Phys Rev Lett. 2012; 109:248107. [PubMed: 23368390]

15. Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators. Nature. 2000; 403:335–338. [PubMed: 10659856]

16. Gardner T, Cantor C, Collins J. Construction of a genetic toggle switch in *Escherichia coli*. Nature. 2000; 403:339–342. [PubMed: 10659857]

17. Hasty J, McMillen D, Collins JJ. Engineered gene circuits. Nature. 2002; 420:224–230. [PubMed: 12432407]

18. Ozbudak EM, Thattai M, Lim HN, Shraiman BI, Van Oudenaarden A. Multistability in the lactose utilization network of *Escherichia coli*. Nature. 2004; 427:737–740. [PubMed: 14973486]

19. Frigola D, Casanellas L, Sancho JM, Ibañes M. Asymmetric stochastic switching driven by intrinsic molecular noise. PLoS ONE. 2012; 7:e31407. [PubMed: 22363638]

20. Novak B, Tyson JJ. Modeling the control of DNA replication in fission yeast. Proc Natl Acad Sci USA. 1997; 94:9147–9152. [PubMed: 9256450]

21. Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. Genetics. 1998; 149:1633–1648. [PubMed: 9691025]

22. Thattai M, van Oudenaarden A. Intrinsic noise in gene regulatory networks. Proc Natl Acad Sci USA. 2001; 98:8614–8619. [PubMed: 11438714]

23. Tao Y. Intrinsic noise, gene regulation and steady-state statistics in a two gene network. J Theor Biol. 2004; 231:563–568. [PubMed: 15488533]

24. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB. Gene regulation at the single-cell level. Sci Signal. 2005; 307:1962.

25. Krishnamurthy S, Smith E, Krakauer D, Fontana W. The stochastic behavior of a molecular switching circuit with feedback. Biol Direct. 2007; 2:1–17. [PubMed: 17222345]

26. Munsky B, Trinh B, Khammash M. Listening to the noise: random fluctuations reveal gene network parameters. Mol Syst Biol. 2009; 5:318. [PubMed: 19888213]

27. Dunlop MJ, Cox RS 3rd, Levine JH, Murray RM, Elowitz MB. Regulatory activity revealed by dynamic correlations in gene expression noise. Nat Genet. 2008; 40:1493–1498. [PubMed: 19029898]

28. Stewart-Ornstein J, Weissman JS, El-Samad H. Cellular noise regulons underlie fluctuations in *Saccharomyces cerevisiae*. Mol Cell. 2012; 45:483–493. [PubMed: 22365828]

29. Van Kampen, NG. Stochastic Processes in Physics and Chemistry. 3. North Holland; 2007.

30. Peles S, Munsky B, Khammash M. Reduction and solution of the chemical Master equation using time scale separation and finite state projection. J Chem Phys. 2006; 125:204104. [PubMed: 17144687]

31. Hegland M, Burden C, Santoso L, MacNamara S, Booth H. A solver for the stochastic master equation applied to gene regulatory networks. J Comput Appl Math. 2007; 205:708–724.

32. Macnamara S, Bersani AM, Burrage K, Sidje RB. Stochastic chemical kinetics and the total quasi-steady-state assumption: application to the stochastic simulation algorithm and chemical master equation. J Chem Phys. 2008; 129:095105. [PubMed: 19044893]

33. Smadbeck P, Kaznessis Y. Stochastic model reduction using a modified Hill-type kinetic rate law. J Chem Phys. 2012; 137:234109. [PubMed: 23267473]

34. Waldherr S, Wu J, Allgöwer F. Bridging time scales in cellular decision making with a stochastic bistable switch. BMC Syst Biol. 2010; 4:108. [PubMed: 20696063]

35. Liang J, Qian H. Computational cellular dynamics based on the chemical master equation: A challenge for understanding complexity. Journal of Computer Science and Technology. 2010; 25:154–168. [PubMed: 24999297]

36. Gutierrez PS, Monteoliva D, Diambra L. Cooperative binding of transcription factors promotes bimodal gene expression response. PLoS ONE. 2012; 7:e44812. [PubMed: 22984566]

37. Khanin R, Higham DJ. Chemical Master Equation and Langevin regimes for a gene transcription model. Theor Comput Sci. 2008; 408:31–40.

38. Meister A, Li YH, Choi B, Wong WH. Learning a nonlinear dynamical system model of gene regulation: A perturbed steady-state approach. Ann Appl Stat. 2013; 7:1311–1333.

39. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007; 5:e8. [PubMed: 17214507]

40. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007; 3:78. [PubMed: 17299415]

41. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. Inferring genetic networks and identifying compound mode of action via expression profiling. Science. 2003; 301:102–105. [PubMed: 12843395]

42. di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat Biotechnol. 2005; 23:377–383. [PubMed: 15765094]

43. Michaelis L, Menten ML. Die kinetik der invertinwirkung. Biochem Z. 1913; 49:333–369.

44. Hill AV. The combinations of haemoglobin with oxygen and with carbon monoxide. Biochem J. 1913; 7:471–480. [PubMed: 16742267]

45. Ackers GK, Johnson AD, Shea MA. Quantitative model for gene regulation by lambda phage repressor. Proc Natl Acad Sci USA. 1982; 79:1129–1133. [PubMed: 6461856]

46. Shea MA, Ackers GK. The OR control system of bacteriophage lambda: A physicalchemical model for gene regulation. J Mol Biol. 1985; 181:211–230. [PubMed: 3157005]

47. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Phillips R. Transcriptional regulation by the numbers: models. Curr Opin Genet Dev. 2005; 15:116–124. [PubMed: 15797194]

48. Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R. Transcriptional regulation by the numbers: applications. Curr Opin Genet Dev. 2005; 15:125–135. [PubMed: 15797195]

49. Rao CV, Arkin AP. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. J Chem Phys. 2003; 118:4999–5010.

50. Walker, JA. Dynamical Systems and Evolution Equations. New York: Plenum Press; 1939.

51. Kubo R, Matsuo K, Kitahara K. Fluctuation and relaxation of macrovariables. J Stat Phys. 1973; 9:51–96.

52. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977; 81:2340–2361.

53. Gillespie DT. The chemical Langevin equation. J Chem Phys. 2000; 113:297.

54. Komorowski M, Finkenstädt B, Harper CV, Rand DA. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. BMC Bioinformatics. 2009; 10:343. [PubMed: 19840370]

55. Choi, B. PhD thesis. Department of Applied Physics, Stanford University; Stanford: 2012. Learning networks in biological systems.

56. Planck, M. Verband Deutscher Physikalischer Gesellschaften. Physikalische abhandlungen und vorträge. 1958.

57. Rayleigh, Lord. Liii. Dynamical problems in illustration of the theory of gases. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1891; 32:424–445.

58. Einstein A. Eine neue bestimmung der molek uldimensionen. Annalen der Physik. 1906; 324:289–306.

59. Von Smoluchowski M. Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen. Annalen der physik. 1906; 326:756–780.

60. Van Kampen, NG. Fluctuation Phenomena in Solids. New York: Academic Press; 1965. Fluctuations in Nonlinear Systems.

61. Bar-Haim A, Klafter J. Geometric versus energetic competition in light harvesting by dendrimers. J Phys Chem B. 1998; 102:1662–1664.

62. Chickarmane V, Peterson C. A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. PLoS ONE. 2008; 3:e3478. [PubMed: 18941526]

63. Zavlanos MM, Julius AA, Boyd SP, Pappas GJ. Inferring stable genetic networks from steady-state data. Automatica. 2011; 47:1113–1122.

APPENDICES

A Derivation of the Master equation

The following derivation, simplified and adapted from Chapters IV and X of van Kampen's *Stochastic Processes in Physics and Chemistry* [29], is provided here for the reader's convenience.

A.1 The Chapman-Kolmogorov equation

A Markov process is a stochastic process such that for any $t_1 < t_2 < \cdots < t_n$,

$$P(y_n,t_n|y_1,t_1;\cdots;y_{n-1},t_{n-1})| =P(y_n,t_n|y_{n-1},t_{n-1}).$$

Hence a Markov process is completely determined by the functions $P(y_1, t_1)$ and the transition probabilities $P(y_2, t_2|y_1, t_1)$. For example, for any $t_1<t_2<t_3$,

$$\begin{aligned}P(y_1,t_1;y_2,t_2;y_3,t_3)&=P(y_1,t_1;y_2,t_2)P(y_3,t_3|y_1,t_1;y_2,t_2)\\&=P(y_1,t_1)P(y_2,t_2|y_1,t_1)P(y_3,t_3|y_2,t_2).\end{aligned}$$

If we integrate this identity over y_2 and divide both sides by $P(y_1, t_1)$, we obtain the Chapman-Kolmogorov equation, which necessarily holds for any Markov process:

$$\begin{aligned}P(y_1,t_1;y_3,t_3)=P(y_1,t_1)\int P(y_2,t_2|y_1,t_1)P(y_3,t_3|y_2,t_2)\mathrm{d}y_2\\ \Rightarrow P(y_3,t_3|y_1,t_1)\\ =\int P(y_2,t_2|y_1,t_1)P(y_3,t_3|y_2,t_2)\mathrm{d}y_2.\end{aligned} \tag{38}$$

A.2 The Master equation

The Master equation (also known as the Kolmogorov Forward equation) is a differential form of the Chapman-Kolmogorov equation that is often more convenient and easier to relate to physical processes.

In order to derive it, we first assume for convenience that the process is time homogeneous, so we can write the transition probabilities as T_τ, i.e.,

$$T_\tau(y_2|y_1) \equiv P(y_2, t{+}\tau|y_1, t).$$

It can be shown (see van Kampen IV.6) that for small τ', $T_{\tau'}(y_2|y_1)$ has the form

$$T_{\tau'}(y_2|y_1) = (1 - a_0\tau')\delta_{y_2,y_1} + \tau' W(y_2|y_1) + o(\tau'), \quad (39)$$

where $W(y_2|y_1)$ is the $y_1 \to y_2$ transition probability per unit time. The coefficient in front of the delta is the probability that no transition occurs during τ', so

$$a_0(y_1) = \int W(y_2|y_1)\mathrm{d}y_2.$$

Inserting (39) in place of T'_τ in the Chapman-Kolmogorov equation (38) yields

$$\begin{gathered}T_{\tau+\tau'}(y_3|y_1) = \int T_{\tau'}(y_3|y_2)T_\tau(y_2|y_1)\mathrm{d}y_2 \\ = \int((1 - a_0(y_2)\tau')\delta_{y_3,y_2} + \tau' W(y_3|y_2))T_\tau(y_2|y_1)\mathrm{d}y_2 \\ = (1 - a_0(y_3)\tau')T_\tau(y_3|y_1) + \tau' \int W(y_3|y_2)T_\tau(y_2|y_1)\mathrm{d}y_2\end{gathered}$$

$$\begin{gathered}\tfrac{\partial}{\partial\tau}T_\tau(y_3|y_1) = \lim_{\tau' \to 0} \frac{T_{\tau+\tau'}\ (y_3|y_1) - T_\tau(y_3|y_1)}{\tau'} \\ = -a_0(y_3)T_\tau(y_3|y_1) + \int W(y_3|y_2)T_\tau(y_2|y_1)\mathrm{d}y_2 \\ = \int\{W(y_3|y_2)T_\tau(y_2|y_1) - W(y_3|y_2)T_\tau(y_3|y_1)\}\mathrm{d}y_2.\end{gathered}$$

We can rewrite this equation as (7) from the main text as follows:

$$\begin{gathered}P(y_3, \tau) = \int T_\tau(y_3|y_1)P(y_1, 0)\mathrm{d}y_1 \text{ as } \tau \to 0 \\ \Rightarrow \tfrac{\partial P(y_3,\tau)}{\partial\tau} = \int \tfrac{\partial}{\partial\tau}T_\tau(y_3|y_1)P(y_1, 0)\mathrm{d}y_1 \\ = \iint P(y_1, 0)\{W(y_3|y_2)T_\tau(y_2|y_1) - W(y_2|y_3)T_\tau(y_3|y_1)\mathrm{d}y_1\mathrm{d}y_2 \\ = \int\{W(y_3|y_2)P(y_2, \tau) - W(y_2|y_3)P(y_3|\tau)\}\mathrm{d}y_2.\end{gathered}$$

Or change the names of the variables:

$$\frac{\partial P(y,t)}{\partial t} = \int\{W(y|y')P(y', t) - W(y'|y)P(y, t)\}\mathrm{d}y'.$$

B van Kampen's expansion of the Master equation

This calculation is adapted from van Kampen, Chapter X [29]. It is simplified from the original by assuming a birth-and- death process, and provided here for the reader's convenience. The Master equation for a birth-and-death process is given by

$$\frac{\partial P(X,t)}{\partial t}=W(X|X-1)P(X-1,t)+W(X|X+1)P(X+1,t)-[W(X+1|X)+W(X-1|X)]P(X,t).$$

Assume that the transition probabilities have the special form:

$$W_{\Omega}(X+r|X)=\Omega\Phi_0\left(\frac{X}{\Omega};r\right),$$

and define

$$a_v(x)=\sum_r r^v\Phi_0(x;r).$$

For birth-and-death processes, *r* only takes the values ±1, so we have

$$\begin{gathered}W_{\Omega}(X+1|X)=\Omega\Phi_0\left(\tfrac{X}{\Omega};+1\right)\equiv\Omega\Phi_0^+\left(\tfrac{X}{\Omega}\right),\\ W_{\Omega}(X-1|X)=\Omega\Phi_0\left(\tfrac{X}{\Omega};-1\right)\equiv\Omega\Phi_0^-\left(\tfrac{X}{\Omega}\right),\\ a_1(\phi)=\Phi_0^+(\phi(t))-\Phi_0^-\phi(t),\\ a_2(\phi)=\Phi_0^+(\phi(t))+\Phi_0^-(\phi(t)),\\ a_v(\phi)=\Phi_0^+(x)+(-1)^v\Phi_0^-(x).\end{gathered}$$

Hence the Master equation becomes

$$\frac{\partial P(X,t)}{\partial t}=\Omega\left\{\Phi_0^+\left(\frac{X-1}{\Omega}\right)P(X-1,t)+\Phi_0^-\left(\frac{X+1}{\Omega}\right)P(X+1,t)-\left(\Phi_0^+\left(\frac{X}{\Omega}\right)+\Phi_0^-\left(\frac{X}{\Omega}\right)\right)P(X,t)\right\}. \quad (40)$$

As discussed in the main text, we make the Ansatz

$$X(t)=\Omega\phi(t)+\Omega^{\frac{1}{2}}\xi$$

and define Π by

$$p(X,t)=P\left(\Omega\phi(t)+\Omega^{\frac{1}{2}}\xi\right)\equiv\Pi(\xi,t).$$

The partial derivatives Π are given by

$$\begin{gathered}\frac{\partial^v\Pi}{\partial\xi^v}=\Omega^{\frac{v}{2}}\frac{\partial^v P}{\partial X^v}\\ \frac{\partial\Pi}{\partial t}=\frac{\partial P}{\partial t}+\Omega\frac{\mathrm{d}\phi}{\mathrm{dt}}\frac{\partial P}{\partial X}=\frac{\partial P}{\partial t}+\Omega^{\frac{1}{2}}\frac{\mathrm{d}\phi}{\mathrm{dt}}\frac{\partial\Pi}{\partial\xi}.\end{gathered}$$

Therefore we can rewrite (40) as

$$\begin{gathered}\frac{\partial\Pi}{\partial t}-\Omega^{\frac{1}{2}}\frac{\mathrm{d}\phi}{\mathrm{d}t}\frac{\partial\Pi}{\partial\xi}\\ =\Omega\left\{\Phi_0^+\left(\phi(t)+\Omega^{-\frac{1}{2}}\left(\xi-\Omega^{-\frac{1}{2}}\right)\right)\Pi\left(\xi-\Omega^{-\frac{1}{2}},t\right)\right.\\ +\Phi_0^-\left(\phi(t)+\Omega^{-\frac{1}{2}}\left(\xi+\Omega^{-\frac{1}{2}}\right)\right)\Pi\left(\xi+\Omega^{-\frac{1}{2}},t\right)\\ \left.-\left(\Phi_0^+\left(\phi(t)+\Omega^{-\frac{1}{2}}\xi\right)+\Phi_0^-\left(\phi(t)+\Omega^{-\frac{1}{2}}\xi\right)\right)\Pi(\xi,t)\right\}\end{gathered}$$

Taylor expanding $\Phi_0^{\pm}\left(\phi+\Omega^{-\frac{1}{2}}\left(\xi-\Omega^{-\frac{1}{2}}\right)\Pi\left(\xi-\Omega^{-\frac{1}{2}}\right)\right.$ about ξ allows us to approximate (40) in terms of the jump moments a_1, a_2:

$$\begin{gathered}\frac{\partial\Pi}{\partial t}-\Omega^{\frac{1}{2}}\frac{\mathrm{d}\phi}{\mathrm{d}t}\frac{\partial\Pi}{\partial\xi}=-\Omega^{\frac{1}{2}}\frac{\partial}{\partial\xi}\left[\alpha_1\left(\phi(t)+\Omega^{-\frac{1}{2}}\xi\right)\Pi(\xi,t)\right]\\ +\frac{\Omega^0}{2!}\frac{\partial^2}{\partial\xi^2}\left[\alpha_2\left(\phi+\Omega^{-\frac{1}{2}}\xi\right)\Pi(\xi,t)\right]\\ -\frac{\Omega^{-\frac{1}{2}}}{3!}\frac{\partial^3}{\partial\xi^3}\left[\alpha_3\left(\phi+\Omega^{-\frac{1}{2}}\xi\right)\Pi(\xi,t)\right]\\ +O(\Omega^{-1}).\end{gathered}$$

A second Taylor expansion of $\alpha_1\left(\phi+\Omega^{-\frac{1}{2}}\xi\right)$ about φ gives

$$\begin{gathered}\frac{\partial\Pi}{\partial t}-\Omega^{\frac{1}{2}}\frac{\mathrm{d}\phi}{\mathrm{d}t}\frac{\partial\Pi}{\partial\xi}\\ =-\Omega^{\frac{1}{2}}\alpha_1(\phi)\frac{\partial\Pi}{\partial\xi}-{\alpha'}_1(\phi)\frac{\partial\xi\Pi}{\partial\xi}-\frac{1}{2}\Omega^{-\frac{1}{2}}{\alpha''}_1(\phi)\frac{\partial\xi^2\Pi}{\partial\xi}\\ +\frac{1}{2}\alpha_2(\phi)\frac{\partial^2\Pi}{\partial\xi^2}+\frac{1}{2}\Omega^{-\frac{1}{2}}{\alpha'}_2(\phi)\frac{\partial^2\xi\Pi}{\partial\xi^2}\\ -\frac{\Omega^{-\frac{1}{2}}}{3!}\alpha_3(\phi)\frac{\partial^3\Pi}{\partial\xi^3}+O(\Omega^{-1}).\end{gathered}$$

We can cancel the $O\left(\Omega^{\frac{1}{2}}\right)$ terms on the right- and left-hand- sides by choosing

$$\begin{gathered}\frac{\mathrm{d}\phi}{\mathrm{d}t}=\alpha_1(\phi)\\ \Rightarrow\frac{\partial\Pi}{\partial t}=-{\alpha'}_1(\phi)\frac{\partial\xi\Pi}{\partial\xi}+\frac{1}{2}\alpha_2(\phi)\frac{\partial^2\Pi}{\partial\xi^2}\\ +\frac{1}{2}\Omega^{-\frac{1}{2}}\left({\alpha'}_2(\phi)\frac{\partial^2\xi\Pi}{\partial\xi^2}-{\alpha''}_1(\phi)\frac{\partial\xi^2\Pi}{\partial\xi}-\frac{1}{3}\alpha_3(\phi)\frac{\partial^3\Pi}{\partial\xi^3}\right)+O(\Omega^{-1}).\end{gathered}$$

This is the final form of the expansion. It can be truncated at any level of detail desired and translated back into the original variables to yield various approximations of the Master equation. Note that it is only applicable for systems with a single stable steady-state [29].

C Mean first-passage time

For a birth-and-death process with states 0, 1, 2, …, we can derive a simple formula for the mean first-passage time. Suppose the system starts at state *m* and we want to find the mean first-passage time to state *n*. Let τ_i denote the expected time to reach state *n* starting from state *i*. Clearly $\tau_n = 0$, and the quantity of interest is τ_m. Let g_k, r_k denote the birth and death rates of the chain, respectively, and t_k denote the expected waiting time in state k before a transition. The waiting times and transition probabilities are related to the rates as follows:

$$t_k=\frac{1}{g_k+r_k}, \mathbb{P}(k \rightarrow k+1)=t_k g_k, \mathbb{P}(k \rightarrow k-1)=t_k r_k.$$

Then we have

$$\begin{gathered}\tau_k=t_k(r_k\tau_{k-1}+g_k\tau_{k+1}+1), k=0,\ldots,n-1\\ \Rightarrow \tau_{k+1}-\tau_k=\tfrac{1}{g_k}[r_k(\tau_k-\tau_{k-1})-1],\\ \text{by nothing that } \tau_k(g_k+r_k)=1\\ \Rightarrow \tau_{k+1}-\tau_k=\tfrac{1}{g_k}\sum_{i=0}^{k}\prod_{j=k}^{i-1}\tfrac{g_j}{r_j}=\tfrac{1}{g_k p_k^s}\sum_{i=0}^{k}p_i^s\\ \Rightarrow \tau_m=\sum_{k=m}^{n-1}(\tau_{k+1}-\tau_k)=\sum_{k=m}^{n-1}\tfrac{1}{g_k p_k^s}\sum_{i=0}^{k}p_i^s,\end{gathered}$$

where p^s is the stationary distribution (11). Observe that if n, m are stable points and l with $m \le l \le n$ is an unstable point, then the stationary distribution will have peaks at n and m and a valley at l. The most important terms in the sum are therefore those with P_l^s being the denominator, and the inner sum is then π_m, which is $O(1)$. Hence the escape rate is on the order of P_l^s; that is,

$$\tau_m \sim O\left(\frac{1}{P_l^s}\right) \sim O(e^{\Omega}).$$

The escape time scales as e^{Ω} since the stationary distribution is approximately a mixture of Gaussians with peaks of order Ω at the stable points, so P_l^s is $O(e^{-\Omega})$.

D Coefficients of the bistable two gene system

The coefficients of system (37) are given by

$$\left[\begin{array}{cccc} b_1 & c_1 & b_2 & c_2 \end{array}\right]=\left[\begin{array}{llll} 0.3991 & 1 & 0.0557 & 1\\ 0.2271 & 0.6814 & 0.0173 & 0.3009\\ 0.1485 & 0.6703 & 0.0369 & 0.2304\\ 0.0672 & 0.3161 & 0.0127 & 0.0866\\ 0.1035 & 0.3283 & 0.0059 & 0.1531\\ 0.0375 & 0.3821 & 0.1648 & 0.2295 \end{array}\right]$$

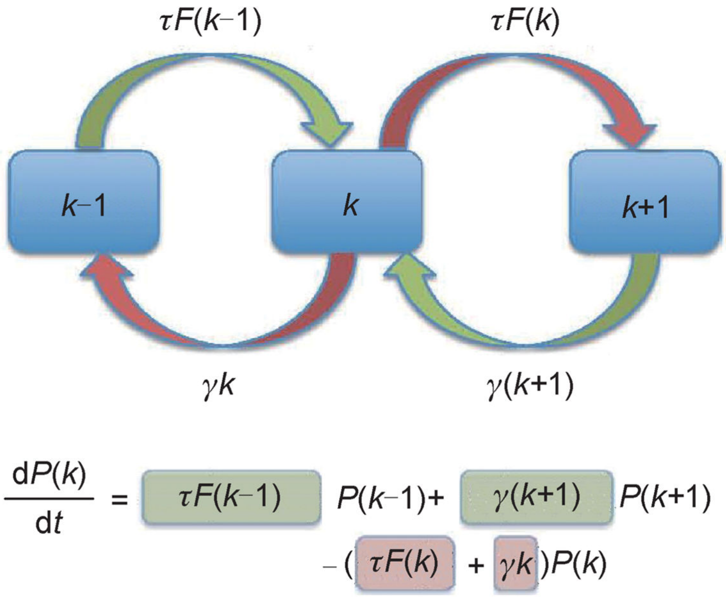

Figure 1. Informal derivation of the Master equation for gene regulation

In a infinitesimal timestep, $P(k; t)$, the probability of k RNA transcripts, increases by $P(k - 1, t)$ times the probability, $F(k - 1)$, of a transcription event (number of transcripts increases by one) plus $P(k + 1, t)$ times the probability, $\gamma(k + 1)$, of a degradation event (number of transcripts decreases by one). It decreases by $P(k)$ times the probability of transcription plus $P(k)$ time the probability of degradation.

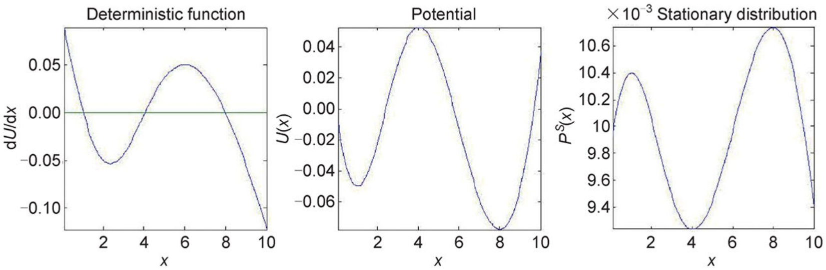

Figure 2. Bistability in a stochastic system modeled by a Fokker-Planck equation of the form (28), corresponding to deterministic equation $\frac{\mathrm{d}x}{\mathrm{d}t}=\frac{\mathrm{d}U}{\mathrm{d}t}$

The deterministic function $\frac{\mathrm{d}U}{\mathrm{d}t}$ (left) has zeros at the three steady-states $\varphi_a \approx 1$, $\varphi_b \approx 4$, $\varphi_c \approx 8$. The points φ_a and φ_c are stable, while φ_b is unstable. The potential $U(x)$ (center) has minima at φ_a and φ_c and a maximum at φ_b, corresponding to low energy (favorable) at the two steady-states and high energy (unfavorable) at the unstable state. φ_c is more stable than φ_a since its potential well is deeper and wider. The stationary distribution (right), to which the stochastic system will eventually converge, is bimodal with peaks at φ_a and φ_c. The peak at φ_c is higher since φ_c is more stable than φ_a.

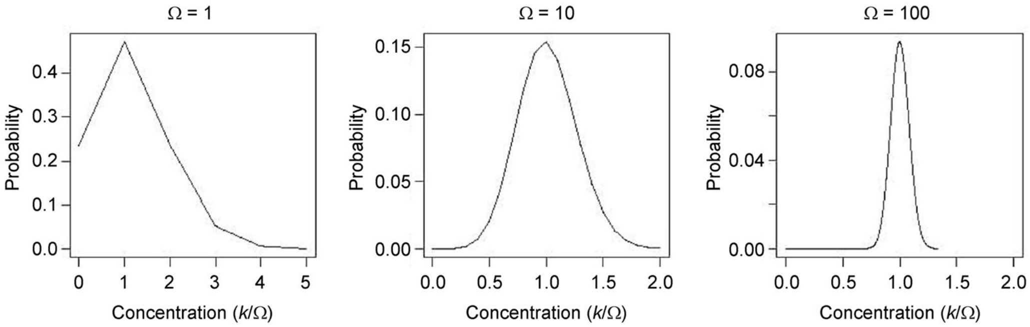

Figure 3. Steady-state probability distributions of a one gene system with one steady-state (31) for increasing system sizes Ω = 1,10,100

The distribution always peaks at the deterministic steady-state solution ($y = 1$), and the variance decreases as Ω increases. For smaller values of Ω, it is clear that the mean lies slightly above the deterministic solution, but as Ω increases, the distribution becomes quite symmetric.

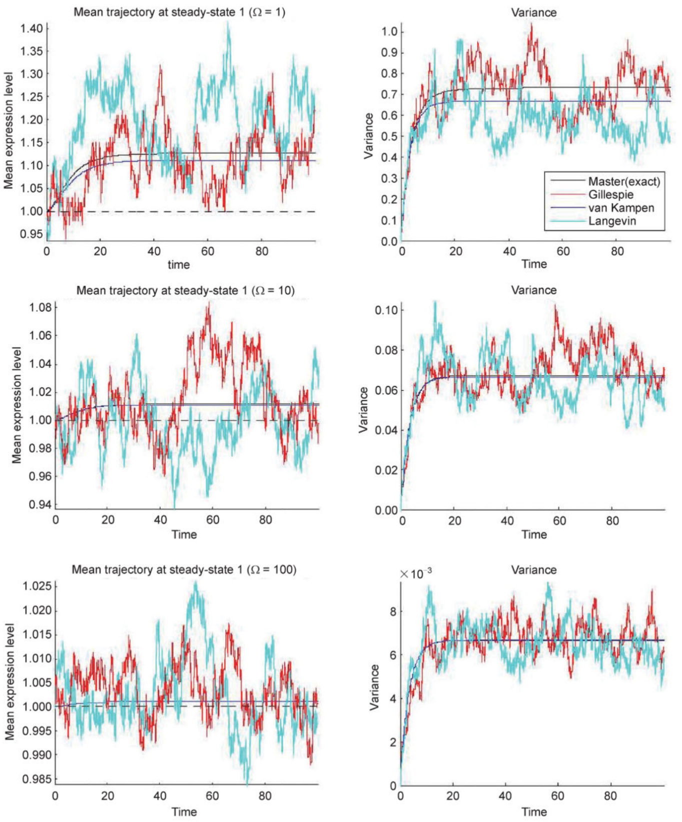

Figure 4. One gene system with one steady-state (31)

Mean (left) and variance (right) trajectories via Master equation (black), van Kampen approximation (blue), and average of 100 trajectories of the Gillespie (red) and Langevin simulation (cyan) with Ω = 1,10,100 (top to bottom, respectively). There is excellent agreement between simulations, van Kampen approximation, and exact Master equation for both mean and variance. Discrepancy between the stochastic mean and deterministic trajectory and magnitude of the variance are both $O(\Omega^{-1})$.

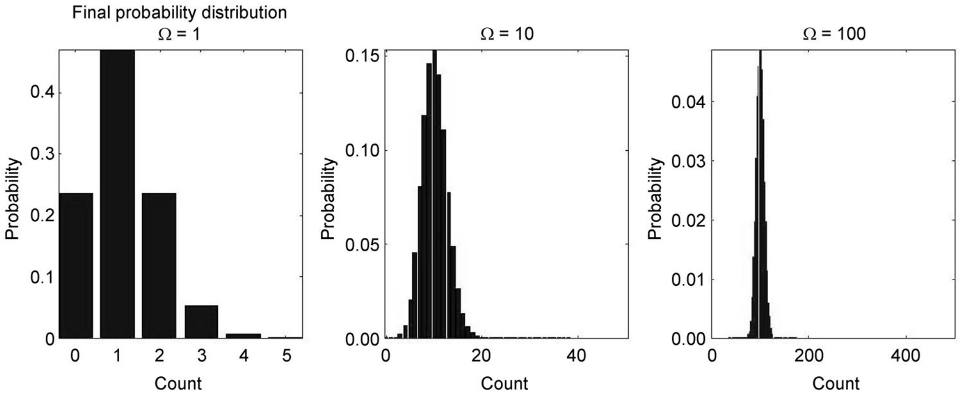

Figure 5. Final probability distributions of the exact Master equation for a one gene system with one steady-state (31), with Ω = 1,10,100

The probabilities converge to approximately Gaussian steady-state distributions peaked near the deterministic steady-state. For larger system sizes, the distribution is more Gaussian and the peak is sharper.

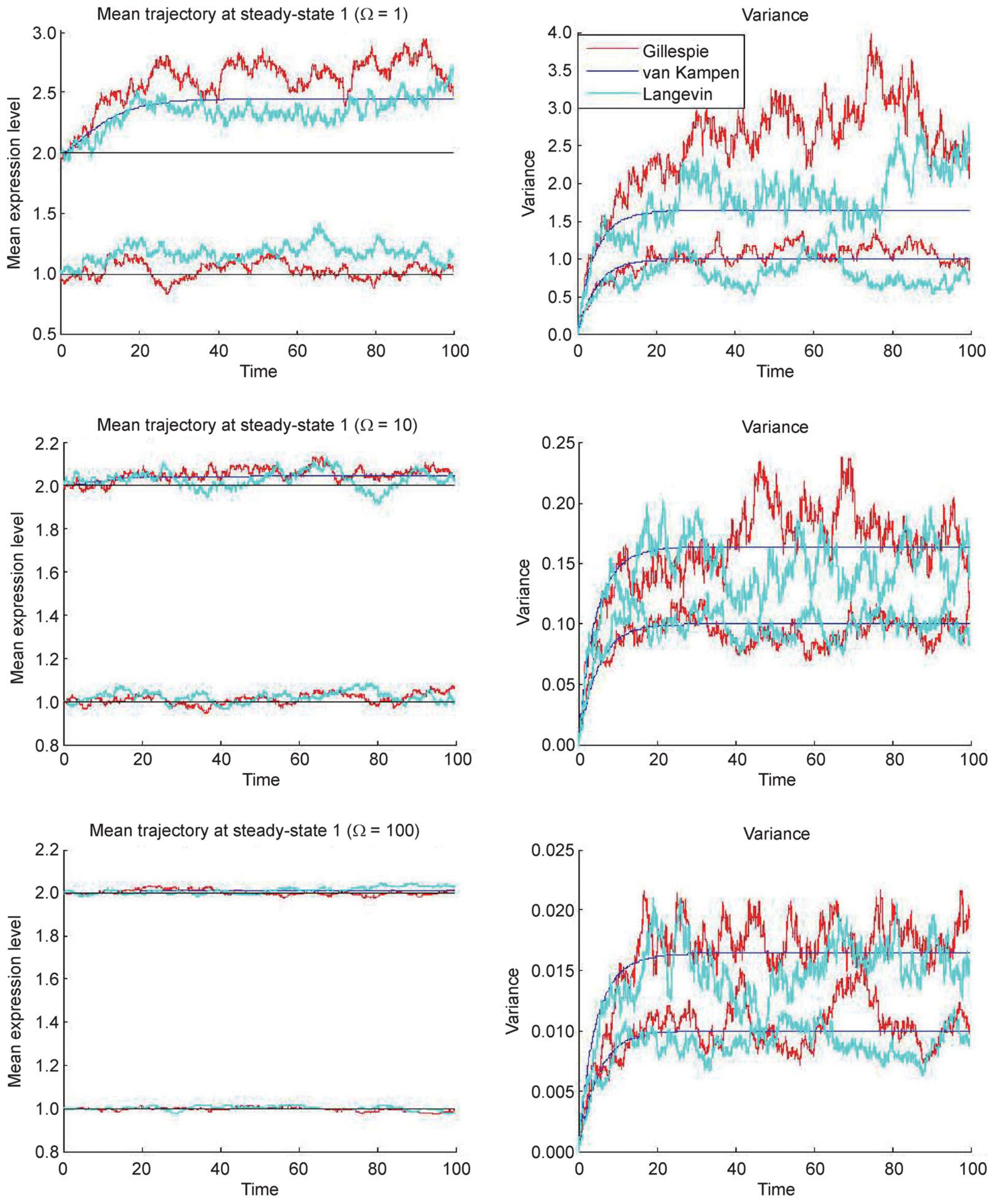

Figure 6. Two gene system with one stable steady-state (34)

Mean (left) and variance (right) trajectories of van Kampen approximation (blue) and average of 100 trajectories of Gillespie (red) and Langevin simulation (cyan) with Ω = 1,10,100 (top to bottom, respectively). As with the one gene system, agreement between the simulations and the van Kampen approximation is excellent, and both the variance and the discrepancy between the mean and deterministic trajectory are $O(\Omega^{-1})$. The only exception is for Ω = 1, where slight inaccuracy of the Langevin simulation and van Kampen expansion arises from the non-Gaussianity of the probability distribution.

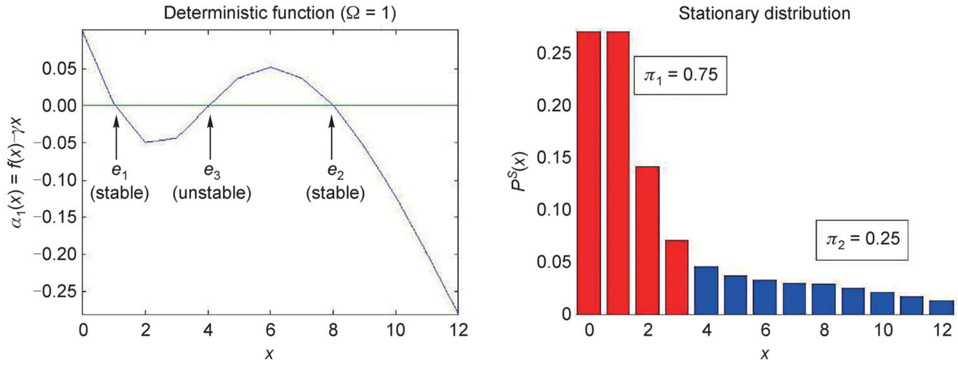

Figure 7. The deterministic function $\alpha_1(x) = f(x) - \gamma x$ for the system (36), with Ω= 1, has three zeros corresponding to the three deterministic steady-states, e_1, e_2, e_3

The derivative of the deterministic function is negative (dα_1 = dt < 0) at the stable steady-states e_1, e_2, and positive at the unstable steady-state e_3. The stationary distribution (computed with Equation (12)) has a strong peak at e_1 and a weaker one at e_2. The system is much more likely to be in the domain of e_1 ($x < e_3$) than in the domain of e_2 ($x > e_3$): specifically, $\pi_1 \approx 0.75$, and $\pi_2 \approx 0.25$. The steady-state mean is given by $\pi_1 e_1 + \pi_2 e_2 \approx 2.78$.

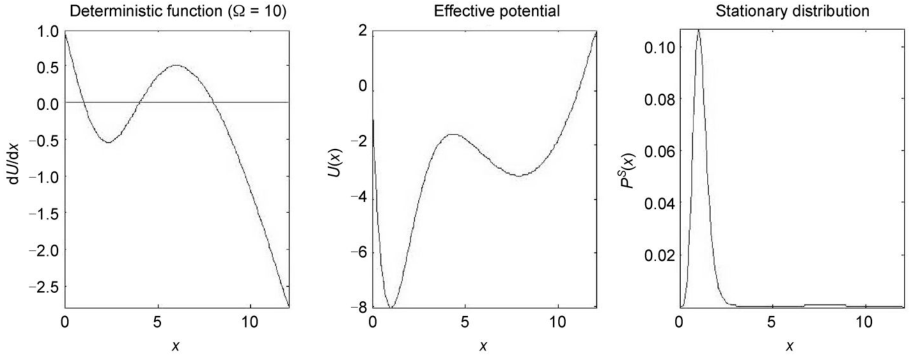

Figure 8. The deterministic function, effective potential, and (approximate) stationary distribution for system (36) with Ω = 10, computed with the Fokker-Planck approximation and Equations (29,30)

(The result is nearly identical to what we would have obtained with the explicit equation (12)). The deterministic function and stationary distribution have the same qualitative properties as they did with Ω= 1, except that the e_1 peak in the stationary distribution is now even higher relative to e_2 ($\pi_1 \approx 97\%$; $\pi_2 \approx 3\%$), and the steady-state mean is shifted toward e_1: $\pi_1 e_1 + \pi_2 e_2 \approx 1.24$. The effective potential has minima at the two stable steady-states e_1, e_2, and a maximum at e_3. The more stable steady-state, e_1, has lower “energy”.

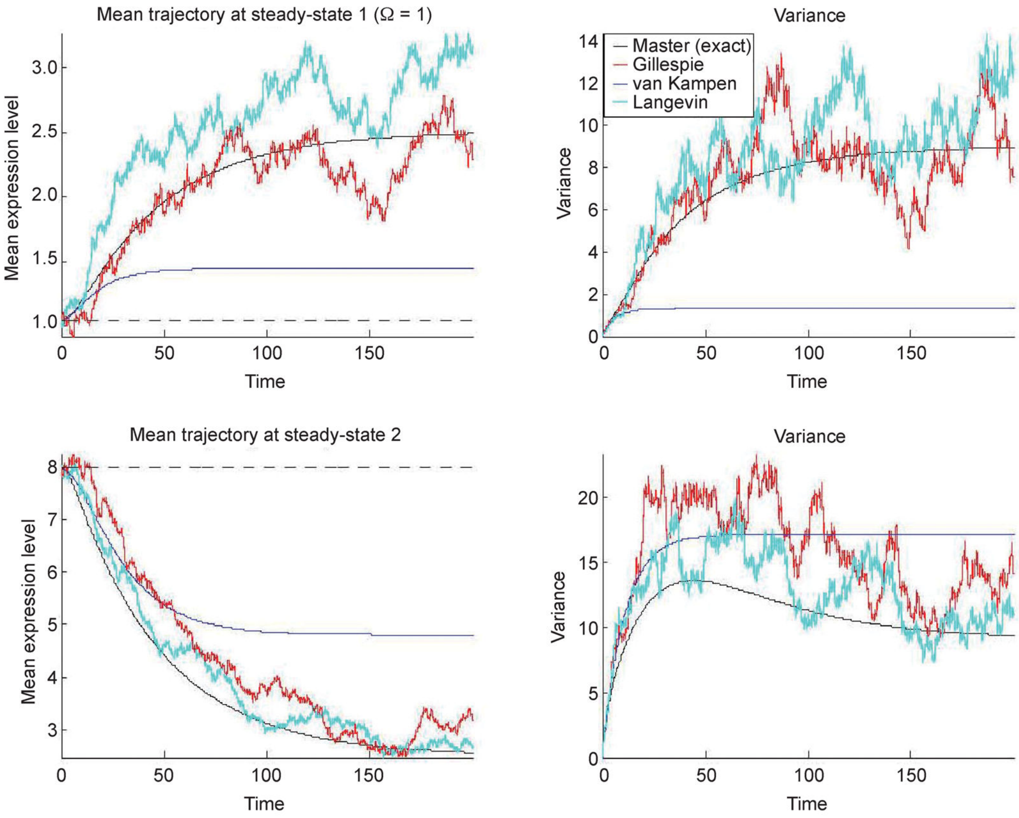

Figure 9. One gene system with two stable deterministic steady-states (36), Ω = 1

Mean (left) and variance (right) trajectories via the Master equation (black), the (improperly applied) van Kampen expansion (blue), and the average of 100 trajectories of the Gillespie (red) and Langevin simulation (cyan). Regardless of the starting point, the stochastic mean trajectory eventually converges to the weighted average of the two deterministic stable steady-states predicted by the analysis of Figure 7: $\pi_1 e_i \approx + \pi_2 e_2 \approx 2.78$. The (improperly applied) van Kampen expansion seriously underestimates the discrepancy between the mean and the deterministic trajectory since, as an expansion about e_1, it effectively ignores e_2, and vice versa; van Kampen's stability analysis is therefore the correct theoretical approach in this case.

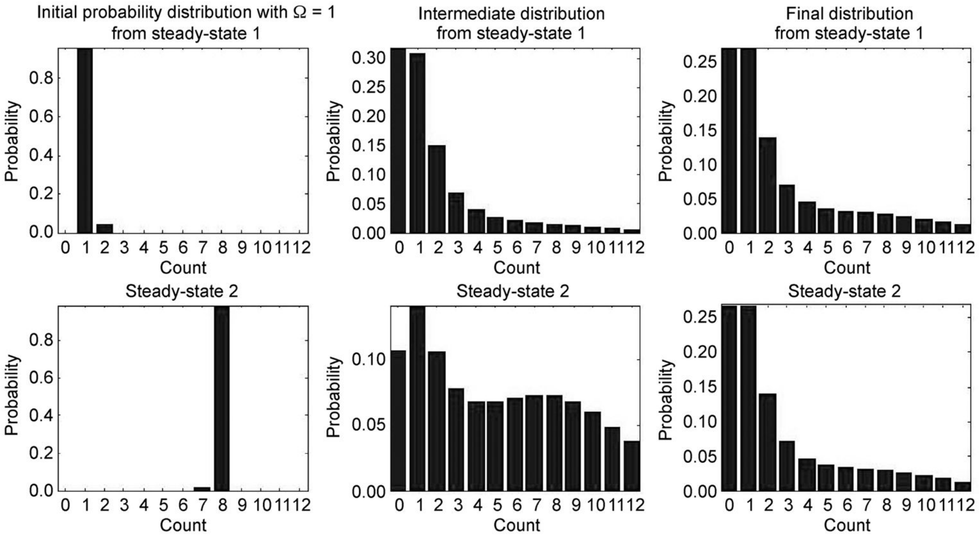

Figure 10. Initial (left), intermediate (center), and final (right) probability distributions of the exact Master equation for the one gene system with two stable steady-states (36), starting from e_1 (top) or e_2 (bottom), with Ω= 1

The probability distributions start out peaked at their respective initial conditions. Over time, some of the probability begins to flow from one deterministic steady-state to the other. Regardless of the initial condition, the system eventually reaches a single bimodal stochastic steady-state (the same distribution shown in Figure 7), with a stronger peak at e_1 (the more stable of the two points) and a weaker peak at e_2.

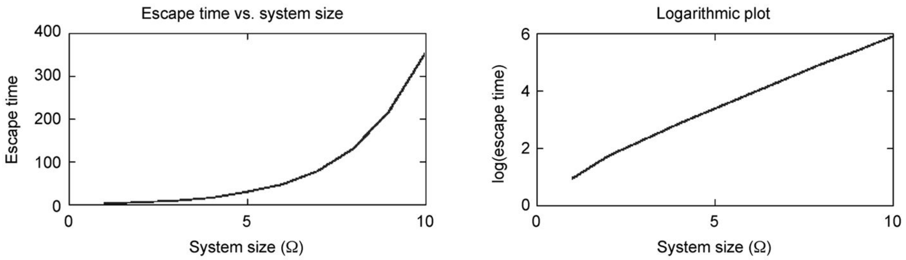

Figure 11.

Escape time $\tau_{2,1}$ versus system size Ω for system (36) (left), computed as mean first-passage time via Equation (27). The plot of log($\tau_{2,1}$) vs. Ω(right) is linear, confirming that the escape time grows exponentially with the system size.

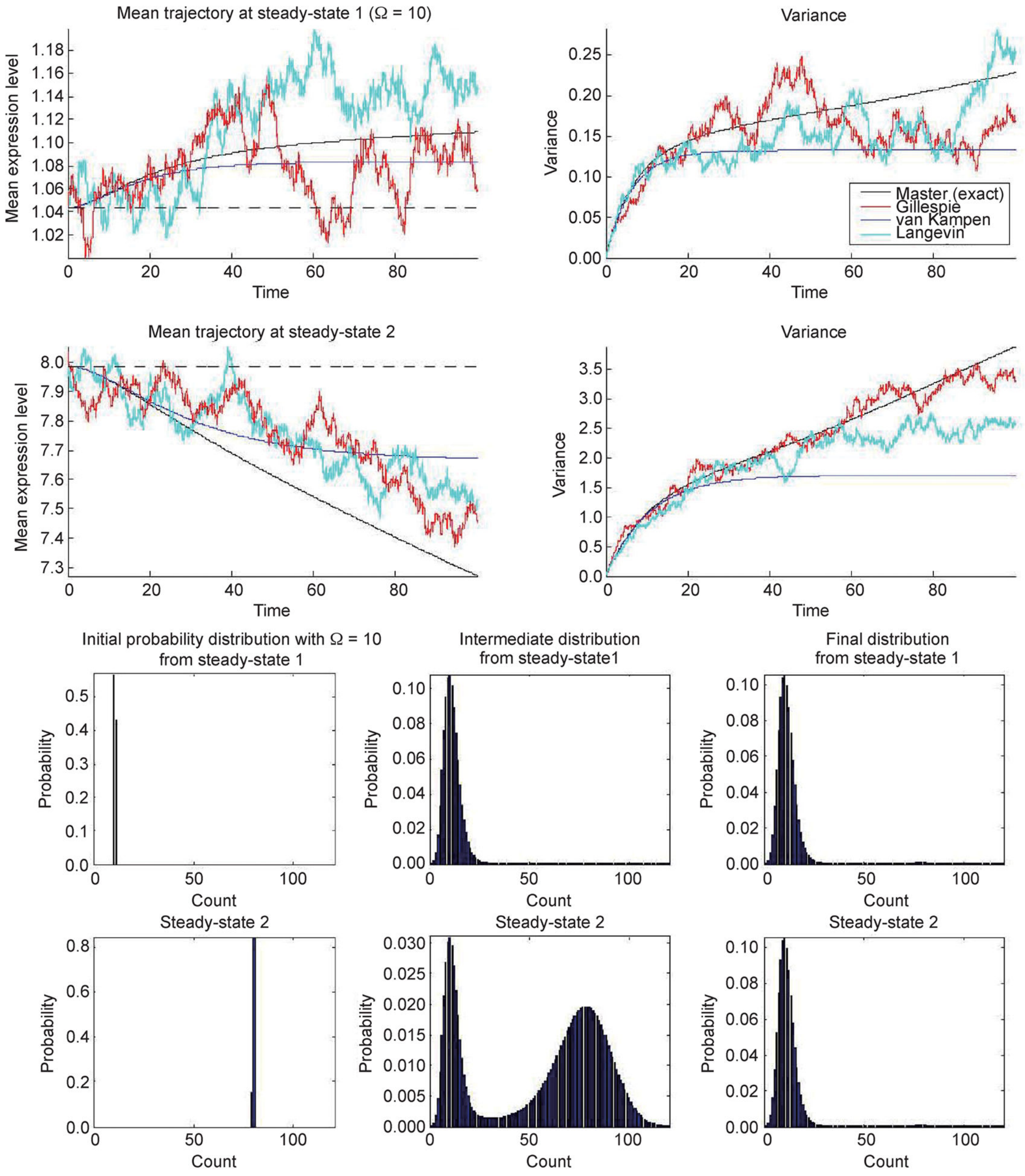

Figure 12. One gene system with two stable steady-states (36), Ω= 10

Just as in Figures 9 and 10, regardless of the initial condition, the probability converges to a bimodal distribution with a strong peak at e_1 (π_1 = 97%) and weaker peak at e_2 (π_2 = 3%), and the mean converges to the weighted average $\pi_1 e_i + \pi_2 e_2 \approx 1.24$ predicted in our stability analysis for Ω= 10.

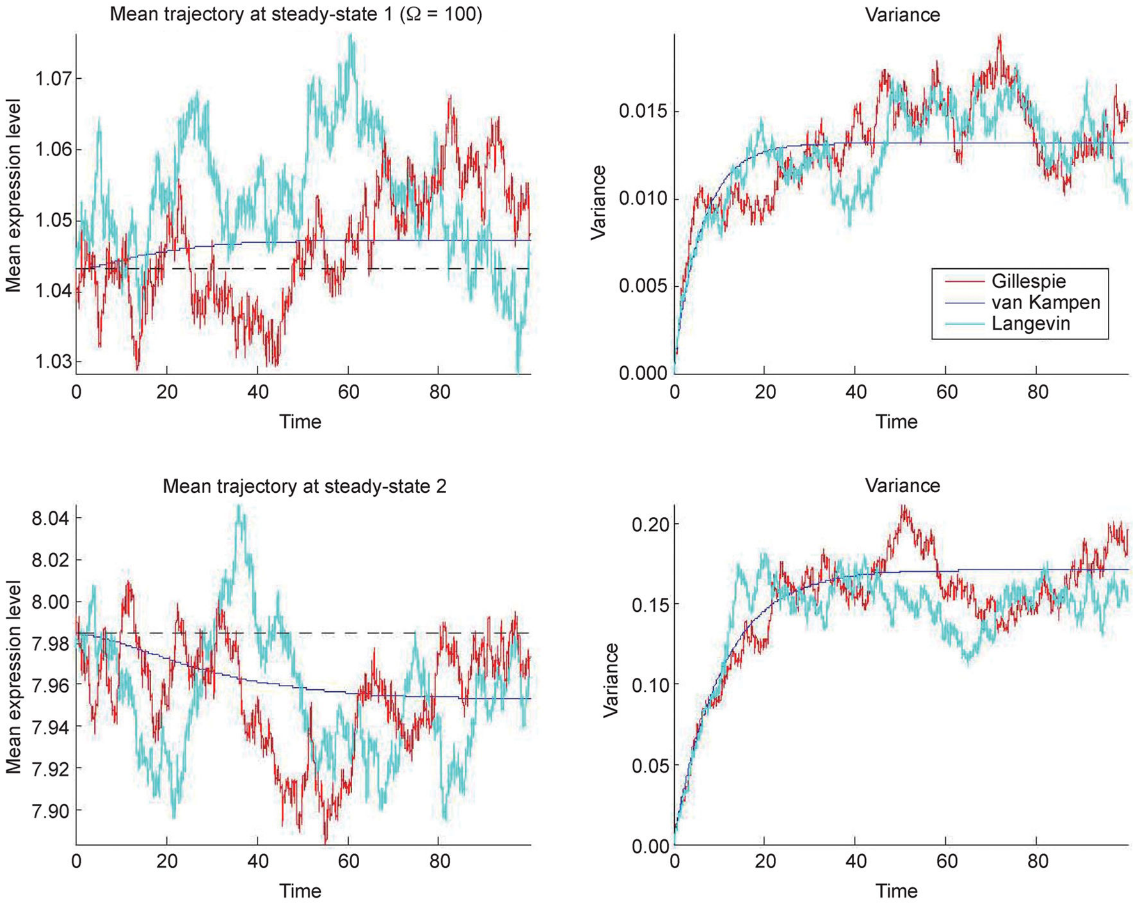

Figure 13. One gene system with two stable steady-states (36), Ω= 100

Mean (left) and variance (right) trajectories via (improperly applied) van Kampen approximation (blue) and average of 100 trajectories of the Gillespie (red) and Langevin simulation (cyan). The exact Master equation calculation suffered from instability (oscillations) so the trajectory is not shown here. Since escape time scales exponentially with the system size, the escape time for this system far exceeds the length of the simulation. Therefore the stochastic trajectories remain close to the deterministic steady-state where they originated for the duration of the simulation. Since the two deterministic steady-states operate mostly independently of each other in the simulation timeframe, the van Kampen approximation agrees quite well, unlike for smaller system sizes. The variance and difference between the mean and deterministic trajectory are both on the order of $O(\Omega^{-1})$.

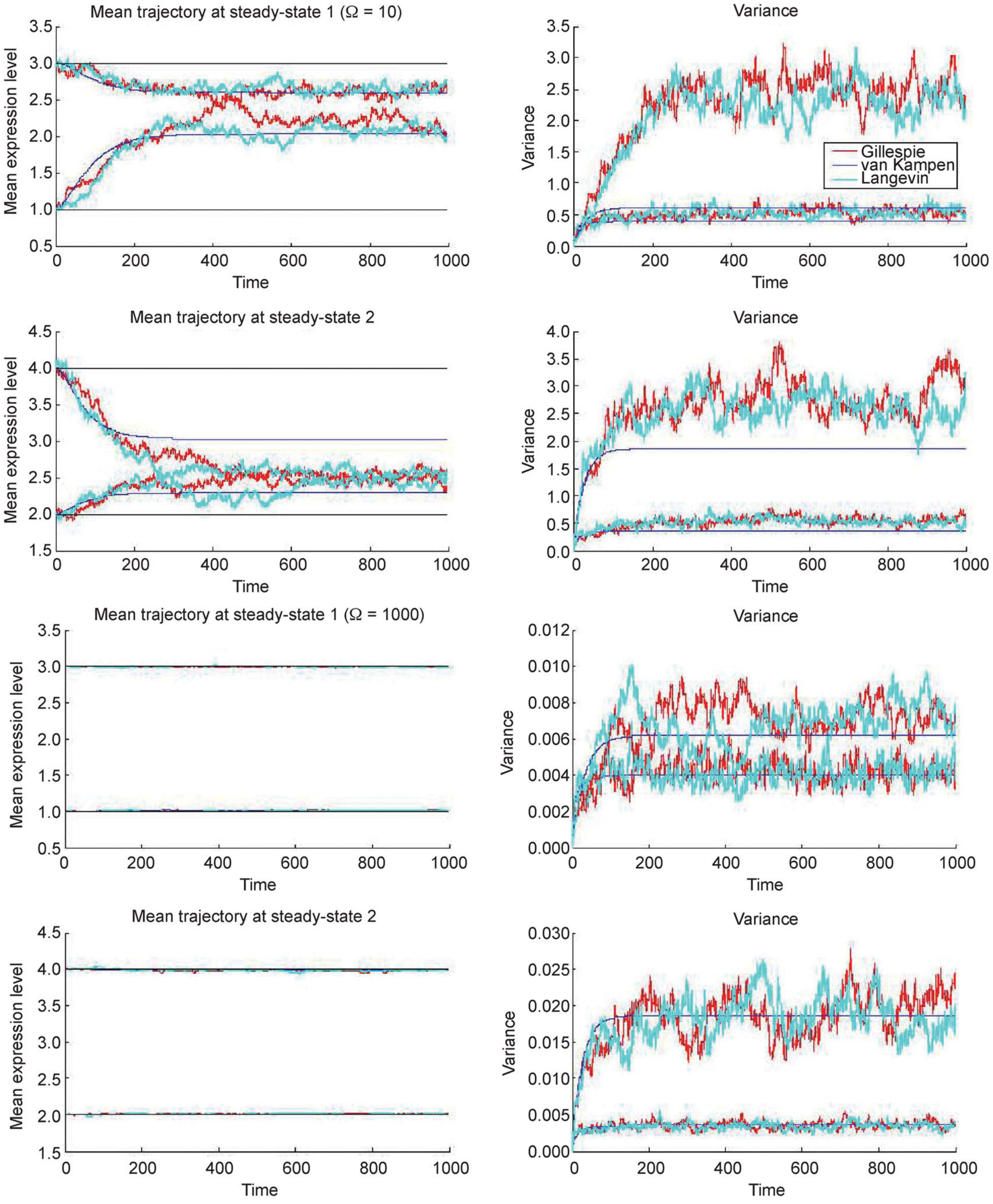

Figure 14. Two gene system with two stable steady-states (37), with Ω= 10 (top) and Ω= 1000 (bottom)

mean and variance via van Kampen approximation (blue), and average over 100 simulations of the Gillespie (red) and Langevin simulation (cyan). For small systems (Ω= 10), the stochastic mean trajectory converges to a weighted average of e_1 and e_2 corresponding to a bimodal stochastic steady-state. Since escape time scales exponentially with system size, the escape time for the large system (Ω= 1000) is very long and the trajectories remain near their initial conditions for the duration of the simulation, hence the van Kampen approximation is quite accurate (though technically not applicable) and the variance and mean-deterministic discrepancy are both $O\left(\Omega^{-1}\right)$.