Chapter 11: Modeling Cellular Variability*

Brian Munsky

CCS-3, B-9, and the Center for NonLinear Studies

Los Alamos National Laboratory

Los Alamos, NM 87545, USA

munsky@lanl.gov

November 16, 2011

1 Introduction

There are a number of reasons why phenotypical diversity may arise despite clonal genetics. Many of these are due to fluctuations in the environment–cells nearer to nutrient sources grow faster; those subjected to heat, light or other inputs will respond accordingly; and so on. But even cells in carefully controlled, homogenous environments can exhibit diversity, and a strong component of this diversity arises from the rare and discrete nature of genes and molecules involved in gene regulation. In particular, many genes have only one or two copies per cell and may be inactive for large portions of the cell cycle. The times at which these genes turn on or off depend upon many random or chaotic events, such as thermal motion, molecular competitions, and upstream fluctuations.

To illustrate how diversity may arise despite identical genetics and initial conditions, Fig. 1 shows a cartoon of a simple gene regulatory system. At the initial time (Fig. 1A), the cell could have a single copy of an important gene, such as *gfp* (green fluorescent protein) and a single copy of an unstable activator molecule, A. The dynamics of the process is a race–will A bind and activate g first (Fig. 1C), or will A degrade before it has a chance to bind (Fig. 1B)? Although real biological systems are far more complex than this toy cartoon, the principles remain the same: genes may be

*This chapter has been submitted for inclusion in *Quantitative Biology From Molecular to Cellular Systems*, Edited by Michael E. Wall at the Los Alamos National Laboratory, and to be published by Taylor and Francis, Inc. Please contact Brian Munsky (brian.munskygmail.com) for an updated preprint and copyright information.

active or inactive simply due to a chance reaction with another molecule. Regulatory molecules may undergo many different reactions (degradation, dimerization, folding, etc...), which may impede or help them to bind to the proper gene regulatory site. Smaller copy numbers typically result in more variable responses, as a single molecule events represent a much higher relative change. In particular, with one gene copy, switches can be all-or-nothing; with more copies, the response can be much more graded.

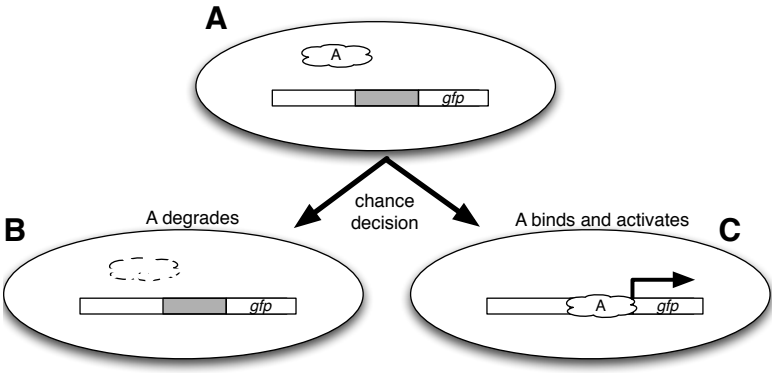

Figure 1: Cartoon depiction of stochastic gene regulation. Panel A: The cell begins with a single gene and a single, unstable activator protein (A). Panel B: Through a chance event, the activator molecule degrades, and the gene remains inactive. Panel C: through a different chance event, the activator may bind and activate the gene making the cell active.

Once rarity and discreteness causes variability to arise in a single gene or regulatory molecule, it can affect the system's downstream elements as the products of one gene activate or repress another [1, 2, 3, 4, 5, 6, 7]. How the system survives and even exploits this variability depends upon the mechanisms of the underlying gene regulatory network. In many cases, variability is undesirable, in which case evolution will have favored mechanisms that diminish the level of variability. For example, negative feedback (especially auto-regulation) mechanisms can reduce variability for a given mean level signal [8, 9, 10], and such auto-regulatory mechanisms are present in about 40% of the transcription factors in *E. coli* [11]. Section 5.1 considers a model of gene transcription and translation in which a protein represses the activation of it's own gene. This auto-regulation enables the system to exhibit less variability for the same mean level of expression. In another context, dynamics in one part of a regulatory network can help to filter out certain fluctuation frequencies coming from other sources via low-pass or band-pass filters [12, 13]. For example, the simple system examined in Section 5.2 acts as a low-pass filter.

In other circumstances, discrete variations may be exploited to realize different desired cellular behaviors. When passed through certain nonlinear processes, such as sequential molecular binding

events or oligomerizaytion, external signals can be amplified [14] or damped as a result of system stochasticity. Section 5.2 illustrates such an example, where the nonlinearity arrises as a result of the binding of gene regulatory factors. Other mechanisms use fluctuations to excite and/or improve the robustness of resonant behaviors [15]. An example of this behavior will be presented in Section 5.4.

One of the most important and obvious of stochastic phenomena is that of stochastic switching, where cells can express two or more very different phenotypes despite identical genotypes and environments [16, 17, 18, 19]. In single cell organisms, the ability to switch at random is an important evolutionary survival trait in competitive and/or uncertain environments. If a species behavior is completely uniform, then a single new mechanism may be sufficient for a competitor to overcome that species (i.e., a single antibiotic would destroy all bacteria). If a species switches too predictably between phenotypes, then competitors can evolve their own switching strategies to outperform that organism in every circumstance (i.e., a given sequence of antibiotics would destroy all bacteria). But if a given cell type switches at random among many unpredictable phenotypes, then such strategies become much more difficult to devise, and perhaps no strategy would suffice (i.e., some bacteria would survive no matter what sequence of antibiotics are applied). Even in the absence of direct competition, the ability to switch at random is also important for survival in an uncertain environment [20]. To illustrate stochastic switching behavior, Section 5.3 provides an example of the analysis of a genetic toggle switch that has been used to sense and record environmental conditions such as UV radiation [21, 22].

1.1 Measurement of single cell variability.

There are a number of well-established experimental techniques with which one can measure the phenotypic and/or molecular variability of single cells [23]. In many of these techniques, cells are prepared so that traits of interest (i.e., gene expression, protein-membrane localization, etc...) are made visible due to the presence or activation of fluorescent markers. For example, fluorescent dyes can be attached to antibodies, which bind to specific cellular proteins or phosphoproteins of interest. Using fluorescence *in situ* hybridization (FISH) techniques, fluorophores can also be attached via oligomers to DNA and RNA molecules [24, 25, 26]. By combining multiple dyes, researchers can simultaneously measure multiple different molecule types, or examine colocalization

or conformational changes in molecules via fluorescence (Förster) resonance energy transfer (FRET) techniques. Alternatively, genes in cell strains may be cloned to include coding for fluorescent proteins such as green, yellow or cyan fluorescent protein (*gfp, yfp,cfp*) instead of, or in addition to, their naturally expressed proteins. Like FRET, introduction of split GFP [27, 28, 29] enables a similar ability to measure the colocalization of important proteins.

In addition to the many means of marking individual biological molecules in a cell, there are also many different ways of measuring these markings within a cell. A natural approach is fluorescence microscopy, with which one can simply look at the cells and directly observe which cells are active and which are not. With confocal microscopy, it is possible to resolve individual fluorescently tagged molecules such as DNA and RNA in fixed cells [26]. For example, with single molecule FISH and automated 3-dimensional image processing software, it is then possible to count mRNA molecules in hundreds to thousands of different cells–thereby obtaining precise distributions in carefully controlled experimental conditions. With time lapse microscopy and non-invasive reporters like fluorescent proteins, it is possible to track individual living cells as they change over time. In either confocal or time-lapse microscopy approach, each cell reveals a huge amount of data in the form of multiple multi-color images, which must be manually or automatically processed.

Another common technique is flow cytometry [30]. In this approach, individual cells pass through an excitation laser (or lasers) and detectors record how much light is reflected, scattered, or absorbed and re-emitted at various wavelengths of light. The high throughput nature of this approach enables researchers to measure millions of cells in a minute. With auto-sampling techniques, researchers can test hundreds of cultures in an hour–each with with different inputs or conditions. More recent technologies are currently under development to combine the strengths of microscopy with the high through-put nature of flow cytometry. In particular, new imaging flow cytometers can capture multiple fluorescent images of each cell as it passes through the excitation lasers–in some cases these images may be used to resolved to count individual fluorescently tagged molecules.

1.2 Using measurements of cellular variability to infer system properties.

Each of these experimental approaches enables thorough quantitative measurements of gene expression. However, the resulting data are vast and often difficult to analyze. In many computational studies, cell-to-cell variability has been viewed as a computational nuisance. Certain cellular behav-

iors can be understood only in the context of intrinsic variability, but including this variability in a computational model results in an explosion of computational complexity. Researchers have made significant progress on developing methods to handle this computational challenge including kinetic Monte Carlo algorithms [31, 32, 33, 34, 35, 36], linear noise approximations [37, 38], moment closure [39, 40] and matching [41] techniques, moment generating functions [42], spectral methods [43] and finite state projection approaches [44, 45, 46, 47]. While no single computational approach applies to all biological systems, the growing arsenal of tools makes it more likely that some approach may suffice for a given system of interest. In Sections 3 and 4, we review a couple of these approaches.

By integrating stochastic modeling approaches with experimental measurements of single-cell variability, it becomes possible to obtain a better understanding of the dynamics of biochemical networks. Such analyses provide a new tool with which to compare and contrast different possibilities for evolutionary design [20]. These analyses of cellular variability may also help to determine what mechanisms are being employed by a particular biological system [48, 49, 50]. For example, different logical structures such as AND or OR gates can be discovered in two component regulatory systems by examining the stationary transmission of the cell variability through the network [48], or correlations of different aspects of cell expression at many time points can reveal different causal relationships between genes within a network [49]. Similarly, measuring and analyzing the statistics of gene regulatory responses in certain conditions can help to identify system parameters and develop quantitative, predictive models for certain systems [51, 22, 52, 53]. Section 6 provides such an example on the identification of a gene regulation model from single-cell flow cytometry data.

1.3 Chapter Focus

The focus of this chapter is to discuss phenomena of cell-to-cell variability in biological systems and illustrate a few computational analyses of these phenomena. In Section 2, we discuss the mesoscopic scale for modeling intracellular processes as discrete state Markov processes; we derive the chemical master equation that describes such processes; and we review a few kinetic Monte Carlo algorithms that are often used to simulate these processes. In Section 3 we describe the Finite State Projection (FSP) approach to solving the chemical master equation, and in Section 4, we include step-by-step examples on using the FSP approach to analyze and identify stochastic models of gene regulation.

Each of these approaches is further illustrated with graphical user interface Matlab software, which is included in the book CD and which can be downloaded from http://cnls.lanl.gov/~munsky or can be requested from the author at munsky@lanl.gov. In Section 5, we illustrate the use of the FSP approach and software on a few examples of stochastic gene regulation, and in Section 6 we use this software and flow cytometry data to identify a model of *lac* regulation in *E. coli*. Finally, in Section 7 we finish with a brief summary on the state of the art in the analysis and identification of single cell variability in gene regulatory systems.

2 Mesoscopic Modeling of Bio-molecular Reactions.

One could attempt to analyze biochemical reaction networks at many different scales. At the microscopic scale, one can use molecular dynamics simulations to explore how individual protein molecules move, fold and interact with surrounding molecules. At the macroscopic scale, large-volume chemical processes are treated with continuous-valued concentrations that evolve according to deterministic ordinary differential equations. However, single-cell data of the types discussed in 1.2 require an intermediate approach, typically referred to as the *mesoscopic* scale. At this scale, each chemical species, $\{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$, is described with an integer population, *i.e.*, the population of $\mathcal{S}_k$ is denoted by the integer $\xi_k \geq 0$. At any point in time the state of the system is then given by the integer population vector $[\xi_1, \ldots, \xi_N]$, and reactions correspond to transitions from one such state to another. Typically, the process is assumed to evolve according to Markovian dynamics, meaning that reaction rates depend only upon the current state of the system and not upon how that state has been reached. In order for this assumption to hold, the system must be well-mixed in some sense as follows:

Gillespie's 1992 paper [54] provides a derivation of Markovian chemical kinetics based upon a literal sense of a well-mixed chemical solution. To understand this argument, consider two spherical molecules, s_1 and s_2 in a volume of Ω. A reaction occurs when the two molecule centers come within a certain distance, r, of one another. During an infinitesimal time period dt, the molecule s_1 moves with an average speed u, covers a distance udt, and sweeps a region $d\Omega \approx \pi r^2 udt$ relative to the center of molecule s_2. Assuming that the system is physically well-mixed, the probability that the two molecules react is $\pi r^2 u \Omega^{-1} dt$. If there were ξ_1 molecules of s_1 and ξ_2 molecules of s_2, then the probability that *any* such reaction will occur is given by $w(\xi_1, \xi_2)dt = \xi_1 \xi_2 \pi r^2 u \Omega^{-1} dt$. In this

formulation, the key result is that the infinitesimal probability of reaction has the form $w(\xi_1, \xi_2)dt$ which depends only upon the population $\{\xi_1, \xi_2\}$ at the current time and not upon the history of these populations. We note that according to this derivation, reaction propensities are restricted to at most second order.

At face value the Markovian description derived by Gillespie does not seem to apply to most biological systems, where molecules are spatially concentrated and non-spherical. However, the "well-mixed" concept need not be so literal–the memoryless nature of Markovian dynamics can also result from the overwhelming complexity of biochemical reactions [55]. Many biochemical reactions, such as transcription, translation, degradation, protein assembly and folding are comprised of numerous sub-steps. Each of these sub-steps adds a new opportunity for the given reaction to reverse, abort, or otherwise fail to complete. As more and more of these sub-reactions occur, the distribution of the corresponding sub-states quickly equilibrates to a quasi-steady distribution. Thus after a short transient period, the system's transition probabilities attain a "well-mixed" quasi-steady equilibrium, which is defined by the current coarse state of the system. Unlike the formulation in [54], this concept of well-mixedness supports far more complicated formulation for the stochastic reaction rates, including Michaelis-Menten, Hill and other more complicated functions.

For the purposes of this article, we assume the most general Markov form for a discrete-value, continuous time chemical process. The reaction rates are given by propensity function $w(\xi_1, \ldots, \xi_N, t)dt$, where w can be any non-linear function of the species populations and the current time. For later convenience, we will refer to specific Markov processes with the notation, $\mathcal{M}$, and we will think of them as random walks on a discrete lattice as shown in as shown in Fig. 2a. In the next subsection, we present the (chemical) Master Equation, which describes the dynamics of the probability distributions for such a process.

2.1 The chemical Master Equation.

We describe a chemical solution of N species, $\{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$ by its state $\mathbf{x} = [\xi_1, \ldots, \xi_N]$. Each μ^{th} reaction is a transition from some state $\mathbf{x}_i$ to some other state $\mathbf{x}_j = \mathbf{x}_i + \nu_\mu$. Here, ν_μ is known as the *stoichiometric vector* and it describes how the μ^{th} reaction changes the system's state. For example, the reaction $s_1 + s_2 \rightarrow s_3$ has the stoichiometric vector $\nu = [-1, -1, 1]^T$. As described above, each reaction has a *propensity function*, $w_\mu(\mathbf{x}, t)dt$, which is the probability that the μ^{th}

reaction will happen in a time step of length dt. This function of the system population vector and the current time is allowed to have any arbitrary form, provided that it does not allow for reactions that lead to negative numbers.

The stoichiometry and propensity functions for each of the M possible reactions fully define the system dynamics and are sufficient to find sample trajectories with the kinetic Monte Carlo methods as discussed in Section 2.2. However, for many interesting gene regulatory problems individual system trajectories are not the best description. Instead, it is desirable to analyze the dynamics in terms of probability distributions. For this it is useful to derive the chemical master equation.

Suppose that one knows the probability of all states $\mathbf{x}_i$ at time t, then the probability that the system will be in the state $\mathbf{x}_i$ at time, $t + dt$, is equal to the sum of (i) the probability that the system begins in the state $\mathbf{x}_i$ at t and remains there until $t + dt$, and (ii) the probability that the system is in a different state at time t and will transition to $\mathbf{x}_i$ in the considered time step, dt. This probability can be written as:

$$p(\mathbf{x}_i; t + dt) = p(\mathbf{x}_i; t) \left(1 - \sum_{\mu=1}^{M} w_\mu(\mathbf{x}, t) dt \right) + \sum_{\mu=1}^{M} p(\mathbf{x}_i - \nu_\mu; t) w_\mu(\mathbf{x}_i - \nu_\mu, t) dt. \tag{1}$$

If one enumerates all possible $\mathbf{x}_i$ and defines the probability distribution vector

$$\mathbf{P}(t) = [p(\mathbf{x}_1; t), p(\mathbf{x}_2; t), \ldots]^T,$$

then it is relatively easy to derive the set of linear ordinary differential equations, known as the chemical master equation (CME) [37]:

$$\dot{\mathbf{P}}(t) = \mathbf{A}(t)\mathbf{P}(t). \tag{2}$$

Although the master equation is linear, its dimension can be extremely large or even infinite, and it is unusually impossible to solve exactly. In many cases, the master equation can only be solved by using kinetic Monte Carlo to simulate numerous trajectories for its dynamics. Such approaches are discussed in the following subsection. In other cases, certain projection approaches make it possible to obtain approximate solutions for the master equation, as is discussed in Section 3.

2.2 Kinetic Monte Carlo methods (Stochastic Simulation Algorithm)

The majority of analyses at the mesoscopic scale have been conducted using kinetic Monte Carlo (MC) algorithms. The most widely used of these algorithms is Gillespie's Stochastic Simulation Algorithm (SSA) [31], which is very easy to apply. Each step of the SSA begins at a state $\mathbf{x}$ and a time t and is comprised of three tasks, (*i*) generate the time until the next reaction, (*ii*) determine which reaction happens at that time, and (*iii*) update the time and state to reflect the previous two choices. For a single reaction with propensity function, $w(\mathbf{x})$, the time of the next reaction, τ, is an exponentially distributed random variable with mean $1/w(\mathbf{x})$. For M different possible reactions with propensities $\{w_\mu(\mathbf{x})\}$, τ is the minimum of M such random variables, or, equivalently an exponentially distributed random variable with mean equal to $\left(\sum_{\mu=1}^{M} w_\mu(\mathbf{x})\right)^{-1}$. To determine which of the M reactions occurs at $t+\tau$, one must generate a second random variable from the set $\mu = \{1, 2, \ldots, M\}$ with the probability distribution given by $P(\mu) = w_\mu(\mathbf{x}) \left(\left(\sum_{\mu=1}^{M} w_\mu(\mathbf{x})\right)^{-1}\right.$. Once τ and μ have been chosen, the system can be updated to $t = t + \tau$ and $\mathbf{x} = \mathbf{x} + \nu_\mu$.

Researchers have proposed many accelerated approximations of the SSA. In the first such approximation, the system is partitioned into slow and fast portions. In [34] the system is separated into slow "primary" and fast "intermediate" species. This method uses three random variables at each step: first, the primary species' populations are held constant, and the population of the intermediate species is generated as a random variable from its quasi-steady-state (QSS) distribution. The dynamics of the "primary" species are then found with two more random variables, similar to the SSA above but with propensity functions depending upon the chosen populations of the intermediates species. The Slow-Scale SSA (ssSSA) [35] is very similar in that the system is again separated into sets of slow and fast species. The ssSSA differs in that it does not explicitly generate a realization for the fast species, but instead uses the QSS distribution to scale the propensities of the slow reactions. Each of these QSS assumption based approaches lead to a reduced process where all states must retain exponential waiting times. In contrast, similar reductions based upon concepts of stochastic path integrals and moment generating functions have yielded coarse-grained realizations that allow for non-exponential waiting times and thereby preserve the statistical characteristics of the original dynamics [42]. So-called hybrid methods such as [56] and [57] also separate the system into fast and slow reactions, but these methods do not then rely upon a QSS approximation. Instead, the fast reactions are approximated with deterministic ODEs or as continuous

valued Markov processes using Langevin equations, and the slow reactions are treated in a manner similar to the SSA except now with time varying propensity functions.

In a second approach to accelerating the SSA, researchers frequently assume that propensity functions are constant over small time intervals. With this "τ leap assumption" one can model each of the M reaction channels as an independent Poisson random process [32]. Beginning at time t and state $\mathbf{x}(t)$, the state at the end of a time step of length τ is approximated as $\mathbf{x}(t+\tau) = \mathbf{x}(t) + \sum_{\mu=1}^{M} k_\mu \nu_\mu$, where each k_μ is a random variable chosen from the Poisson distribution $k_\mu \in \mathcal{P}(w_\mu(\mathbf{x}(t)), \tau)$. The accuracy of τ leaping methods depends only upon how well the τ leap assumption is satisfied. Naturally, the τ leap assumption is best satisfied when all species have sufficiently large populations and all propensities functions are relatively smooth. Otherwise small changes in populations could result in large relative changes in propensities. Ignoring these changes can easily lead to unrealistic predictions of negative populations and/or numerical stiffness. One may avoid negative populations by using a Binomial τ leap strategy [58] or by adaptively choosing the size of each τ leap [59]. One can also ameliorate the problem of numerical stiffness using implicit methods such as that in [60].

When the populations are very large, and the propensity functions are very smooth, the chemical species may be more easily modeled with continuous variables using the *chemical Langevin equation* [61]. In this solution scheme, one assumes that many reactions will occur in the *macroscopic infinitesimal* times step dt without violating the τ leap assumption. One can therefore replace the Poisson distributions with Gaussian distributions, and treat the resulting process as a stochastic differential equation driven by white noise [61].

A single simulation using kinetic Monte Carlo algorithms, such as the SSA and its modifications, describes a possible trajectory of one cell as it changes over time. These trajectories may then be compared directly to experimental data such as time lapse fluorescence microscopy studies, with which it is possible to track the dynamics of single cells. Unfortunately, because these trajectories are random, two identical cells may show very different trajectories, and this comparison can be difficult to make or even misleading. To avoid these problems, it is often useful to collect statistics from many such trajectories and try to make comparisons on the levels of these statistics rather than at the level of a single trajectory. In the next section, we discuss and alternate approach that can directly generate these statistics for certain systems.

3 Analyzing Population Statistics with FSP Approaches.

As discussed above, there are a number of experimental techniques to measure and quantify cell to cell variability. In particular, many of these approaches such as flow cytometry and FISH are capable of taking only images or measurements from a given cell at a single time point in its development. With these approaches, one cannot measure trajectories of a given cell, but it is very easy to establish probability distributions for a population of cells. Thus, to better compare models and data, it is useful to use modeling approaches to generate these distributions. This is equivalent to solving the master equation at certain instances in time. With the KMC approaches described above, this corresponds to running many different simulations and collecting the ensemble statistics. Alternatively, one could attempt to directly solve for the solve for the populations statistics or distributions. In this section, we discuss one such approach, namely the Finite State Projection approach, to solve the master equation.

3.1 Notation for the FSP

In order to describe Finite State Projection approach, we must first introduce some convenient notation in addition to that presented above. As above, the population of the system is comprised of the integer populations of the different species, $\{\xi_1, \ldots, \xi_N\} \in \mathbb{Z}_{\geq 0}$. The states can be enumerated, meaning that each can be assigned a unique index i such that the state $\mathbf{x}_i$ refers to the population vector, $[\xi_1^{(i)}, \ldots, \xi_N^{(i)}]$.

Let $J = \{j_1, j_2, j_3, \ldots\}$ denote a set of indices in the following sense. If $\mathbf{X}$ is an enumerated set of states $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots\}$, then $\mathbf{X}_J$ denotes the subset $\{\mathbf{x}_{j_1}, \mathbf{x}_{j_2}, \mathbf{x}_{j_3}, \ldots\}$. Let J' denote the complement of the set J. Furthermore, let $\mathbf{v}_J$ denote the subvector of $\mathbf{v}$ whose elements are chosen according to J, and let $\mathbf{A}_{IJ}$ denote the submatrix of $\mathbf{A}$ such that the rows have been chosen according to I and the columns have been chosen according to J. For example, if I and J are defined as $\{3, 1, 2\}$ and $\{1, 3\}$, respectively, then:

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}_{IJ} = \begin{bmatrix} g & k \\ a & c \\ d & f \end{bmatrix}.$$

For convenience, we will let $\mathbf{A}_J := \mathbf{A}_{JJ}$. With this notation, we are now ready to state the main result of the Finite State Projection approach [44, 46], which we will present as it was described in

[62].

We define the infinite state Markov process, $\mathcal{M}$, as the random walk on the configuration set $\mathbf{X}$, as shown in Fig. 2a. The master equation for this process is $\dot{\mathbf{P}}(t) = \mathbf{A}(t)\mathbf{P}(t)$, with initial distribution $\mathbf{P}(0)$ as described in Section 2. We can define a new Markov process $\mathcal{M}_J$ such as that in Fig. 2b, comprised of the configurations indexed by J plus a single absorbing state. The master equation of $\mathcal{M}_J$ is given by

$$\begin{bmatrix} \dot{\mathbf{P}}_J^{FSP}(t) \\ \dot{g}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_J & \mathbf{0} \\ -\mathbf{1}^T\mathbf{A}_J & 0 \end{bmatrix} \begin{bmatrix} \mathbf{P}_J^{FSP}(t) \\ g(t) \end{bmatrix}, \tag{3}$$

with initial distribution,

$$\begin{bmatrix} \mathbf{P}_J^{FSP}(0) \\ g(0) \end{bmatrix} = \begin{bmatrix} \mathbf{P}_J(0) \\ 1 - \sum \mathbf{P}_J(0) \end{bmatrix}.$$

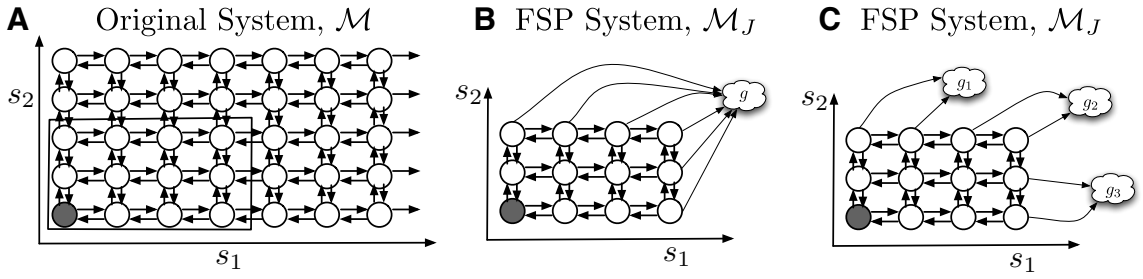

Figure 2: (a): A Markov chain for a two species chemically reacting system, $\mathcal{M}$. The process begins in the configuration shaded in grey and undergoes four reactions to increase/decrease the two different species populations. The dimension of the Master equation is equal to the total number of configurations in $\mathcal{M}$, and is too large to solve exactly. (b) In the FSP algorithm a configuration subset $\mathbf{X}_J$ is chosen, and all remaining configurations are projected to a single absorbing point g. This results in a small dimensional Markov process, $\mathcal{M}_J$. (c) By using multiple absorbing sites, one can keep track of how the probability measure leaves the projection space [62].

3.2 FSP Theorems and Results

The finite state process $\mathcal{M}_J$ has a clear relationship to the original $\mathcal{M}$. First, the scalar $g(t)$ is the *exact probability* that the system has been in the set $\mathbf{X}_{J'}$ at *any* time $\tau \in [0, t]$. Second, the vector $\mathbf{P}_J^{FSP}(t)$ are the *exact joint probabilities* that the system (i) is in the corresponding states $\mathbf{X}_J$ at time t, and (ii) the system has remained in the set $\mathbf{X}_J$ *for all* $\tau \in [0, t]$. Note that $\mathbf{P}_J^{FSP}(t)$ also provides a finite dimensional approximation of the solution to the CME as follows:

First, it is guaranteed that $\mathbf{P}_J(t) \geq \mathbf{P}_J^{FSP}(t) \geq \mathbf{0}$ for any index set J and any initial distribution $\mathbf{P}(0)$. This is a concequence of $\mathbf{P}_J^{FSP}(t)$ being a more restrictive joint distribution than $\mathbf{P}_J(t)$.

Second, the actual 1-norm distance between $\mathbf{P}(t)$ and $\mathbf{P}^{FSP}(t)$ is easily computed as

$$\begin{aligned}\left|\left[\begin{array}{c}\mathbf{P}_J(t_f)\\ \mathbf{P}_{J'}(t_f)\end{array}\right]-\left[\begin{array}{c}\mathbf{P}_J^{FSP}(t_f)\\ \mathbf{0}\end{array}\right]\right|_1 &= \left|\mathbf{P}_J(t_f)-\mathbf{P}_J^{FSP}(t_f)\right|_1+\left|\mathbf{P}_{J'}(t_f)\right|_1,\\ &= \left|\mathbf{P}_J(t_f)\right|_1-\left|\mathbf{P}_J^{FSP}(t_f)\right|_1+\left|\mathbf{P}_{J'}(t_f)\right|_1,\\ &= 1-\left|\mathbf{P}_J^{FSP}(t_f)\right|_1,\\ &= g(t). \qquad (4)\end{aligned}$$

3.3 The FSP Algorithm.

The formulation above suggests an FSP algorithm [44], which examines a sequence of finite projections of the ME. For each projection set, one can obtain an accuracy guarantee using Eqn. (4). If this accuracy is insufficient, more configurations can be added to the projection set, thereby monotonically improving the accuracy. The full algorithm can be stated as follows:

The Original Finite State Projection Algorithm

Inputs Propensity functions and stoichiometry for all reactions.
Initial probability density vector, $\mathbf{P}(0)$.
Final time of interest, t_f.
Total amount of acceptable error, $\varepsilon > 0$.

Step 0 Choose an initial finite set of states, $\mathbf{X}_{J_o}$, for the FSP.
Initialize a counter, $i = 0$.

Step 1 Use propensity functions and stoichiometry to form $\mathbf{A}_{J_i}$.
Compute $g(t_f)$ by solving Eqn. 3

Step 2 If $g(t_f) \leq \varepsilon$, **Stop**.
$\mathbf{P}^{FSP}(t)$ approximates $\mathbf{P}(t_f)$ to within a total error of ε.

Step 3 Add more states to find $\mathbf{X}_{J_{i+1}}$.

Increment i and return to **Step 1**.

In this FSP algorithm, there are many way to choose and expand the projections space in Steps 0 and 3, respectively. In the following subsections, we will present a couple such approach, although others may be equally good.

3.3.1 Choosing the initial projection space.

A number of different approaches have been proposed for choosing the initial guess for the projection space. In previous work [44], the initial projection set $\mathbf{X}_{J_0}$ was an arbitrarily chosen set of configurations reachable from the initial condition. The most obvious choice is for $\mathbf{X}_{J_0}$ to contain only the initial configuration: $\mathbf{X}_{J_0} = \{\mathbf{x}(0)\}$. The problem with this approach is that the initial projection space is likely to be far too small. In [63] we proposed initializing $\mathbf{X}_{J_0}$ with a set of states determined by running a few trial SSA trajectories. If we use more SSA runs, $\mathbf{X}_{J_0}$ will likely be larger and therefore retain a larger measure of the probability distribution in the specified time interval. As one uses more SSA runs in the initialization portion of Step 0, fewer iterations of the FSP algorithm are necessary, but there is an added computation cost for running and recording the results of the SSA runs. In this study and in the codes provided, we utilize a mixture of the two approaches.

First, we define a projection space that is defined by a set of nonlinear inequalities:

$$\mathbf{X}_J = \{\mathbf{x}_i\}, \text{ such that } \{f_k(\mathbf{x_i}) \leq b_k\} \text{ for all constraints } k = \{1, 2, \ldots, K\}, \tag{5}$$

where the functions $\{f_k(\mathbf{x})\}$ are fixed functions of the populations and where the bounds $\{b_k\}$ are changed in order to expand or contract the projection space. For example, in the two species $\{\xi_1, \xi_2\}$ systems below, we will use the projection shape functions:

$$\begin{aligned}
&f_1 = -\xi_1,\ f_2 = -\xi_2,\ f_3 = \xi_1,\ f_4 = \xi_2,\\
&f_5 = \max(0, \xi_1 - 4)\max(0, \xi_2 - 4),\\
&f_6 = \max(0, \xi_1 - 4)^2\max(0, \xi_2 - 4),\\
&f_7 = \max(0, \xi_1 - 4)\max(0, \xi_2 - 4)^2.
\end{aligned}$$

We note that with $b_1 = b_2 = 0$, the first of these two constraints specify that the both species must have non-negative populations. The third and fourth constraints specify the max populations of each species, and the remaining constraints specify additional upper bounds on various products of the population numbers. For all constraints, it is important that increases in the values $\{b_k\}$ correspond to relaxations of the associated constraints and increases in the projections space. In practice, these constraints functions are easily changed–the best choice of constraints remains an open problem that will differ from one system to another. Next, we run a single SSA simulation and record all of the states (ξ_1, ξ_2) that are visited in that simulation. Finally, we increase the boundary values $\{b_k\}$ until the inequalities in (5) are satisfied. Thus, we arrive at an initial guess for the projection space, the next step is to expand that projection space until the FSP error meets the specified tolerance.

3.3.2 Updating the projection space.

In Step 3 of the FSP algorithm it is necessary to expand the projection space. In [44] the space was expanded to include all of the states that are reachable in one reaction from the current set. Because not all reactions are equal, this is a very inefficient approach to expanding the projection space–it can lead to expanding too far in one direction or too little in another. Here we tailor an approach similar to that in [63] in order to match our definition of the projection space given in Eq. 5. For this, we choose K absorbing points $\{g_1, \ldots, g_K\}$ where each $g_k(t)$ corresponds to the probability that the system has left the set $\mathbf{X}_J$ in such a way as to violate the k^{th} boundary condition. To do this, we simply split the index set, J', into K different subsets, $\{J'_1, \ldots, J'_K\}$ where J'_k is the set of states that satisfy the first $(k-1)$ boundary constraints, but not the k^{th} boundary constraint:

$$J'_k = \{i\} \text{ such that } \{f_1(\mathbf{x}_i) \leq b_1, \ldots, f_{k-1}(\mathbf{x}_i) \leq b_{k-1}, f_k(\mathbf{x}_i) > b_k\}.$$

With these index sets, we arrive at a new projection for the master equation:

$$\frac{d}{dt}\begin{bmatrix} \mathbf{P}_J^{FSP}(t) \\ g_1(t) \\ \vdots \\ g_K(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_J(t) & \mathbf{0} \\ -\sum \mathbf{A}_{J'_1 J}(t) & \mathbf{0} \\ \vdots & \mathbf{0} \\ -\sum \mathbf{A}_{J'_K J}(t) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{P}_J^{FSP}(t) \\ g_1(t) \\ \vdots \\ g_K(t) \end{bmatrix}. \tag{6}$$

The solution of (6) at a time t_f yields all of the same information as above. In particular, the sum of the vector $\mathbf{g}(t)$ provides the exact distance between the FSP approximation and the true solution, as was observed in Eqn. (4). In addition, each element, $g_k(t)$, is the probability that the k^{th} boundary condition was violated and that this violation occurred before the $\{1, 2, \ldots, k-1\}^{th}$ boundary conditions were violated. This knowledge is easily incorporated into Step 3 of the FSP algorithm. If the k^{th} one boundary condition is violated with high probability, we expand $\mathbf{X}_J$ by increasing b_k to relax that boundary condition.

3.4 Advancements to the FSP Approach.

Linear systems theory provides many tools with which the order of the chemical master equation may be reduced and the efficiency of the FSP may be improved. In most of these reductions, one seeks to approximate the vector $\mathbf{P}(t) \in \mathbb{R}^n$ (here, n may or may not be finite) as some linear transformation of a lower dimensional vector, $\mathbf{\Phi P}(t) = \mathbf{q}(t) \in \mathbb{R}^{m \leq n}$. For example, the original FSP itself is one such projection in which the elements of $\mathbf{q}(t)$ correspond to $\mathbf{P}_J(t)$. There are many other possible projection choices, each of which takes advantage of a different common trait of discrete state Markov processes.

In [64] and [62], we use control theory concepts of controllability and observability to obtain a minimal basis set for the space in which the solution to the FSP evolves. This approach takes into account that not all points in the full probability distribution are necessary, and one may only be interested in solving for a coarser level behavior, such as population means, variances, or extreme value events. In this case, the vector $\mathbf{q}$ corresponds to the observable and controllable states of the master equation, and can have a very low dimension [62].

Alternatively, in [65, 66], one may use time scale separations to project the system onto a spaces defined by the system's fast or slow dynamics. In this case, the projection operator, $\mathbf{\Phi}$ is spanned by the appropriately chosen sets of eigenvectors, and $\mathbf{q}(t)$ refers to the dynamics in that space. For long times, $\mathbf{\Phi}$ consists of the eigenvectors corresponding to the slow eigenvalues. Conversely, for short times, $\mathbf{\Phi}$ should consists of the eigenvectors that correspond to the fast eigenvalues. This approach is similar to the ssSSA [35] discussed above, in that the existence of "fast" and "slow" species do indeed result in a separation of time scales in the ME. However, time-scale based reductions to the master equation are more general in that they may be possible even in the absence of clear

separations of fast and slow species.

For a third projection-type reduction, one can assume that the probability distribution varies smoothly over some portions of the configuration space, solve the FSP problem on a coarse grid, and then interpolate to find the distributions at intervening points. This approach has shown great promise for certain problems [46], particularly for systems with high populations, where the full FSP may have an exorbitant dimension size. Furthermore, it is relatively easy to formulate an algorithm to systematically refine the interpolation grid to attain more precise solutions, which are better tailored to a given system.

For some systems, probability distributions may drift over large portions of the state space, yet remain relatively tight in that they are sparsely supported during any given instant in time [63, 45]. By splitting the full time interval into many small subintervals, one can reduce computational effort by considering much smaller portions of the state space during each time increment. For further improvements in efficiency, many of these these multiple time interval solutions of the FSP can readily be combined with the projection based reductions discussed above.

4 Description of the FSP two-species software.

Before studying any specific biological problem, it is useful to introduce the FSP Toolkit through a simple tutorial example.[1]

4.1 System initialization

The first task in analyzing any system is to specify the system mechanics, parameters, and initial conditions. For this first example, let us consider a process of gene transcription and translation [3], where the reaction mechanisms are defined:

$$R_1: \quad \emptyset \rightarrow R, \qquad R_2: \quad R \rightarrow \emptyset,$$

$$R_3: \quad R \rightarrow R + P, \qquad R_4: \quad P \rightarrow \emptyset.$$

[1] All examples shown in this tutorial, can be accessed by typing "FSP_ToolKit_Main" in the Matlab command window and then clicking the appropriate button in the resulting graphical user interface.

The propensity functions of these reactions are

$$w_1 = k_R, \qquad w_2 = \gamma_R x,$$
$$w_3 = k_P x, \qquad w_4 = \gamma_P y,$$

where the rates are $\{k_R = 5, \gamma_R = 1, k_P = 5, \gamma_P = 1\}$ and the initial condition is given as five molecules of mRNA ($x(0) = 5$) and two protein molecules ($y(0) = 2$). For this example, we have also chosen a final time of 10 time units. For the Finite State Projection approach, it is necessary to specify the maximum allowable 1-norm error in the solution of the master equation. For all examples presented in Sections 4 and 5, we have set a strict accuracy requirement of $\varepsilon = 10^{-6}$. For the FSP Toolkit, these mechanisms and parameters can be entered and/or changed directly in the graphical interface, or they can be pre-defined in a user-specified file.

4.2 Generating stochastic trajectories

Once the system is specified, one can solve it with many different approaches. Perhaps the simplest such approach is to run a stochastic simulation to find a sample trajectory for the process. In the FSP Toolkit, this can be accomplished simply by pressing the button "Run SSA". Fig. 3 illustrates three different aspects of sample trajectories obtained from this approach. Figs. 3A and 3B show the populations of x and y, respectively, as functions of time, and Fig. 3C shows the trajectories in the x-y plane. Because the process is stochastic, each trajectory achieved in this manner will be different. Although a few SSA runs do not provide a solution to the master equation, they do provide a good sense of the system's dynamics. Furthermore, the SSA runs help to choose an initial projections space for use in the FSP solution.

4.3 Solving the Master Equation

In the FSP Toolkit, all aspects of the FSP solution can be acquired simply by pressing the button marked "FSP–Solve It" on the graphical interface. for the casual user, this is sufficient to specify and solve the master equation for many problems. However, for the advanced user, it is import to understand how this solution is obtained. This process is described as follows, beginning with the definition of the master equation.

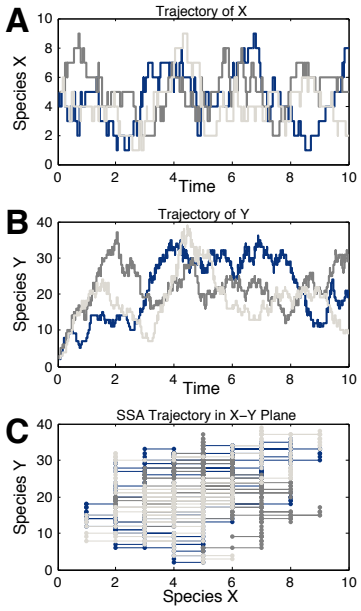

Figure 3: Three stochastic trajectories for the example system in Section 4. Panel A) Trajectories for species X versus time. Panel B) Trajectories for species Y versus. Panel C) Trajectories on the X-Y plane.

4.3.1 Defining the Full Master Equation

With the definition of the mechanisms and parameters, one can also define the master equation, $\dot{\mathbf{P}}(t) = \mathbf{A}(t)\mathbf{P}(t)$. For this, the infinitesimal generator matrix, $\mathbf{A} = \{A_{i,j}\}$ is defined as

$$\mathbf{A}_{ij} = \left\{ \begin{array}{ll} -\sum_{\mu=1}^{M} w_\mu(\mathbf{x}_i) & \text{for } (i = j) \\ w_\mu(\mathbf{x}_j) & \text{for all } j \text{ such that } (\mathbf{x}_i = \mathbf{x}_j + \nu_\mu) \\ 0 & \text{Otherwise} \end{array} \right\}. \tag{7}$$

For the initial distribution, we use the distribution $\mathbf{P}(0) = \{P_i(0)\}$ given as:

$$P_i(0) = \left\{ \begin{array}{l} 1, \text{ if } \mathbf{x}_i = [5, 2] \\ 0, \text{ otherwise} \end{array} \right\},$$

which corresponds to a specific initial condition of five mRNAs and 2 proteins.

4.3.2 Defining the Projected Master Equation

The finite state projection space is governed by the boundary shape functions, $\{f_k\}$, as defined in Eqn. 5. To initialize the constraints, $\{b_k\}$, we use the previous SSA runs as follows. If $\mathbf{X}_{SSA}$ refers

to the set of all states, $\{\mathbf{x}\}$, that were visited during the SSA run(s), then the initial value for each b_k is set to:

$$b_k = \max_{\mathbf{x} \in \mathbf{X}_{SSA}} f_k(\mathbf{x}).$$

In turn, the index sets for the projection are defined by the functions $\{f_k\}$ and the constraints $\{b_k\}$ as follows:

$$\begin{aligned}
J &= \{i\} \text{ such that } \{f_1(\mathbf{x}_i) \leq b_1, \ldots, f_k(\mathbf{x}_i) \leq b_k\}, \\
J'_1 &= \{i\} \text{ such that } \{f_1(\mathbf{x}_i) > b_1\}. \\
J'_2 &= \{i\} \text{ such that } \{f_1(\mathbf{x}_i) \leq b_1, f_2(\mathbf{x}_i) > b_2\}, \\
\vdots &= \vdots \\
J'_K &= \{i\} \text{ such that } \{f_1(\mathbf{x}_i) \leq b_1, \ldots, f_{K-1}(\mathbf{x}_i) \leq b_{k-1}, f_k(\mathbf{x}_i) > b_k\}.
\end{aligned}$$

With these index sets we can define the projections matrix $\mathbf{A}_J$ and the row vectors $\left\{\sum \mathbf{A}_{J'_k J}(t)\right\}$ for use in Eqn. 6.

4.3.3 Solving the Projected Master Equation

Once defined, Eqn. 6 can be solved in a number of different ways depending upon the system. For systems with time varying reaction rates (see Section 5.2), a more general stiff ODE solver is necessary. For systems where the matrix $\mathbf{A}$ is constant, the solution can be found with Krylov subspace methods included in Roger Sidje's expokit ([67]–http://www.expokit.org). With either solution scheme, the solution of (6) is solved incrementally in time from initial time t_0 to time $t = min(t_f, t_v)$, where t_v is the time at which the FSP error tolerance is first observed to be violated. This definition of t_v is necessary only for efficiency reasons–at time t_v the algorithm already knows that the current projection is insufficient and it knows which of the K boundary conditions have been violated. By exiting early from the ODE solver, the solution need not be computed over the interval (t_v, t_f), and there is a significant computational savings.

4.3.4 Updating the Projection

Upon solving Eqn. 6, there are two possibilities: Either the truncation error was small enough ($\sum_{k=1}^{K} g_k(t_f) \leq \varepsilon$), and the solution is acceptable, or the projection must be expanded to encompass more states. In the latter case, the values of $\{g_k(t_v)\}$ are used to increase the boundary constraint constants. For the examples shown here, we use the simple expansion rule:

$$\text{If } g_k(t_v) \geq \varepsilon/K \text{ then } b_k + 0.05|b_k| \rightarrow b_k.$$

Once the boundary constants have been updated in this manner, the next projection can be defined, and the FSP algorithm may continue.

4.3.5 Analyzing FSP Solutions.

Once the FSP error tolerance has been met, there are a number of ways to represent and understand the acquired solution. Fig. 4 shows a couple of these representations that are automatically plotted using the "FSP Toolkit." Fig. 4A shows a contour plot of the joint probability distribution for the populations of mRNA and protein molecules and Figs. 4C,D show marginal distributions for each of these species separately. Fig. 4B shows a plot of the projection space that was found during the FSP algorithm, where the region in white is included in the projection, while the rest in black is excluded.

All of the results in Fig. 4 correspond to the solution at the final time of $t_f = 10$ time units. however, once a projection is found to be sufficient for the final time, t_f, that projection will also be satisfactory for all times between t_0 and t_f. With this in mind, we can compute the dynamics of the means for each of the species as functions of time. For example, the trajectory for species Y is plotted in Fig. 6A below. To generate these plots, one can simply press the button "Show FSP Dynamics" in "FSP Toolkit." In addition to showing trajectories of the means and and standard deviation, this will also create a movie of the joint distribution as a function of time.

5 Examples of Stochastic Analyses

In this section, we will utilize the stochastic analysis tools described above to illustrate some important stochastic phenomena in biological systems. All of these examples can be implemented in

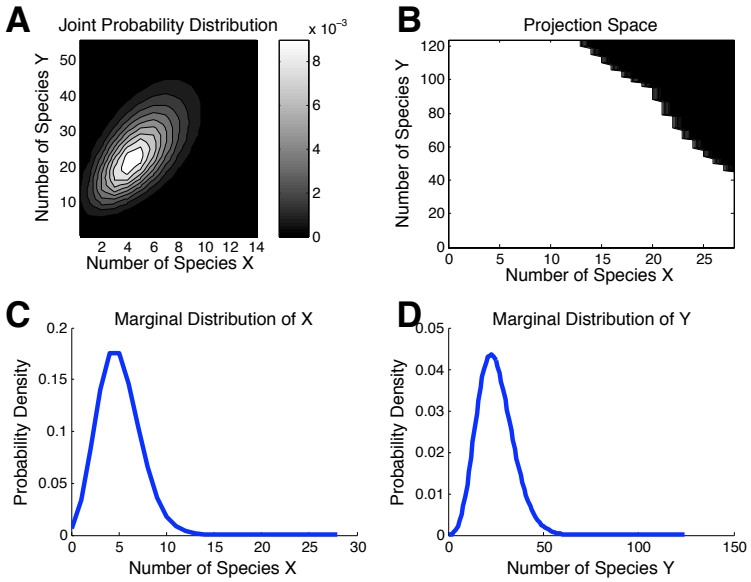

Figure 4: The FSP solution of the master equation at the the final time. Panel A) The joint probability distribution of species X and Y. Panel C,D) the marginal distributions of Species X and Y, respectively. Panel D) The automatically chosen projection space that satisfies a stopping criteria of $\sum(g_{t_f}) \leq = 10^{-6}$.

Matlab using the codes provided, and with minimal user input. The reader is strongly encouraged to work through each example using this software.

5.1 Example 1: Using autoregulation to reduce variability in gene expression

In order to illustrate the importance of feedback, the fist example considers the auto-regulation of a single gene whose protein inhibits its own transcription. The simplified model is comprised of four simple reactions:

$$R_1: \quad \emptyset \rightarrow R, \qquad\qquad R_2: \quad R \rightarrow \emptyset,$$

$$R_3: \quad R \rightarrow R + P, \qquad\qquad R_4: \quad P \rightarrow \emptyset.$$

The propensity functions of these reactions are

$$w_1 = k_R/(1 + k_f y), \qquad\qquad w_2 = \gamma_R x,$$

$$w_3 = k_P x, \qquad\qquad w_4 = \gamma_P y,$$

where the term k_f denotes the strength of the negative feedback. For the nominal model without feedback, we have chosen a parameter set of $\{k_R = 5, \gamma_R = 1, k_P = 5, \gamma_P = 1, k_f = 0\}$, in non-

dimensional time units. For the feedback model, we have set the feedback term to unity, $k_f = 1$, and adjusted the basal transcription rate to $k_R = 120$ in order to maintain the same mean levels of 5 mRNA transcripts and 25 proteins at the final time. For each parameter set, the distribution of mRNAs and proteins after $t_f = 10$ time units is plotted in Fig. 5, where the solid lines correspond to the nominal system and the dashed lines correspond to the system with feedback.

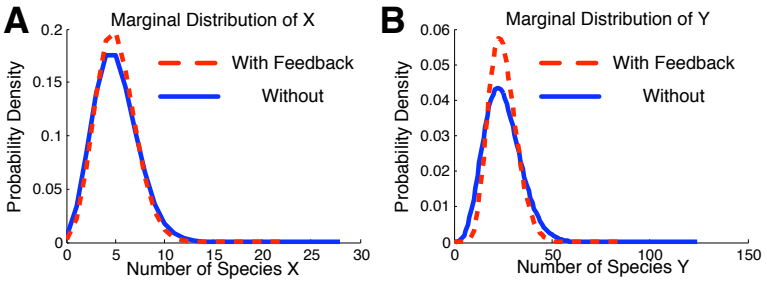

Figure 5: The effect of feedback in gene regulation. Panel A) The marginal distributions of Species X. Panel B) The marginal distributions of Species Y. Panel C) Trajectory for the mean level of species Y in the absence of feedback. Panel D) Trajectory for the mean level of species Y with auto-regulatory feedback.

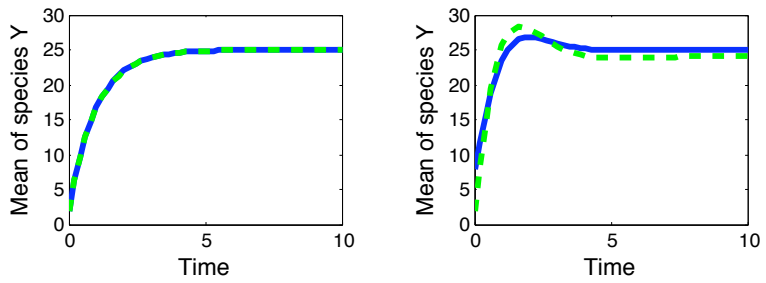

Figure 6: The effect of feedback in gene regulation. Panel A) Trajectory for the mean level of species Y in the absence of feedback. Panel B) Trajectory for the mean level of species Y with auto-regulatory feedback. The dashed lines correspond to the solution to the deterministic ODE model, and the solid lines correspond to the mean of the stochastic model.

By adding negative feedback, the variance in the mRNA levels is reduced by about 20%, while the variance in the protein levels is reduced by about 40% (see Fig. 5). Fig. 6 shows the trajectories for the mean level of proteins for the two systemw with and without feedback. Although both systems eventually achieve the same mean level of protein, the actual dynamics are slightly different. In addition to having less variance, the auto-regulated system has a faster response time. However, the increased speed and lower variability in the feedback mechanism comes at a cost of protein overproduction, which may be costly in terms of cellular resources.

5.2 Example 2: Using nonlinearities and stochasticity to amplify/damp external signals.

For systems with *linear* reaction rates, stochastic models will exhibit the exact same mean level behavior as the corresponding deterministic description. This connection rests on the fact that the expected value of a linear function is equal to the same linear function applied to the mean of its argument. In other words, $\mathbb{E}\{f(\tilde{x})\} = f(\{\mathbb{E}\{\tilde{x}\})$, for any distribution of $\tilde{x}$, when $f(.)$ is linear. For nonlinear functions, this equivalence typically does not hold and non-linearities can result in big differences between the stochastic and deterministic representations of the system. Of particular interest is when $f(x)$ has a significant curvature over the support of the variable x. According to Jensen's inequality, if $f(.)$ is convex over the support of x, then $\mathbb{E}\{f(x)\} \geq f(\{\mathbb{E}\{x\})$. In this case, nonlinearities and stochasticities combine to amplify the signal. If the function is concave, then the inequality is reversed and the signal is effectively damped by the nonlinear and stochastic effects. In this example, we show how this nonlinearity can help to amplify or damp a system's response to an external signal.

In this example, we consider a two species system comprised of four simple reactions:

$$R_1: \quad \emptyset \to X, \qquad\qquad R_2: \quad X \to \emptyset,$$
$$R_3: \quad \emptyset \to Y, \qquad\qquad R_4: \quad Y \to \emptyset.$$

In this system, we assume that the production of the first species, X, is modulated by an external signal according to $w_1 = k_x(1 + \mathrm{e}^{-i(\Omega t - \pi/2)})$, and the degradation of X is given by the standard $w_2 = \gamma_x x$. We consider three subcases for the production of species Y:

(a) species X *activates* the production of species Y according to the *linear* propensity function $w_3^{-} = k_y x/2$;

(b) species X *activates* production of species Y according to the *concave* propensity function $w_3^{\cap} = k_y x/(1+x)$;

(c) species X *represses* the production of species Y according to the *convex* propensity function $w_3^{\cup} = k_y/(1+x)$.

In all three sub-cases, the degradation of species Y is given by $w_4 = \gamma_y y$. The parameters are

kept the same in all three cases: $\{k_x = 10\text{s}^{-1},\ \gamma_x = 10\text{s}^{-1},\ k_y = 15\text{s}^{-1},\ \gamma_y = 1\text{s}^{-1}, \Omega = 1\text{s}^{-1}\}$.

Using the FSP toolKit, one can solve for and generate movies of the distributions of X and Y as functions of time. The solid lines in Fig. 7 illustrate the dynamics of the mean of X for each of the different sub-cases where the propensity for the production is linear, convex or concave. For comparison, the dashed lines show the corresponding solution to the deterministic ordinary differential equation:

$$\frac{dx}{dt} = w_1(x,t) - w_2(x); \qquad \frac{dy}{dt} = w_3(x) - w_4(y).$$

As expected, the mean level of the system with the linear reaction rates is exactly the same as the solution of the deterministic ODEs (see Fig. 7A). For the concave activation of Y, we see that the nonlinear stochastic effects dampen the response to the external signal (see Fig. 7B). Finally, for the convex repression of Y, we see that the nonlinear stochastic effects amplify the response to the external signal (see Fig. 7C).

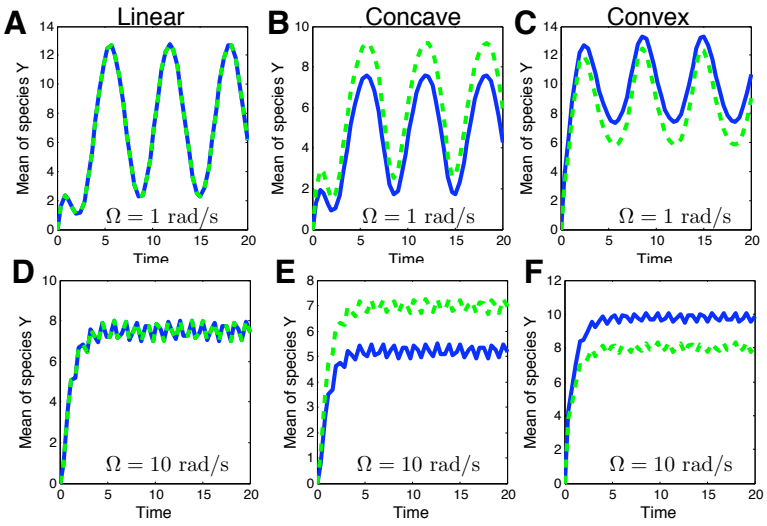

Figure 7: The effects of nonlinearities and stochasticity on signal transduction. Panel A) Trajectory for the mean level of species Y with when species X activates Y through linear regulation. Panel B) Trajectory for the mean level of species Y when species X activates Y with a concave function. Panel C) Trajectory for the mean level of species Y when species X represses Y with a convex function. Panels A-C correspond to a system where the external signal varies with a frequency of 1 rad/s. Panels D-F correspond to a system where the external signal varies with a frequency of 10 rad/s. In all plots, the dashed lines correspond to the solution to the deterministic ODE model, and the solid lines correspond to the man of the stochastic model.

As a side note, the system considered here also acts as a low-pass filter of the external signal. At a frequency of 1 rad/s, the external signal is easily passed through the system. However, if the external fluctuations are much higher in frequency, say 10 rad/s, then the fluctuations become much smaller (see Fig. 7D-F).

5.3 Example 3: Stochastic Toggle Switch.

One of the most obvious of stochastic phenomena in biological systems is that of stochastic switching. To illustrate this phenomenon and how it could be analyzed, we consider the toggle switch composed of genes *lacI* and λcI, which inhibit each other. This system was experimentally constructed in [68] and later used as a sensor of UV light in the environment [21]. In the switch the proteins λcI and LacI inhibit each other as shown in Fig. 8. The switch works as a sensor because the degradation rate of λcI is sensitive to various factors in the external environment, and the system is tuned so that its phenotype is sensitive to changes in this degradation rate. With low degradation rates, λcI will out-compete LacI–in high λcI degradation conditions, LacI will win the competition. In [21], a GFP reporter has been used to quantify the expression level of LacI. While many models are capable of describing this and other toggle switches (see for example [21, 69, 19, 70]), we consider a relatively simple model from [22]. This model is described as follows.

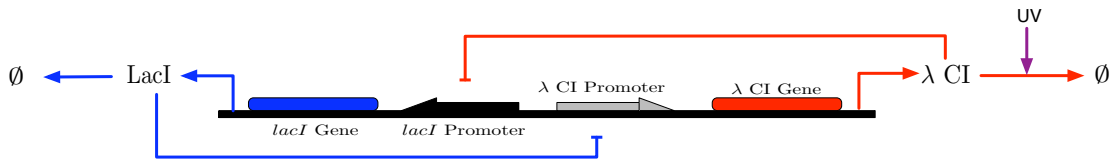

Figure 8: Schematic of the toggle model. Two proteins, λcI and LacI inhibit each other. Environmental influences (ultraviolet radiation) increase the degradation rate of λcI and affect the tradeoff between the two regulators.

We assume that four non-linear production / degradation reactions can change the populations of λcI and LacI according to:

$$\begin{array}{llll} \mathrm{R}_1 & ; \ \mathrm{R}_2 & ; \ \mathrm{R}_3 & ; \ \mathrm{R}_4 \\ \emptyset \rightarrow \lambda\mathrm{cI} & ; \ \lambda\mathrm{cI} \rightarrow \emptyset & ; \ \emptyset \rightarrow \mathrm{LacI} & ; \ \mathrm{LacI} \rightarrow \emptyset. \end{array} \tag{8}$$

The rates of these reactions, $\mathbf{w}(\lambda\mathrm{cI}, \mathrm{LacI}, \mathbf{\Lambda}) = [w_1(\lambda\mathrm{cI}, \mathrm{LacI}, \mathbf{\Lambda}), \ldots, w_4(\lambda\mathrm{cI}, \mathrm{LacI}, \mathbf{\Lambda})]$ depend upon the populations of the proteins, λcI and LacI, and the parameters in

$\mathbf{\Lambda} = \{k_{\lambda\mathrm{cI}}^{(0,1)}, \alpha_{\mathrm{LacI}}, \eta_{\mathrm{LacI}}, k_{\mathrm{LacI}}^{(0,1)}, \alpha_{\lambda\mathrm{cI}}, \eta_{\lambda\mathrm{cI}}, \delta_{\mathrm{LacI}}, \delta_{\lambda\mathrm{cI}}(\mathrm{UV})\}$, according to:

$$w_1 = k_{\lambda\mathrm{cI}}^{(0)} + \frac{k_{\lambda\mathrm{cI}}^{(1)}}{1 + \alpha_{LacI}[\mathrm{LacI}]^{\eta_{\mathrm{LacI}}}}; \qquad w_2 = \delta_{\lambda\mathrm{cI}}(\mathrm{UV})[\lambda\mathrm{cI}];$$

$$w_3 = k_{\mathrm{LacI}}^{(0)} + \frac{k_{\mathrm{LacI}}^{(1)}}{1 + \alpha_{\lambda\mathrm{cI}}[\lambda\mathrm{cI}]^{\eta_{\lambda\mathrm{cI}}}}; \qquad w_4 = \delta_{\mathrm{LacI}}[\mathrm{LacI}],$$

In the model, the λcI degradation parameter, $\delta_{\lambda\text{cI}}$, takes on different values depending upon the UV radiation level, while the remaining parameters are assumed to be independent of environmental conditions. As in [22], we have chosen a reference parameter set as follows:

$$\begin{array}{lll} k_{\lambda\text{cI}}^{(0)} = 6.8\times10^{-5}\ \text{s}^{-1} & k_{\lambda\text{cI}}^{(1)} = 1.6\times10^{-2}\ \text{s}^{-1} & \alpha_{\text{LacI}} = 6.1\times10^{-3}\ \text{N}^{-\eta_{\text{LacI}}} \\ k_{\text{LacI}}^{(0)} = 2.2\times10^{-3}\ \text{s}^{-1} & k_{\text{LacI}}^{(1)} = 1.7\times10^{-2}\ \text{s}^{-1} & \alpha_{\lambda\text{cI}} = 2.6\times10^{-3}\ \text{N}^{-\eta_{\lambda\text{cI}}} \\ \eta_{\text{LacI}} = 2.1\times10^{-0} & \eta_{\lambda\text{cI}} = 3.0\times10^{-0} & \delta_{\text{LacI}} = 3.8\times10^{-4}\ \text{N}^{-1}\text{s}^{-1}, \end{array} \tag{9}$$

where the notation N corresponds to the integer number of molecules of the relevant reacting species. We have also assumed the following for the degradation rate of λcI:

$$\delta_{\lambda\text{cI}}(\text{UV}) = 3.8^{-4} + \frac{0.002\text{UV}^2}{1250 + \text{UV}^3},$$

which has been chosen to approximate the values of

$$\{\delta_{\lambda\text{cI}}(0) = 0.00038\text{s}^{-1}, \delta_{\lambda\text{cI}}(6) = 0.00067\text{s}^{-1}, \delta_{\lambda\text{cI}}(12) = 0.0015\text{s}^{-1}\},$$

as used in [22].

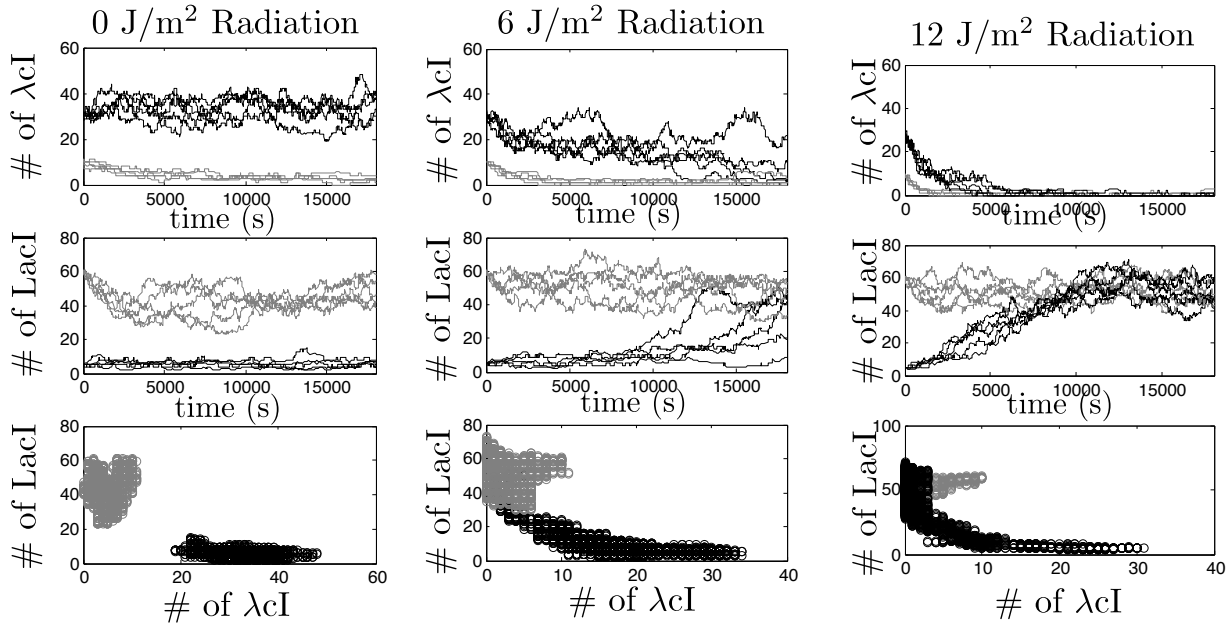

Figure 9: Trajectories of the genetic toggle switch. Two separate initial conditions are considered corresponding to (λcI,LacI) = (30,5) in black and (λcI,LacI) = (10,60) in gray. three different UV radiation levels are considered: 0J/m$^-$2, 6J/m$^-$2, and 12J/m$^-$2 in the left, center and right columns, respectively. Different aspects of the trajectories are show in the rows: λcI versus time (top), LacI versus time (middle), λcI versus LacI (bottom).

Fig. 9(left) shows numerous trajectories of the system beginning at two different initial conditions in the absence of UV radiation, one corresponding to high expression of λcI=30 and low expression of LacI=5, and the other corresponding to high expression of LacI=60 and low expression of λcI=10. Both conditions are quite stable over a period of 5 hours when there is no UV radiation. When UV radiation is applied, the degradation rate of λcI increases, and the stability of the high λcI state is decreased leading to switching to the high LacI state. This can be observed in the center and right columns corresponding to 6 and 12J/m^2 UV radiation levels. As the radiation level increases, the rate of switching also increases (compare center and right columns).

To quantify the probability of switching from the high to low λcI state, we can solve the master equation using the FSP approach. This is easily solved for using the FSP ToolKit. Fig. 10 shows the probability distribution for the amount of LacI after 1, 2, 4 and 8 hours for each of the three different UV levels. In all cases, we assume that the system began with 30 molecules of λcI and 5 molecules of LacI.

5.4 Example 4: Stochastic Resonance.

To illustrate the phenomenon of stochastic resonance, we turn to a very simple theoretical model of circadian rhythm. The model consists of a singe gene that can have two states, s_1 and s_2. When it is the system is in state s_1, it is active and rapidly produces a protein denoted as Y. This protein is assumed to bind to the s_1 with a very high cooperativity factor transforming the gene into the state s_2. The state s_1 is active and allows for the production of Y, and the state s_2 is assumed to be inactive. In addition to binding and forming state s_2, the product Y also stabilize the s_2 state, also with a high cooperativity. Mathematically, these reactions are described by:

$$R_1: \quad s_1 \to s_2, \qquad\qquad R_2: \quad s_2 \to s_1,$$

$$R_3: \quad s_1 \to s_1 + Y, \qquad\qquad R_4: \quad Y \to \emptyset.$$

The propensity functions of these reactions are

$$w_1 = \frac{s_1 y^{10}}{1000^{10} + y^{10}}, \qquad\qquad w_2 = \frac{100 s_2}{1 + y^{10}},$$

$$w_3 = 20 s_2, \qquad\qquad w_4 = 0.02 y,$$

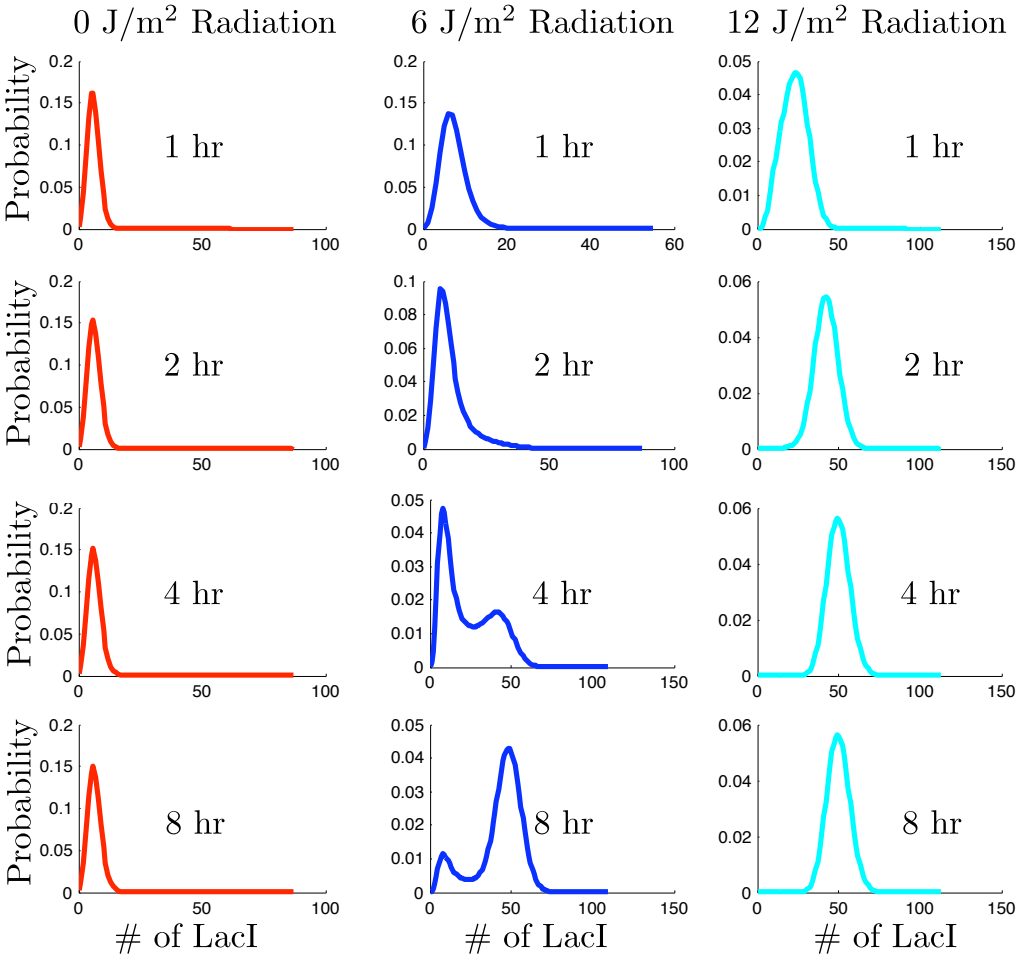

Figure 10: Distributions of LacI at different times (rows) and UV radiation levels (columns). All cells start with a high λcI and low LacI expression level (λcI=30, LacI=5). Without radiation (left column) the low LacI state is very stable, and very few cells switch. At high UV radiation (right column) almost all cells switch to the high expression state within about two hours. At a moderate UV level (center), all cells will eventually switch, but the time to do so is much longer–a significant low LacI population still exists after eight hours).

We assume that the system begins at an initial condition where the gene is in the s_1 state, and there are 100 molecules of Y. With these reactions and initial conditions, Fig. 11A, illustrates two stochastic trajectories of the system. From the figure, it is clear that the process maintains a strong oscillatory behavior, although the deterministic model of the same process reaches a steady state in only one oscillation (see the dashed lines in Fig. 11B). The oscillations of the stochastic process would continue in perpetuity, although without restoring interactions, different trajectories become desynchronized after numerous cycles. To illustrate this de-synching of the processes, the solid line in Fig. 11B shows the mean level of the stochastic process over time as computed with the FSP approach. Even after five cycles, the mean level of the population is still showing significant oscillations.

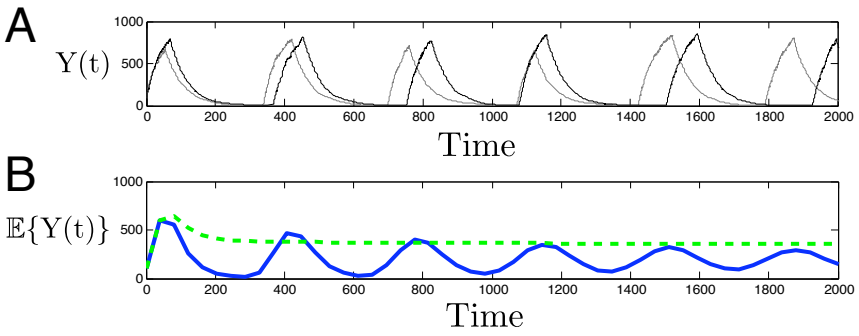

Figure 11: The effect of noise to induce and maintain oscillations. A) Two stochastic trajectories of the theoretical circadian rhythm model. B) The mean level of species Y as a function of time as computed with the stochastic model (solid line) or with the deterministic ODE model (dashed line).

6 Identifying stochastic models of gene regulation.

As has been seen already in the previous examples, different mechanisms or parameters can cause biochemical systems to exhibit different behaviors with respect to their fluctuations and cell-to-cell variability. As a result, these fluctuations contain additional information about the underlying system, which might not be obtainable from the mean level behavior [52]. In turn, this information could enable researchers to identify mechanisms and parameters of gene regulatory constructs [52, 22]. To illustrate this approach, we will identify a model of regulation of the *lac* operon in *E. coli* under the induction of IPTG. For this identification, we utilize experimental data from [52], but we attempt to fit a simpler model to this data.

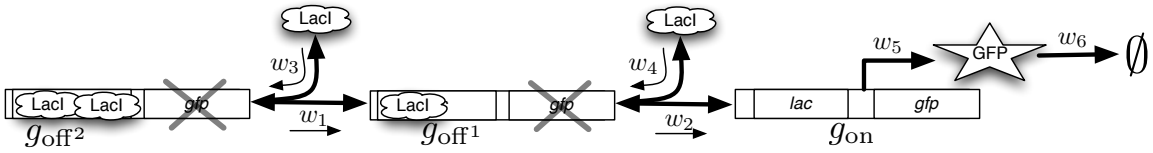

Figure 12: Schematic of the *lac* induction model with cooperative activation by IPTG.

The chosen model corresponds to a gene that can be in three distinct states, denoted as g_{off^2}, g_{off^1} and g_{on} corresponding to when two, one or zero molecules of LacI are bound to the *lac* operon (see Fig. 12). The unbinding of LacI is assumed to be the same for both molecules and can be described by the reactions:

$$\mathcal{R}_1 : g_{\text{off}^2} \xrightarrow{w_1} g_{\text{off}^1}, \quad \mathcal{R}_2 : g_{\text{off}^1} \xrightarrow{w_2} g_{\text{on}}$$

where each unbinding transition is assumed to depend upon the level of IPTG according to

$$w_1 = (\kappa_0 + \kappa_1[\text{IPTG}])\, g_{\text{off}^2}, \qquad w_2 = (\kappa_0 + \kappa_1[\text{IPTG}])\, g_{\text{off}^1}.$$

In this model, the total level of LacI, $[\text{LacI}]_{\text{Tot}}$, is assumed to be constant, but the effective amount of LacI free to bind is diminished through the action of IPTG according to the expression:

$$[\text{LacI}]_{\text{eff}} = \frac{[\text{LacI}]_{\text{Tot}}}{\beta + [\text{IPTG}]}.$$

The binding rate is then simply a constant times the effective LacI level. In order to capture the effect of cooperatively, this constant is allowed to depend upon whether it is the first or second LacI molecule to bind.

$$w_3 = \frac{\alpha_1}{\beta + [\text{IPTG}]} g_{\text{off}^1}, \qquad w_4 = \frac{\alpha_2}{\beta + [\text{IPTG}]} g_{\text{on}},$$

For the current study, the level of IPTG is assumed to be constant for all times $t > 0$.

Production of GFP occurs only when the gene is in the g_{on} state and has the rate: $w_5 = k_{\text{G}} g_{\text{on}}$. The propensity for degradation of GFP is the standard linear degradation: $w_6 = \gamma_{\text{GFP}} y$. Because GFP is known to be a stable protein, its degradation rate is set to the dilution rate of $\gamma_{\text{GFP}} = 3.8 \times 10^{-4}$, and there remain five unknown positive real parameters for the regulatory system:

$$\mathbf{\Lambda} = \{\kappa_0, \kappa_1, \alpha_1, \alpha_2, \beta, k_{\text{G}}\} \in \mathbb{R}^6_+.$$

In addition to these parameters which describe the evolution of the probability distribution for the GFP population, it is also necessary to account for the background fluorescence and variability in the fluorescence of individual GFP molecules. The values for these quantities were previously obtained in [52]. In particular, the background fluorescence was assumed to be independent of the IPTG levels, and was measured at each instant in time [52]. The mean, $\mu_{GFP} = 220$ AU, and standard deviation, $\sigma_{GFP} = 390$ AU, in the fluorescence per GFP molecule are also taken from [52]. It is important to note that the parameters μ_{GFP} and σ_{GFP} are highly dependent upon the flow cytometer and its measurement settings, particularly the thresholds for event detection and the amplification of the fluorescence detector.

The current model is fit to the measurements of GFP fluorescence at $\{5, 10, 20, 40, 100\}\mu$M IPTG and at times of $\{0, 3, 4, 5\}$ hours after induction. These data are in arbitrary units of fluorescence,

and have been collected into sixty logarithmically distributed increments between 10 and 10^5. The two separate data sets are shown with solid lines in Fig. 13, where the columns correspond to different time points and the rows correspond to different IPTG induction levels. For the initial conditions, we assumed that every cell begins in the g_{off^2} state with zero molecules of GFP and no IPTG at a time of three hours before induction ($t = -3\text{hr}$). The fit is begun with an initial guess of unity for β and 10^{-4} for the remaining five parameters, and the search is run using numerous iterations of Matlab's *fminsearch* and a simulated annealing algorithm. The objective for the fit is to match the fluorescence distribution as close as possible in the 1-norm sense for all times and IPTG levels. The electronic data and the codes for fitting this data can be downloaded from http://cnls.lanl.gov/~munsky or can be requested from the author at munsky@lanl.gov.

For the chosen model, Fig. 13 shows the distributions of the measured and fitted GFP fluorescence levels in the various experimental conditions. From the figure, one can see that this simplified model does indeed capture the qualitative and quantitative features of the distributions at the different times and IPTG induction levels (compare dashed lines to solid lines). The final parameter values of the fit were found to be:

$$\kappa_0 = 1.91 \times 10^{-5}\ \text{s}^{-1}, \qquad \kappa_1 = 3.21 \times 10^{-6}\ \mu\text{M}^{-1}\text{s}^{-1},, \qquad \beta = 4.88 \times 10^{2}\ \mu\text{M},$$
$$\alpha_1 < 1.0 \times 10^{-10}\ \mu\text{M}s^{-1}, \qquad \alpha_2 = 5.36 \times 10^{-1}\ \mu\text{M}s^{-1}, \qquad k_{\text{G}} = 8.09 \times 10^{-2}\ \text{s}^{-1},$$

It is useful to note that the parameter, α_1 is many orders of magnitude smaller than the parameter α_2, which has the same units. Furthermore, the time scale of reaction three is on the order of $w_3^{-1} \approx \beta/\alpha_1 \gg 10^{10}s^{-1}$, which is far longer than the experimental time. This suggests that this parameter is not necessary for the model to fit the data. Indeed, setting $\alpha_1 = 0$ results in no appreciable difference in the model fits for any of the cases. In other words, the state g_{off^2} is not needed in the current model to capture the data, suggesting that this state is unobservable from the current data set.

For the fits shown in Fig. 13, we have used all of the data and found a single parameter set. It is also interesting to determine how well smaller subsets of the data would do to (i) constrain the model parameters and (ii) enable predictions of the other conditions. To examine this, we attempt to identify the model from each possible combination of three or four different IPTG levels. Table 1 and Fig. 14 show the parameters that have been identified with each of these data sets and Table

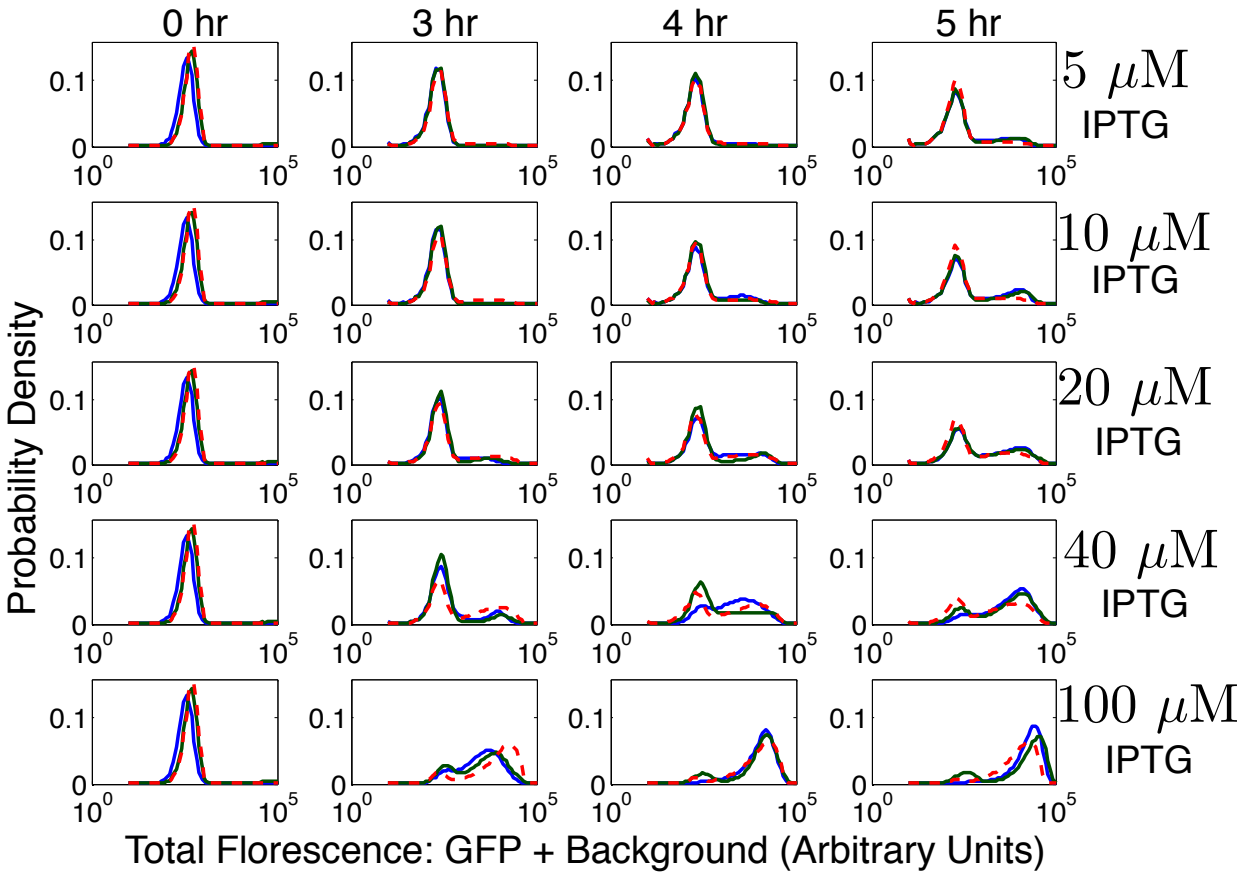

Figure 13: Measured (solid lines) and computed (dashed lines) histograms of *gfp* expression under the control of the *lac* operon and induced with IPTG. The columns correspond to different measurement times (0,3,4,5)hr after induction. The rows correspond to different levels of extra-cellular IPTG induction (5,10,20,40,100)μM. Experimental data is reproduced from [52], but a different model is used to fit this data as described in the text.

2 shows the one norm errors in each of the different data sets, where 1-norm prediction errors are shown in bold face. From the fits resulting from the various data subsets, it is possible to determine which data sets are the most predictive of the remaining conditions. In particular, when four data sets are available, the best overall fit is found when all but the 40μM IPTG concentration is used, meaning that the information learned from that condition is redundant to the information contained in the other conditions. Leaving two data sets out, shows that the 20μM IPTG concentration is also easily predicted from the remaining data sets. By leaving out one or two data sets, we are able to characterize the uncertainty in the parameter values. With three different IPTG concentrations,

IPTG Levels Used for Fit (μM)	**Best Fit Parameter Values**					
	κ_0 s^{-1}	κ_1 $\mu M^{-1}s^{-1}$	β μM	α_1 $\mu M s^{-1}$	α_2 $\mu M s^{-1}$	k_G s^{-1}
$\{5, 10, 20\}\mu M$	1.22e-05	3.94e-06	7.76e+02	<1e-10	5.52e-01	5.53e-02
$\{5, 10, 40\}\mu M$	1.74e-05	3.34e-06	7.84e+02	<1e-10	1.66e+00	1.29e-01
$\{5, 10, 100\}\mu M$	1.89e-05	3.43e-06	6.23e+02	<1e-10	7.53e-01	8.54e-02
$\{5, 20, 40\}\mu M$	1.18e-05	3.60e-06	1.72e+03	<1e-10	1.59e+00	7.13e-02
$\{5, 20, 100\}\mu M$	1.38e-05	3.45e-06	1.63e+03	<1e-10	1.64e+00	8.15e-02
$\{5, 40, 100\}\mu M$	1.32e-05	3.39e-06	2.12e+02	<1e-10	3.29e-01	9.18e-02
$\{10, 20, 40\}\mu M$	2.44e-05	2.88e-06	1.03e+04	<1e-10	9.14e+00	7.19e-02
$\{10, 20, 100\}\mu M$	2.43e-05	3.03e-06	8.04e+02	<1e-10	7.17e-01	7.65e-02
$\{10, 40, 100\}\mu M$	2.57e-05	3.07e-06	2.00e+02	<1e-10	2.80e-01	8.79e-02
$\{20, 40, 100\}\mu M$	1.00e-07	4.37e-06	6.87e+02	<1e-10	7.53e-01	7.92e-02
$\{5, 10, 20, 40\}\mu M$	1.43e-05	3.61e-06	2.47e+03	<1e-10	2.33e+00	7.18e-02
$\{5, 10, 20, 100\}\mu M$	1.99e-05	3.22e-06	3.89e+02	<1e-10	3.81e-01	7.52e-02
$\{5, 10, 40, 100\}\mu M$	2.03e-05	3.27e-06	2.06e+02	<1e-10	3.13e-01	9.07e-02
$\{5, 20, 40, 100\}\mu M$	1.38e-05	3.46e-06	6.90e+02	<1e-10	7.94e-01	8.38e-02
$\{10, 20, 40, 100\}\mu M$	2.35e-05	3.05e-06	4.77e+02	<1e-10	4.91e-01	7.89e-02
$\{5, 10, 20, 40, 100\}\mu M$	1.91e-05	3.21e-06	4.88e+02	<1e-10	5.36e-01	8.09e-02

Table 1: Parameter Sets for the various data subsets for the *lac* regulation model.

the uncertainty on the parameters can be estimated as:

$$\kappa_0 = 1.62 \times 10^{-5} \pm 7.75 \times 10^{-6}\ \mathrm{s}^{-1}, \qquad \kappa_1 = 3.45 \times 10^{-6} \pm 4.44 \times 10^{-7}\ \mu\mathrm{M}^{-1}\mathrm{s}^{-1},$$
$$\beta = 1.78 \times 10^{3} \pm 3.05 \times 10^{3}\ \mu\mathrm{M}, \qquad \alpha_2 = 1.74 \times 10^{0} \pm 2.65 \times 10^{0}\ \mu\mathrm{M}s^{-1},$$
$$k_\mathrm{G} = 8.30 \times 10^{-2} \pm 1.91 \times 10^{-2}\ \mathrm{s}^{-1},$$

where the values are listed as the mean plus or minus one standard deviation. From these results it is clear that the values of κ_0, κ_1 and k_G are well determined from just three different IPTG concentrations, but the other values are more poorly constrained. By adding a fourth IPTG concentration, the uncertainty drops considerably for all six parameters as follows

$$\kappa_0 = 1.84 \times 10^{-5} \pm 4.18 \times 10^{-6}\ \mathrm{s}^{-1}, \qquad \kappa_1 = 3.32 \times 10^{-6} \pm 2.20 \times 10^{-7}\ \mu\mathrm{M}^{-1}\mathrm{s}^{-1},$$
$$\beta = 8.47 \times 10^{2} \pm 9.26 \times 10^{2}\ \mu\mathrm{M}, \qquad \alpha_2 = 8.63 \times 10^{0} \pm 8.43 \times 10^{-1}\ \mu\mathrm{M}s^{-1},$$
$$k_\mathrm{G} = 8.01 \times 10^{-2} \pm 7.44 \times 10^{-3}\ \mathrm{s}^{-1}.$$

From these values it is clear that the addition of the fourth concentration goes a long way toward helping to constrain the parameters.

The code "'FSP Fit_Tools" provides a simple graphical user interface with which these different fits can be obtained and plotted for each of the different subsets of data. In this example, we have

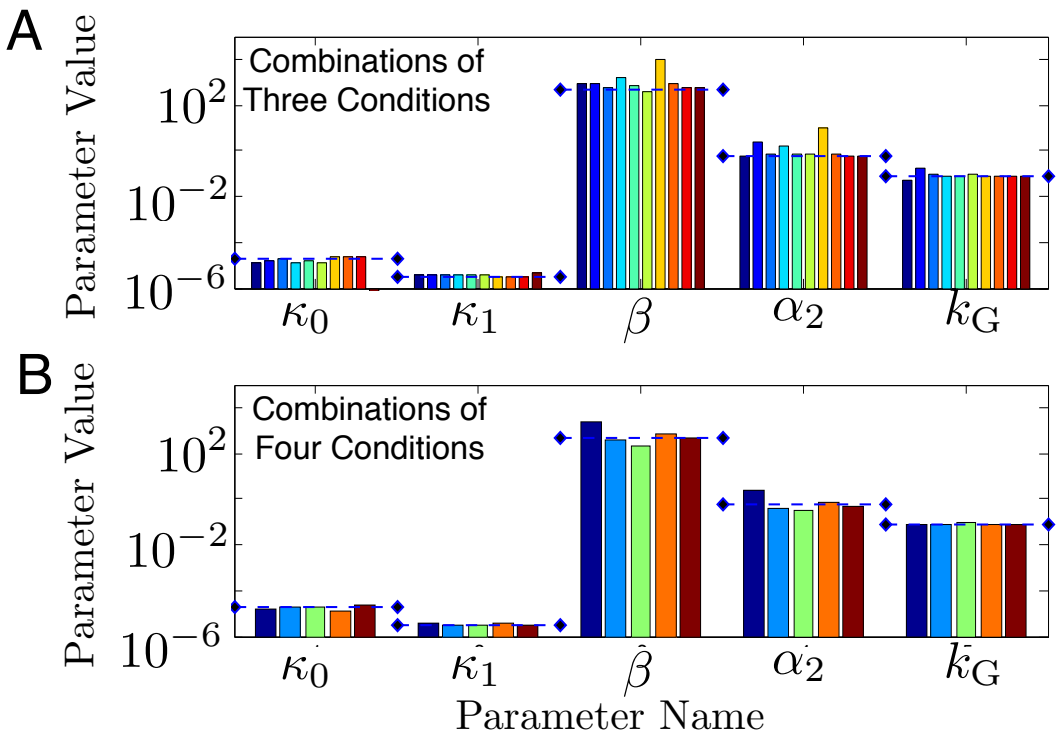

Figure 14: Identified parameters for the induction of *lac* with IPTG. A) Parameters identified with every possible combination of three different IPTG concentration from $\{5, 10, 20, 40, 100\}\mu$M. B) Parameters identified with every possible combination of four different IPTG concentrations. In each set of bars, the diamonds and the horizontal dashed lines correspond to the parameter set identified from all five IPTG levels.

considered only a single model for the IPTG induction of the *lac* operon, and we have obtained a single parameter set with which this model does a good job of capturing the observed experimental behavior. Other models will perform better or worse than that presented here. We encourage the interested reader to use the provided FSP Toolkit codes to propose and test alternate models for this system. Also included in the online software is an example of identifying a model of the toggle switch (Section 5.3) from simulated data of the marginal and full distributions at different levels of UV radiation (see also [22]).

7 Summary

This chapter has presented a few of the phenomena that result from discrete stochastic reactions, including stochastic amplifications, stochastic damping, stochastic resonance and stochastic switching. For each of these phenomena, we have used Finite State Projection analysis tools to illustrate these behaviors. We showed how different mechanisms and parameters lead to different responses in the face of stochasticity, and we illustrated how it is possible to use information about the variability

IPTG Levels Used for Fit (μM)	**One Norm Differences from Data**					
	5μM	10μM	20μM	40μM	100μM	Total
$\{5, 10, 20\}\mu$M	8.35e-01	1.02e+00	1.03e+00	**1.69e+00**	**1.93e+00**	6.51e+00
$\{5, 10, 40\}\mu$M	8.47e-01	1.02e+00	**1.07e+00**	1.64e+00	**1.82e+00**	6.40e+00
$\{5, 10, 100\}\mu$M	8.57e-01	1.01e+00	**1.05e+00**	**1.65e+00**	1.78e+00	6.35e+00
$\{5, 20, 40\}\mu$M	8.34e-01	**1.05e+00**	1.04e+00	1.65e+00	**1.81e+00**	6.38e+00
$\{5, 20, 100\}\mu$M	8.36e-01	**1.04e+00**	1.04e+00	**1.65e+00**	1.79e+00	6.36e+00
$\{5, 40, 100\}\mu$M	8.34e-01	**1.06e+00**	**1.07e+00**	1.64e+00	1.79e+00	6.39e+00
$\{10, 20, 40\}\mu$M	**8.85e-01**	1.00e+00	1.04e+00	1.65e+00	**1.84e+00**	6.42e+00
$\{10, 20, 100\}\mu$M	**8.91e-01**	1.00e+00	1.04e+00	**1.66e+00**	1.78e+00	6.37e+00
$\{10, 40, 100\}\mu$M	**8.97e-01**	1.01e+00	**1.07e+00**	1.65e+00	1.77e+00	6.40e+00
$\{20, 40, 100\}\mu$M	**9.20e-01**	**1.12e+00**	1.03e+00	1.65e+00	1.88e+00	6.60e+00
$\{5, 10, 20, 40\}\mu$M	8.39e-01	1.03e+00	1.04e+00	1.66e+00	**1.81e+00**	6.37e+00
$\{5, 10, 20, 100\}\mu$M	8.59e-01	1.01e+00	1.04e+00	**1.66e+00**	1.78e+00	6.35e+00
$\{5, 10, 40, 100\}\mu$M	8.60e-01	1.01e+00	**1.07e+00**	1.65e+00	1.78e+00	6.36e+00
$\{5, 20, 40, 100\}\mu$M	8.35e-01	**1.04e+00**	1.04e+00	1.65e+00	1.79e+00	6.36e+00
$\{10, 20, 40, 100\}\mu$M	**8.80e-01**	1.00e+00	1.05e+00	1.65e+00	1.78e+00	6.36e+00
$\{5, 10, 20, 40, 100\}\mu$M	8.55e-01	1.02e+00	1.04e+00	1.65e+00	1.78e+00	6.35e+00

Table 2: One norm errors in the distributions for the various fits. Values shown in regular fonts correspond to the differences for data used in the fitting procedure, whereas values in bold face correspond to errors in the predicted distributions.

of individual cells to help infer regulatory mechanisms and parameters from single cell data. For the readers' convenience, all examples included in this work can be reproduced using FSP Toolkit codes, which can be downloaded from http://cnls.lanl.gov/∼munsky or can be requested from the author at munsky@lanl.gov. .

Acknowledgments

The author would like to thank Mustafa Khammash for help in the development of the FSP methodology in Section 3; Brooke Trinh for the experimental data in Section 6; Chunbo Lou, Stephen Payne, Alvin Tamsir and other students at the 4^{th} annual q-bio Summer Summer School on Cellular Information Processing for their feedback on some of the examples, and CNLS for providing a stimulating environment in which to pursue this research.

References

[1] McAdams, M., and A. Arkin. 1999. Its a noisy business! *Tren. Gen.* 15:65–69.

[2] Elowitz, M., A. Levine, E. Siggia, and P. Swain. 2002. Stochastic gene expression in a single cell. *Science.* 297:1183–1186.

[3] Thattai, M., and A. van Oudenaarden. 2001. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci.* 98:8614–8619.

[4] Hasty, J., J. Pradines, M. Dolnik, and J. Collins. 2000. Noise-based switches and amplifiers for gene expression. *PNAS*. 97:2075–2080.

[5] Ozbudak, E., M. Thattai, I. Kurtser, A. Grossman, and A. van Oudenaarden. 2002. Regulation of noise in the expression of a single gene. *Nature Genetics*. 31:69–73.

[6] Federoff, N., and W. Fontana. 2002. Small numbers of big molecules. *Science*. 297:1129–1131.

[7] Kepler, T., and T. Elston. 2001. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys. J.* 81:3116–3136.

[8] Becskei, A., and L. Serrano. 2000. Engineering stability in gene networks by autoregulation. *Nature*. 405:590–593.

[9] Dublanche, Y., K. Michalodimitrakis, N. Kummerer, M. Foglierini, and L. Serrano. 2006. Noise in transcription negative feedback loops: simulation and experimental analysis. *Molecular Systems Biology*. 2.

[10] Nevozhay, D., R. Adams, K. Murphy, K. Josic, and G. Balazsi. 2009. Negative autoregulation linearizes the dose-response and suppresses the heterogeneity of of gene expression. *Proc. Nat. Acad. Sci. USA*. 106:5123–5128.

[11] Thieffry, D., A. Huerta, E. Perez-Rueda, and J. Collado-Vides. 1998. From specific gene regulation to genomic networks. *Bioseeays*. 20:433–440.

[12] Tan, C., F. Reza, and L. You. 2007. Noise-limited frequency signal transmission in gene circuits. *Biophysical Journal*. 93:3753–3761.

[13] Sohka, T., R. Heins, R. helan, J. Greisler, C. Townsend, and M. Ostermeier. 2009. An externally tunable bacterial band-pass filter. *Proc. Nat. Acad. Sci.* 106:10135–10140.

[14] Paulsson, J., O. Berg, and M. Ehrenberg. 2000. Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation. *PNAS*. 97:7148–7153.

[15] Li, H., Z. Hou, and H. Xin. 2005. Internal noise stochastic resonance for intracellular calcium oscillations in a cell system. *Phys. Rev. E*. 71.

[16] Arkin, A., J. Ross, and M. H. 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*. 149:1633–1648.

[17] Wolf, D., and A. Arkin. 2002. Fifteen minutes of fim: Control of type 1 pili expression in e. coli. *OMICS: A Journal of Integrative Biology*. 6:91–114.

[18] Munsky, B., A. Hernday, D. Low, and M. Khammash. 2005. Stochastic modeling of the pap-pili epigenetic switch. *Proc. FOSBE*. :145–148.

[19] Tian, T., and K. Burrage. 2006. Stochastic models for regulatory networks of the genetic toggle switch. *PNAS*. 103:8372–8377.

[20] Cagatay, T., M. Turcotte, M. Elowitz, J. Garcia-Ojalvo, and G. Suel. 2009. Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell*. 139:512–522.

[21] Kobayashi, H., M. Kaern, M. Araki, K. Chung, T. Gardner, C. Cantor, and J. Collins. 2004. Programmable cells: Interfacing natural and engineered gene networks. *PNAS*. 101:8414–8419.

[22] Munsky, B., and M. Khammash. 2010. Guidelines for the identification of a stochastic model for the genetic toggle switch. *IET Systems Biology*. 4:356–366.

[23] Raj, A., and A. van Oudenaarden. 2009. Single-molecule approaches to stochastic gene expression. *Annual Review of Biophysics*. 38:255–270.

[24] Pardue, M., and G. J. 1969. Molecular hybridization of radioactive dna to the dna of cytological preparation. *Proc Nat Acad Sci USA*. 64:600–604.

[25] John, H., M. Birnstiel, and K. Jones. 1969. Rna-dna hybrids at the cytological level. *Nature*. 223:582–587.

[26] Raj, A., P. van den Bogaard, S. Rifkin, A. van Oudenaarden, and S. Tyagi. 2008. Imaging individual mrna molecules using multiple singly labeled probes. *Nature Methods*. 5:877–887.

[27] Ghosh, I., A. Hamilton, , and L. Regan. 2000. Antiparallel leucine zipper-directed protein reassembly: Application to the green fluorescent protein. *J. American Chemical Society*. 122:5658–5659.

[28] Cabantous, S., T. Terwilliger, and G. Waldo. 2004. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nature Biotechnology*. 23:845–854.

[29] Cabantous, S., and G. Waldo. 2006. In vivo and in vitro protein solubility assays using split gfp. *Nature Methods*. 3:845–854.

[30] Shapiro, H. 2003. Practical Flow Cytometry, 4th Ed. Wiley-Liss.

[31] Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340–2360.

[32] Gillespie, D. T. 2001. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* 115:1716–1733.

[33] Allen, R., P. Warren, and P. Rein ten Wolde. 2005. Sampling rare switching events in biochemical networks. *Phys. Rev. Lett.* 94.

[34] Rao, C. V., and A. P. Arkin. 2003. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the gillespie algorithm. *J. Chem. Phys.* 118:4999–5010.

[35] Cao, Y., D. Gillespie, and L. Petzold. 2005. The slow-scale stochastic simulation algorithm. *J. Chem. Phys.* 122.

[36] Warmflash, A., P. Bhimalapuram, and A. Dinner. 2007. Umbrella sampling for nonequilibrium processes. *J. Chem. Phys.* 127.

[37] van Kampen, N. 2007. Stochastic Processes in Physics and Chemistry, 3rd Ed. Elsevier.

[38] Elf, J., and M. Ehrenberg. 2003. Fast evaluations of fluctuations in biochemical networks with the linear noise approximation. *Genome Research*. 13:2475–2484.

[39] Nasell, I. 2003. An extension of the moment closure method. *Theoretical Population Biology*. 64:233–239.

[40] Gmez-Uribe, C., and G. Verghese. 2007. Mass fluctuation kinetics: Capturing stochastic effects in systems of chemical reactions through coupled mean-variance computations. *JCP*. 126.

[41] Singh, A., and J. Hespanha. 2007. A derivative matching approach to moment closure for the stochastic logistic model. *Bulletin of Mathematical Biology*. 69:1909–1925.

[42] Sinitsyn, N., N. Hengartner, and I. Nemenman. 2009. Adiabatic coarse-graining and simulations of stochastic biochemical networks. *Proc. Nat. Acad. Sci. U.S.A.* 106:10546–10551.

[43] Walczak, A., A. Mugler, and C. Wiggins. 2009. A stochastic spectral analysis of transcriptional regulatory cascades. *Proc. Nat. Acad. Sci.* 106:6529–6534.

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

[45] Burrage, K., M. Hegland, S. Macnamara, and R. Sidje. 2006. A krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems. *Proc. of The A.A.Markov 150th Anniversary Meeting.* :21–37.

[46] Munsky, B., and M. Khammash. 2008. The finite state projection approach for the analysis of stochastic noise in gene networks. *IEEE Trans. Automat. Contr./IEEE Trans. Circuits and Systems: Part 1.* 52:201–214.

[47] Munsky, B. 2008. The finite state projection approach for the solution of the chemical master equation and its application to stochastic gene regulatory networks. Ph.D. thesis, Univ. of California at Santa Barbara, Santa Barbara.

[48] Warmflash, A., and A. Dinner. 2008. Signatures of combinatorial regulation in intrinsic biological noise. *Proc. Nat. Acad. Sci. USA.* 105:17262–17267.

[49] Dunlop, M., R. Cox III, J. Levine, R. Murray, and M. Elowitz. 2008. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature Genetics.* 40:1493–1498.

[50] de Ronde, W., B. Daniels, A. Mugler, N. Sinitsyn, and I. Nemenman. 2009. Mesoscopic statistical properties of multistep enzyme-mediated reactions. *IET Syst. Biol.* 3:429–437.

[51] Mettetal, J., D. Muzzey, J. Pedraza, E. Ozbudak, and A. van Oudenaarden. 2006. Predicting stochastic gene expression dynamics in single cells. *Proc. Natl. Acad. Sci. U S A.* 103:7304–7305.

[52] Munsky, B., B. Trinh, and M. Khammash. 2009. Listening to the noise: random fluctuations reveal gene network parameters. *Molecular Systems Biology.* 5.

[53] Thorsley, D., and E. Klavins. 2010. Approximating stochastic biochemical processes with wasserstein pseudometrics. *IET Systems Biology.* 4:193–211.

[54] Gillespie, D. T. 1992. A rigorous derivation of the chemical master equation. *Physica A.* 188:404–425.

[55] Bel, G., B. Munsky, and I. Nemenman. 2010. Simplicity of completion time distributions for common complex biochemical processes. *Physical Biology.* 7.

[56] Haseltine, E., and J. Rawlings. 2002. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.* 117:6959–6969.

[57] Salis, H., and Y. Kaznessis. 2005. Accurate hybrid stochastic simulation of a system of coupled chemical or biological reactions. *J. Chem. Phys.* 112.

[58] Tian, T., and K. Burrage. 2004. Binomial leap methods for simulating stochastic chemical kinetics. *J. Chem. Phys.* 121:10356–10364.

[59] Cao, Y., D. T. Gillespie, and L. R. Petzold. 2005. Avoiding negative populations in explicit poisson tau-leaping. *J. Chem. Phys.* 123.

[60] Rathinam, M., L. R. Petzold, Y. Cao, and D. T. Gillespie. 2003. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *J. Chem. Phys.* 119:12784–12794.

[61] Gillespie, D. T. 2000. The chemical langevin equation. *J. Chem. Phys.* 113:297–306.

[62] Munsky, B., and M. Khammash. 2008. Transient analysis of stochastic switches and trajectories with applications to gene regulatory networks. *IET Systems Biology.* 2:323–333.

[63] Munsky, B., and M. Khammash. 2007. A multiple time interval finite state projection algorithm for the solution to the chemical master equation. *J. Comp. Phys.* 226:818–835.

[64] Munsky, B., and M. Khammash. 2006. A reduced model solution for the chemical master equation arising in stochastic analyses of biological networks. *Proc. 45th IEEE Conference on Decision and Control.* :25–30.

[65] Peles, S., B. Munsky, and M. Khammash. 2006. Reduction and solution of the chemical master equation using time-scale separation and finite state projection. *J. Chem. Phys.* 125.

[66] Munsky, B., S. Peles, and M. Khammash. 2007. Stochastic analysis of gene regulatory networks using finite state projections and singular perturbation. *Proc. 26th American Control Conference (ACC).* :1323–1328.

[67] Sidje, R. B. 1998. EXPOKIT: Software package for computing matrix exponentials. *ACM Transactions on Mathematical Software.* 24:130–156.

[68] Gardner, T., C. Cantor, and J. Collins. 2000. Construction of a genetic toggle switch in escherichia coli. *Nature*. 403:339–242.

[69] Warren, P., and P. Rein ten Wolde. 2005. Chemical models of genetic toggle switches. *J. Phys. Chem. B*. 109:6812–6823.

[70] Lipshtat, A., A. Loinger, N. Balaban, and O. Biham. 2006. Genetic toggle switch without cooperative binding. *Phys. Rev. Lett.* 96.