Multi-modality in gene regulatory networks with slow gene binding

M. Ali Al-Radhawi\textsuperscript{1}, Domitilla Del Vecchio\textsuperscript{1}, and Eduardo D. Sontag\textsuperscript{2}

\textsuperscript{1}Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA. Emails: \{malirdwi,ddv\}@mit.edu
\textsuperscript{2}Department of Mathematics, Rutgers University, Piscataway, NJ 08854-8019, USA. Email: eduardo.sontag@rutgers.edu

May 8, 2017

Abstract

The choice between deterministic and stochastic modeling can lead to dramatically different theoretical conclusions regarding the steady state behavior of a gene regulatory network (GRN). This is particularly interesting when low-molecular counts and slow TF-gene binding/unbinding lead to the emergence of new phenotypes in the stochastic model that are not reflected in the corresponding deterministic model. This work uncovers a mechanism that underlies this emergence of multiple modes under slow slow promoter kinetics, and studies it theoretically. Mathematical tools from singular perturbation theory are employed in order to analytically characterize stationary distributions of Chemical Master Equations for GRN’s, in the limit of slow switching, as a mixture of Poisson distributions. This approach, which may be interpreted as a finite-dimensional reduction of a countable Markov chain, offers a rigorous framework to explain phenomena such as non-genetic population heterogeneity and transcriptional bursting. As illustrations, the theory is used in order to tease out the role of cooperative binding in stochastic models in comparison to deterministic models, and applications are given to various model systems, including isolated or populations of toggle switches and a trans-differentiation network.

Keywords: Gene Regulatory Networks, Multi-Modality, Singular Perturbations, Slow gene binding, Slow Promoter Kinetics, Markov chains, Master Equation, Cooperativity

1 Introduction

A gene regulatory network (GRN) consists of a collection of genes that transcriptionally regulate each other through their expressed proteins. Through these interactions, which include positive and negative feedback loops, GRN’s play a central
role in the overall control of cellular life. The behavior of such networks is stochastic because of the random nature of transcription, translation, and post-translational protein modification processes, as well as the varying availability of cellular components that are required for gene expression. Stochasticity in GRN’s is a source of phenotypic variation among genetically identical (clonal) populations of cells or even organisms, and is considered to be one of the mechanisms facilitating cell differentiation and organism development. This phenotypic variation may also confer a population advantage when facing fluctuating environments. Stochasticity due to randomness in cellular components and transcriptional and translational processes have been thoroughly researched.

The fast equilibration of random processes sometimes allows stochastic behavior to be “averaged out” through the statistics of large numbers at an observational time-scale, especially if genes and proteins are found at large copy numbers. In those cases, an entire GRN, or portions of it, might be adequately described by a deterministic model. Stochastic effects that occur at a slower time scale, however, render a deterministic analysis inappropriate and might alter the steady-state behaviour of the system. This paper addresses a central question about GRN’s: how many different “stable steady states” can such a system potentially settle upon, and how does stochasticity, or lack thereof, affect the answer? To answer this question, it is necessary to understand the possibly different predictions that follow from stochastic versus deterministic models of gene expression. Indeed, qualitative conclusions regarding the steady-state behavior of gene expression levels in a GRN are critically dependent on whether a deterministic or stochastic model is used (see for a recent review). It follows that the mathematical characterization of phenomena such as non-genetic phenotype heterogeneity, switching behavior in response to environmental conditions, and lineage conversion in cells, will depend on the choice of the model.

In order to make the discussion precise, we must clarify the meaning of the term “stable steady state” in both the deterministic and stochastic frameworks. Deterministic models are employed when molecular concentrations are large, or if stochastic effects can be averaged out. They consist of systems of ordinary differential equations describing averaged-out approximations of the interactions between the various molecular species in the GRN under study. For these systems, steady states are the zeroes of the vector field defining the dynamics, and “stable” states are those that are locally asymptotically stable. The number of such stable states quantifies the degree of “multi-stability” of the system. Stochastic models of GRN’s, in contrast, are based upon continuous-time Markov chains which describe the random evolution of discrete molecular count numbers. Their long-term behavior is characterized by a stationary probability distribution that describes what gene activity configurations and protein numbers are recurrently visited. Under weak ergodicity assumptions, this stationary distribution is unique, so multi-stability in the sense of multiple steady states of the Markov chain is not an interesting notion. A biologically meaningful notion of “multi-stability” in this context, and the one that we employ in our study, is “multi-modality,” meaning the existence of multiple modes (local maxima) of stationary distributions.

Intuitively, given a multi-stable deterministic system, adding noise may help to “shaken” states, dislodging them from one basin of attraction of one stable
state, and sending them into another basin of attraction of another stable state. Therefore, in the long run, we are bound to see the various deterministic stable steady states with higher probability, that is to say, we expect that they will appear as modes in the stationary distribution of the Markov chain of the associated stochastic model. This is indeed a typical way in which modes can be interpreted as corresponding to stable states, with stochasticity responsible for the transitions between multiple stable states \[13\]. However, totally new modes could arise in the stationary distribution of a stochastic system besides those associated to stable states of the deterministic model, and this can occur even if the deterministic model had just a single stable state. This phenomenon of “stochastic multi-stability” has attracted considerable attention lately, both in theoretical and experimental work \[8, 10, 14, 15, 16\]. Stochastic multi-stability has been linked to behaviors such as transcriptional bursting/pulsing \[17, 18\], GRN’s binary response \[19\], and multi-state gene transcription \[2\] has been used to propose explanations for phenotypic heterogeneity in isogenic populations. A common assumption in gene expression models is that gene binding is significantly faster than the rate of protein production and decay \[1\]. However, it has been proposed \[10, 21\] that the emergence of new modes in stochastic systems in addition to those that arise from the deterministic model might be due to low gene copy numbers and to slow promoter kinetics, which means that process of binding and unbinding of transcription factors (TFs) to promoters is slow. Thus, the emergence of multi-modality may be due to the slow TF-gene binding and unbinding, especially in the case of eukaryotic cells, which involve a more sophisticated transcription machinery than prokaryotes. For example, the presence of nucleosomes makes binding sites less accessible to transcription factors and therefore TF-gene binding/unbinding is modulated by the stochastic process of chromatin opening \[21, 10, 22, 23\]. DNA methylation \[24\] is another mechanism that has been reported to slow down TF-gene binding/unbinding. Several experiments have confirmed the role of complex transcription processes such as epigenetic changes in slow promoter kinetics \[18, 25, 26, 24\]. The basic constitutive gene expression model (see §2) has been validated for transcriptional bursting \[18\].

In summary, new modes may appear in the stationary distribution that do not correspond to stable states in the deterministic model. Conversely, multiple steady states in the deterministic model may collapse, being “averaged out” by noise, with a single mode representing their mean. It is a well-established fact that, in general, multi-stability of the deterministic description of a biochemical network and multimodality of the associated stochastic model do not follow from each other \[27\]. This is especially true in low copy numbers regimes and with slow promoter kinetics.

Here, we pursue a mathematical analysis of the role of slow promoter kinetics in producing multi-modality in GRNs. Previous studies of the master equation for single-genes had already observed the emergence of bimodality with slow TF-gene binding/unbinding \[28, 20, 29, 30\], a phenomenon which was also studied by taking the limit of slow promoter kinetics using hybrid stochastic models for single genes \[31, 32, 33\]. However, and despite a wide range of its application relevance, the mathematical analysis of multi-gene networks with slow promoter kinetics has been missing, and only numerical solutions of the master equations have been reported. For example, the emergence of new modes due to slow promoter kinetics has been shown numerically for the the canonical cell-fate circuit \[34, 35\]. In this
work, we first introduce a formalism to model GRN’s with arbitrary numbers of
genes, based on continuous-time Markov chains, and we analyze the stationary
solution of the associated master equation through a systematic application of the
method of singular perturbations \cite{36} in order to study the slow promoter kinetics
limit. We do this by letting the kinetic rate constants of the TF-gene binding and
unbinding reactions approach zero and compute the stationary solution by applying
the method of singular perturbations to the master equation.

Our main result is based on partitioning the state space into weakly-coupled
ergodic classes \cite{12} which, in the limit of slow binding/unbinding, results in the
reduction of the infinite-dimensional Markov chain into a finite-dimensional chain
whose states correspond to ideal “gene states”. In this limit, the stationary distribution
of the network can be expressed as a mixture of Poisson distributions, each corresponding to conditioning of the chain on a certain binding configuration
of gene promoters. The proposed framework enables us to compute the number
of modes, their locations, and their weights in the mixture. Hence, the proposed
framework can be applied to GRNs to predict the different phenotypes that the
network can exhibit with low gene copy numbers and slow promoter kinetics.

In order to illustrate the practical significance of our results, we work out sev-
eral examples, some of which have not been studied before in the literature. As a
first application, we show that, with slow promoter kinetics, a self-regulating gene
can exhibit bimodality even with non-cooperative binding to the promoter site. We
study self-regulation unlike previous works \cite{37} which studied non-cooperative
regulation by an exogenous TF. We then investigate the role of cooperativity. Un-
like deterministic systems, we find that it does not change the number of modes.
Nevertheless, it adds an extra degree of freedom by allowing the network to tune
the relative weights of each mode without changing its location.

As a second application, we revisit the classical toggle switch, under slow gene
binding. It has been reported before that, with fast gene binding, the toggle switch
with single-gene copies can be “bistable” without cooperative binding \cite{38}. We
show that this can also happen with slow gene binding, and that a new mode
having both genes with high concentrations can emerge. We provide a method to
calculate the weight of each mode and suppress the third mode via tuning of the
kinetic rates of the dimerization reactions.

A third application we consider is that of quorum sensing synchronization of
toggle switches. In bacterial populations, quorum sensing has been proposed \cite{39} as
a way for bacterial cells to broadcast their internal states to other cells to facilitate
synchronization. Quorum Sensing communication has been adopted also in as a
tool in synthetic biology \cite{40, 41}. Mathematical analysis of coupled toggle switches
designs usually employs deterministic models \cite{42}. We study a stochastic model
of coupled toggle switches with slow promoter kinetics and compare the resulting
number of modes with deterministic equilibria.

The Reaction Network Structure

In this paper, a gene regulatory network will be formally defined as an intercon-
nection of genes that express proteins constitutively or which have their expression
regulated by one or more transcription factors (TF’s) that bind to their gene pro-
moter. Transcription factors are themselves proteins which appear in the network.
Any additional, external to the network, transcription factors that play a regulatory role will be assumed to be at high concentrations and held approximately constant, so their effects are absorbed into the kinetic rate constants.

The formalism we employ in order to describe GRNs at the elementary level is that of Chemical Reaction Networks (CRNs) \([43]\). A CRN consists of *species* and *reactions*. The set of *species* \(\mathcal{S}\) consists of all reactant and product molecules in the network, while the set of *reactions* \(\mathcal{R}\) describes how reactants transform into products. The network description involves two mathematical elements: the *stoichiometry* and the *kinetics*. Stoichiometry describes the differences in the respective numbers of molecules of reactants and products in each reaction, while kinetics is concerned with the relations the govern the velocity of transformation of reactants into products. We start by defining the species and reactions in the network under consideration.

**Species:** The species in our context consist of promoter (“binding”) configurations for the various genes participating in the network, together with the respective proteins expressed from these genes and some of their multimers. A binding configuration of a gene is its binding configuration is characterized by the possible locations and number of TFs bound to the gene’s promoter at a given time. If the gene is expressed constitutively, then there are two binding configurations, specifies the expression activity state, active or inactive. A multimer is a compound consisting of a protein binding to itself several times. For instance, dimers and tetramers are 2-mers and 3-mers, respectively. If a protein forms an \(n^{th}\)-order multimer then we say that it has a cooperativity index of \(n\).

In our setting, a GRN consists of \(N\) genes and their protein products. For simplicity we assume the following:

**A1** Each gene can have up to **two binding sites**, and each binding site can be occupied by only one TF.

**A2** Each protein has a fixed cooperativity index, i.e, it can not act as a TF with two different cooperativity indices.

**A3** Each gene is present at only a **single copy**.

All the above assumptions can be relaxed. We make these assumptions only in order to simplify the notations and mathematical derivations, and all our examples satisfies them. The supplementary material contains generalizations of the results to arbitrary numbers of binding sites and gene copy numbers.

Consider the \(i^{th}\) gene. The expression rate of a gene is dependent on its current binding configuration. We call the set of all possible such configurations the **binding-site set** \(B_i\). Each member of \(B_i\) corresponds to a binding configuration species \(D_i^j, j \in B_i\). If a gene has just one or no regulatory binding sites, then we let \(B_i = \{0, 1\}\). Hence, the gene binding configuration can be represented by **two species**: the bound species \(D_i^0\) and the unbound species \(D_i^1\). On the other hand, if the gene has **two binding sites** then \(B_i = \{00, 01, 10, 11\}\). The first digit in a member of \(B_i\) species specifies whether the first binding site is occupied, and the second digit specifies the occupancy of the second binding site. Hence, the gene

---

\(^{1}\text{We interpret the elements of the binding set as integers in binary representation.}\)
can be represented by four species $D_{i00}^i, D_{i10}^i, D_{i01}^i, D_{i11}^i$. Note that in general we need to define $2^\kappa$ species for a gene with $\kappa$ binding sites.

The species that denotes the protein produced by the $i^{th}$ gene is $X_i$. A protein’s multimer is denoted by $X_{ic}$. If protein does not form a multimer then $X_{ic} := X_i$.

Therefore, the set of species in the network is $S = \bigcup_i \left( \{D_{ij}^i, j \in B_i\} \cup \{X_i, X_{ic}\} \right)$.

Reactions: A reaction is the process of transforming a set of reactant species into a set of product species. In our context, the reactions consists of TFs binding and unbinding from genes and the respective protein expression (with transcription and translation combined in one step), decay, and $n$-merization. A generic reaction $R_j \in R$ takes the form:

$$R_j : \sum_{i=1}^{n} \alpha_{ij} Z_i \rightarrow \sum_{i=1}^{n} \beta_{ij} Z_i,$$

(1)

where $Z_i \in \mathcal{S}$. The reactions that we are going to consider are limited to at most two reactants. The reverse reaction of $R_j$ is the reaction in which the products and reactants are interchanged. If the network contains both the reaction and its reverse then we use the short-hand notation to denote both of them as

$$R_j : \sum_{i=1}^{n} \alpha_{ij} Z_i \leftrightarrow \sum_{i=1}^{n} \beta_{ij} Z_i.$$

(2)

The stoichiometry of a network can be summarized by a stoichiometry matrix $\Gamma$ which is defined element-wise as follows:

$$[\Gamma]_{ij} = \beta_{ij} - \alpha_{ij}.$$ 

The columns of the stoichiometry matrix $\gamma_1, \ldots, \gamma_{|\mathcal{S}|}$ are known as the stoichiometry vectors. A nonzero nonnegative vector $d$ is called a conservation law for the stoichiometry if $d^T \Gamma = 0$.

For each gene, we define a gene expression block. Each block consists of a set of gene reactions and a set of protein reactions as shown in Figure 1.
If a gene switches between two binding configuration constitutively, i.e. without an explicitly modelled TF, then \( B_i = \{0,1\} \) and the gene block consists of:

\[
D_0^i \xrightarrow{\alpha_i} D_1^i.
\]  

(3)

We refer to \( D_0^i, D_1^i \) as the \textit{inactive} and \textit{active} configurations, respectively.

If the gene has one binding site, then also \( B_i = \{0,1\} \) and the gene block consists of just two reactions:

\[
\text{TF} + D_0^i \xrightarrow{\alpha_i} D_1^i, \quad (4)
\]

where \( D_0^i, D_1^i \) are the \textit{unbound} and the \textit{bound} configurations, respectively. Note that we did not designate a specific species as the active one since it depends on whether the TF is an activator or repressor.

Finally if the gene has two binding sites, then \( B_i = \{00,01,10,11\} \) and the gene block is:

\[
\begin{align*}
\text{TF}_1 + D_{00}^i & \xrightarrow{\alpha_{i1}} D_{10}^i, \\ 
\text{TF}_1 + D_{01}^i & \xrightarrow{\alpha_{i2}} D_{11}^i, \\ 
\text{TF}_2 + D_{00}^i & \xrightarrow{\alpha_{i3}} D_{01}^i, \\ 
\text{TF}_2 + D_{10}^i & \xrightarrow{\alpha_{i4}} D_{11}^i.
\end{align*}
\]

(5-8)

The activity of each binding configuration species is dependent on whether the TFs are activators or repressors. Note that by assumption A2 we do not allow “competitive binding,” i.e. two different TFs can not bind to the same location.

The central dogma of molecular biology states that proteins are created in a two-step process from genes. The first step is transcription, in which the genetic code is copied into mRNA by RNA polymerase. The second step is translation, in which proteins are produced by ribosomes based on mRNA. We assume that RNA polymerase and ribosomes are available in high concentration and that we can lump the transcription and translation in one simplified “production” reaction. The rate of production is dependent on the binding configuration. So for each binding configuration \( D_j^i, j \in B_i \) the production reaction is:

\[
D_j^i \xrightarrow{k_{ij}} D_j^i + X_i,
\]

where the kinetic constant \( k_{ij} \) is a non-negative number. \( k_{ij} = 0 \) means that when the gene is in binding configuration \( D_j^i \) there is no protein production.

The character of a TF is manifested in the rate constants of the production reactions: if protein production from a bound gene state has a higher kinetic constant than the protein production from the unbound state, then the TF is \textit{activating}, and if the reverse holds it is \textit{repressing}.

The expressed proteins decay due to dilution resulting from cell growth of the cell or due to the action of specialised proteins known as proteases. We model decay as a single reaction:

\[
X_i \xrightarrow{k_{-i}} \emptyset.
\]

(9)
The expressed proteins can act as transcription factors. They may combine to form dimers, or higher order multimers. Hence, we model the cooperativity reactions as given in Figure 1.

**Kinetics:** The kinetics of the network quantify to the speed of transformation from reactants into products whenever a reaction occurs. In order to keep track of molecule counts, each species $Z_i \in \mathcal{S}$ is associated with a copy number $z_i \in \mathbb{Z}_{\geq 0}$. To each reaction $R_j$ one associates a propensity function $R_j$. Assuming a homogeneous well-stirred isothermal medium with a fixed volume, the most common model of propensities, which we use is the *Mass-Action Kinetics* which is derived from the principal that the likelihood of two reactant molecules colliding and reacting is proportional to their copy numbers. If $R_j$ has a single reactant species $Z_i$ with stoichiometry coefficient $\alpha_{ij}$, then

$$R_j(z_i) = k_j \frac{z_i(z_i - 1)\ldots(z_i - \alpha_{ij})}{\alpha_{ij}!},$$

where $k_j$ is a kinetic rate constant. Note if $\alpha_{ij} = 1$, then $R_j(z_i) = k_j z_i$.

For a bimolecular reaction we only consider the case of unity stoichiometry coefficients of each reactant species, i.e. the left side of the reaction is of the form $Z_{i_1} + Z_{i_2}$. In this case, the propensity function is:

$$R_j(z_{i_1}, z_{i_2}) = k_j z_{i_1} z_{i_2}.$$

**A gene regulatory network:** Consider a set of $N$ genes, binding sets $\{B_i\}_{i=1}^N$, and kinetic constants $k_j$’s. A *gene expression block* is a set of gene reactions and protein reactions as defined above. Each gene block has as an output which is either the protein or its $n$-mer. The input to each gene expression block is the subset of the set of the outputs of all blocks.

Then, a *gene regulatory network* GRN is an arbitrary interconnection of a gene expression blocks as per the interconnection rule expressed above.

A directional graph can be associated to a GRN as follows. Each vertex correspond to a gene expression block. There is a directed edge from vertex A to vertex B if an output of A is an input to B.

In order to simplify the presentation, we assume the following:

**A4** The graph of the gene expression blocks is connected.

Note that if A4 is violated, our analysis can be applied to each connected component.

**Time-Scale Separation:** As mentioned in the introduction, we assume that the gene reactions are considerably slower than the protein reactions. In order to model this assumption, we write the kinetic rates of gene reactions in the form $\varepsilon k_j$, where $0 < \varepsilon \ll 1$ and assume that all other kinetic rates (for protein production, decay and multi-merization) are $O(1)$.

Events in biological cells usually take place at different time-scales [1], and hence singular perturbation techniques are widely used in deterministic settings in order to reduce models for analysis. On the other hand, model-order reduction by
time-scale separation in stochastic processes has been mainly used in the literature for computational purposes, for example to accelerate the stochastic simulation algorithm [46] or to compute finite-space-projection solutions to the master equation [47]. In this work, we use a singular perturbation approach for the analytical purpose of characterizing the form of the stationary distribution in the regimes of slow gene binding.

In the case of a finite Markov chain, the master equation is a finite-dimensional linear ODE, and reduction methods for linear systems can be applied [36], which imply results for Markov chains [48, 49]. For continuous-time Markov chains on a countable space, as needed when analyzing gene networks, there are difficult and open technical issues since exponential stochastic stability [50] need to be established for the stationary solution in order for the asymptotic expansions such as the ones that we will perform [51].

In this paper, we will not delve into technical issues of stochastic stability, simply assume that these expansions exists and the solutions converge to an unique equilibrium solution.

Dynamics and the Master Equation

The dynamics of the network refers to the manner in which the state evolves in time, where by a state $Z(t) \in \mathbb{Z} \subset \mathbb{Z}_{\geq 0}^{S}$ is the vector of copy numbers of the species of the network at time $t$. Since the collision of molecules is random in nature, the time-evolution of states is described mathematically by a stochastic process. The standard stochastic model for a CRN is that of a continuous Markov chain. Let $Z_t$ denote the state space, and consider a time $t$ and let the state be $Z(t) = z \in Z_t$. Then, the probability that the $j^{th}$ reaction fires in an interval $[t, t+\delta]$ is $R_j(z)\delta + o(\delta)$. If $R_j$ fires, then the states changes from $z$ to $z + \gamma_j$, where $\gamma_j$ is the corresponding stoichiometric vector.

As $Z$ is a stochastic process we are interested in characterizing its long-term qualitative behaviour given by the joint probability distribution $p_z(t) = \text{Pr}[Z(t) = z|Z(0) = z_0]$ for any given initial condition $z_0$. The time-evolution of the probability density function can be shown [45] to be given by a system of linear ordinary differential equations known as the forward Kolmogorov equation or the Chemical Master Equation, given by:

$$\dot{p}_z(t) = \sum_{j=1}^{\mathcal{A}} R_j(z - \gamma_j)p_{z-\gamma_j}(t) - R_j(z)p_z(t).$$

(10)

We split the stochastic process $Z(t)$ into two subprocesses: the gene process $D(t)$ and the protein process $X(t)$, as explained below.

For the $i^{th}$ gene, for each binding configuration species $D_i^j \in B_i$ let $D_i^j(t) \in \{0,1\}$ denote its occupancy, i.e. if $D_i^j(t) = 1$, then at time $t$ the $i^{th}$ gene is in a binding configuration $j \in B_i$. It can be seen from gene reactions structure that the network always has conservation law supported on $\{D_i^j, j \in B_i\}$. This implies that

$$\sum_{j \in B_i} D_i^j(t) = 1,$$
which reflects the obvious physical constraint that the gene can be in one binding configuration at a given time.

This conservation law enables us to introduce an equivalent reduced representation. For each gene we define one process $D_i$ such that $D_i(t) \in B_i$. $D_i(t) = j$ if and only if $D^j_i(t) = 1$.

Collecting these into a vector define the gene process $D(t) := [D_1(t), ..., D_N(t)]^T$ where $D(t) \in \prod_{i=1}^N B_i$. The $i^{th}$ gene can be represented by $|B_i|$ states, so $L := \prod_{i=1}^N |B_i|$ is be the total number of binding configurations in the GRN. As an abuse of notation, we write also $D(t) \in \{0, ..., L-1\}$ in the sense of the bijection between $\{0, ..., L-1\}$ and $\prod_{i=1}^N B_i$ defined by interpreting $D_1...D_N$ as a binary representation of an integer. Hence, $d \in \{0, ..., L-1\}$ corresponds to $(d_1, ..., d_N) \in B_1 \times \ldots \times B_N$ and we write $d = (d_1, ..., d_N)$.

Since each gene expresses a corresponding protein, we define $X_{ic}(t) \in \mathbb{Z}_{\geq 0}, i = 1, ..., N$ protein processes. If the multimerized version of the $i^{th}$ protein participates in the network as an activator or repressor then we define $X_{ic}(t)$ as the corresponding multimerized protein process. Then we denote $X_i(t) := [X_{i1}(t), X_{ic}(t)]^T$. If no there is no multimerization reaction then we define $X_i(t) := X_{i1}(t)$. Since not all proteins are necessarily multi-merized, then the total number of protein processes is $N \leq M \leq 2N$. Hence, the protein process is $X(t) \in \mathbb{Z}_+^M$ and the state space can be written as $\mathcal{Z} = \mathbb{Z}_+^M \times \prod_{i=1}^N B_i$.

## 2 Results

### Decomposition of The Master Equation

It is crucial to our analysis to represent linear system given by the master equation of the Markov process as an interconnection of weakly coupled linear systems. We present the appropriate notation in this subsection.

Consider the joint probability distribution:

$$p_{x,d}(t) = \text{Pr}[X(t) = x, D(t) = d],$$

where $d \in \{0, ..., L-1\}, x \in \mathbb{Z}_+^M$. Then, we can define for each fixed $d$:

$$p_d(t) := [p_{dx_0}(t), p_{dx_1}(t), \ldots]^T,$$

where $x_0, x_1, ..$ is an indexing of $\mathbb{Z}_+^M$. Note that $p_d(t)$ can be thought of as an infinite vector with respect to the aforementioned indexing. Finally, let

$$p(t) := [p_1(t)^T, ..., p_L(t)^T]^T.$$  \hspace{1cm} (12)

Note that $p(t)$ is a finite concatenation of infinite vectors.

Consider a given GRN. Then, the master equation (10) can be written as an infinite differential equation with an infinite infinitesimal generator matrix $\Lambda$, which can be written succinctly entry-wise as:

$$\lambda_{\tilde{z}\tilde{z}} := \begin{cases} R_j(z) & \text{if } \exists j \text{ such that } \tilde{z} = z - \gamma_j \\ -\sum_{\tilde{z} \neq z} \lambda_{\tilde{z}\tilde{z}} = -\sum_{j=1}^{[\tilde{z}]} R_j(z) & \text{if } \tilde{z} = z \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (13)
The matrix \( \Lambda \) is stochastic which means that it is Metzler and \( \Lambda^T \Lambda = 0 \). A Metzler matrix is a matrix whose off-diagonal elements are non-negative.

Partitioning the probability distribution vector as in (12) and using a fast-slow decomposition we can state the following basic proposition:

**Proposition 1.** Given a GRN. Then, the master equation (10) can be written as

\[
\dot{p}(t) = \Lambda \varepsilon p(t) = (\hat{\Lambda} + \varepsilon \hat{\Lambda}) p(t),
\]

where \( \hat{\Lambda} \) is the fast matrix, \( \hat{\Lambda} \) is the slow matrix and \( \Lambda_1, \ldots, \Lambda_L \) are stochastic matrices and the entries of all matrices are \( O(1) \).

### Conditional Markov Chains

Since the matrices \( \Lambda_1, \ldots, \Lambda_L \) are stochastic, they can be interpreted as infinitesimal generator matrices for the continuous-Markov chain with a constant \( D(t) \). In other words, fixing \( D(t) = d \in \{1, \ldots, L\} \) the dynamics of the network can be described by a master equation:

\[
\dot{p}_d = \Lambda_d p_d,
\]

and the entries of \( p_d \) can be interpreted as a conditional probabilities \( p_{dx} = \Pr[X(t) = x|D(t) = d] \).

This means that at the slow promoter kinetics limit \( D(t) \) stays constant and hence this makes the dynamics of \( p_d \) independent of \( \{p_d\}_{d \neq d'} \).

We show below that each conditional Markov chain has a simple structure. Fixing \( D(t) = d = (d_1, \ldots, d_N) \), the network consists of uncoupled birth-death process with or without dimerization. So for each \( i \), either the protein reaction takes the simple form of a birth-death process

\[
\emptyset \overset{k_{id_i}}{\underset{k_{-i}}{\longrightarrow}} X_i,
\]

or, if there is a multimerization reaction, it takes the form:

\[
\emptyset \overset{k_{id_i}}{\underset{k_{-i}}{\longrightarrow}} X_i, \quad n_i X_i \overset{\beta_i}{\underset{\beta_{-i}}{\longrightarrow}} X_i c.
\]

Note that the stochastic processes \( X_i(t), i = 1, \ldots, N \) conditioned on \( D(t) = d \) are independent of each other. Hence, the conditional joint stationary distribution can be written as a product of stationary distributions and the individual stationary distributions have Poission expressions as the following proposition states:

**Proposition 2.** Fix \( d \in \{1, \ldots, L\} \). Consider (16), then there exists a conditional joint stationary distribution \( \pi_d \) and it is given by

\[
\pi_d(x) = \prod_{i=1}^{N} \pi_{di}(x_i),
\]
where
\[
\pi_{di}(x_i) = \begin{cases} 
\mathbf{P} \left( x_{i1}, x_{i2}; \frac{k_{id_i}}{k_{-i}}, \frac{n_i k_{id_i} \beta_i}{n_i k_{id_i} \beta_i} \right) & \text{if } X_i \text{ is dimerized}, \\
\mathbf{P} \left( x_i; \frac{k_{id_i}}{k_{-i}} \right) & \text{otherwise}
\end{cases}
\] (20)

where \( \mathbf{P}(x; a) := \frac{a^x e^{-a}}{x!} \), \( \mathbf{P}(x_1, x_2; a_1, a_2) := \frac{a_1^{x_1} a_2^{x_2} e^{-a_1-a_2}}{x_1! x_2!} \).

**Remark 1.** If a production rate \( k_{id_i} \) is zero, then the Poisson distribution for \( \pi_{di} \) reduces to a dirac Delta function.

**Remark 2.** The conditional distribution in (19) is a joint distribution in the protein and dimerized protein processes. If we want to compute a marginal distribution for the protein process only, then we average over \( X_2 \) to get a joint Poisson in \( N \) variables. Hence, the formulae (19) can be replaced by:
\[
\pi^{(m)}_{di}(x) = \prod_{i=1}^{N} \mathbf{P} \left( x_i; \frac{k_{id_i}}{k_{-i}} \right).
\] (21)

**Irreducibility**

An important property in the context of Markov chain analysis is that of irreducibility \[12\], and its significance stems from the fact that it is a necessary condition for the existence of a unique positive stationary distribution. Consider the Markov chain \( Z(t) \) defined on \( Z \) with an associated infinitesimal generator \( \Lambda \) as given in (13). Let \( z, w \in Z \). Then, it is said that \( z \) leads to \( w \) if there exist states \( z_0, ..., z_n \in Z \) such that \( \lambda_{z0} \lambda_{z01} ... \lambda_{znw} > 0 \). A set \( U \subseteq Z \) is said to be a communicating class if for every \( z_1, z_2 \in U \), \( z_1 \) leads to \( z_2 \) and \( z_2 \) leads to \( z_1 \). The state space \( Z \) can always be partitioned into a disjoint union of communicating classes \[12\]. The Markov chain is said to be irreducible if the state space \( \subseteq Z \) is a communicating class. A communicating class \( U \) is said to be closed if \( z \in U \), and \( z \) leads to \( w \) implies \( w \in U \). A Markov chain is said to be weakly irreducible if it has a unique closed communicating class \( U \), and for all \( z \in Z \), \( z \) leads to some element \( U \).

We state the following result, under assumption A4:

**Proposition 3.** Consider a gene regulatory network that consists of \( N \) gene expression blocks. Then the associated Markov chain is weakly irreducible.

**Remark 3.** For finite Markov chains, weak irreducibility with appropriate stochastic stability assumptions are sufficient for the existence of a nonnegative unique stationary distribution \[52\], while irreducibility is usually needed for the existence a positive stationary distribution. Note that not all GRNs are irreducible. However, our subsequent results require weak irreducibility only, and investigation of irreducibility is out of the scope of this paper. Nevertheless, we have developed necessary and sufficient graphical conditions for irreducibility \[53\].
Decomposition of The Stationary Distribution

Recall the slow-fast decomposition in (14). Let $\pi^\varepsilon$ be a unique stationary distribution that satisfies $\Lambda^\varepsilon \pi^\varepsilon = 0$, $\pi^\varepsilon > 0$, and $\sum_z \pi^\varepsilon_z = 1$.

Our objective is to characterize the stationary distribution as $\varepsilon \to 0$. Writing $\pi^\varepsilon$ as an asymptotic expansion in terms of $\varepsilon$, we have, to first order

$$\pi^\varepsilon = \pi^{(0)} + \varepsilon \pi^{(1)} + o(\varepsilon).$$

(22)

Our aim is to find $\pi^{(0)}$. Substituting $\pi^\varepsilon$ in (14), and equating the coefficients of the powers of $\varepsilon$ to zero we obtain the following two equations:

$$\begin{bmatrix}
\Lambda_1 \\
\vdots \\
\Lambda_L
\end{bmatrix}
\begin{bmatrix}
\pi^{(0)} \\
\pi^{(1)} + \hat{\Lambda}\pi^{(0)}
\end{bmatrix} = 0$$

(23)

(24)

Eq. (23) implies that $\pi^{(0)} = \ker \hat{\Lambda}$. We next show how to compute $\ker \hat{\Lambda}$.

Recall the conditional Markov chains with the associated infinitesimal generators as in (16). By the assumptions, for each $d \in \{0, \ldots, L-1\}$ there exists a unique $\pi_d$ such that: $\Lambda_d \pi_d = 0$, $\pi_d > 0$, and $\sum_x \pi_{dx} = 1$. Note that $\pi_d$ is the stationary distribution of the Markov chain conditioned on $D(t) = d$.

Defining the extended conditional distributions for $d = 0, \ldots, L-1$ as:

$$\tilde{\pi}_d := \begin{bmatrix}
0^T & \cdots & 0^T & \pi_d^T & 0^T & \cdots & 0^T
\end{bmatrix}^T,$$

(25)

then $\ker \hat{\Lambda} = \text{span}\{\tilde{\pi}_1, \ldots, \tilde{\pi}_L\}$. Hence, we can write:

$$\pi^{(0)} = \sum_{i=0}^{L-1} \lambda_i \tilde{\pi}_i,$$

for some $\lambda_1, \ldots, \lambda_L \geq 0$. We normalize them to satisfy $\sum_{d=0}^{L-1} \lambda_d = 1$.

In order to satisfy (24), we utilize the fact that each $\Lambda_d$ is an infinitesimal generator which satisfies $1^T \Lambda_d = 0$. Hence, we pre-multiply eq. (4) by the vectors: $[1^T \ 0^T \ \cdots \ 0^T]^T$, $[0^T \ 1^T \ \cdots \ 0^T]^T$, $[0^T \ 0^T \ \cdots \ 1^T]^T$ in order to get the following $L$-dimensional linear system:

$$\Lambda_\ast \lambda := \begin{bmatrix}
1^T & 0^T & \cdots & 0^T \\
0^T & 1^T & \cdots & 0^T \\
\vdots & \vdots & \ddots & \vdots \\
0^T & 0^T & \cdots & 1^T
\end{bmatrix} \hat{\Lambda} \begin{bmatrix}
\tilde{\pi}_1 \\
\tilde{\pi}_2 \\
\vdots \\
\tilde{\pi}_L
\end{bmatrix} \begin{bmatrix}
\lambda_1 \\
\lambda_2 \\
\vdots \\
\lambda_L
\end{bmatrix} = 0.$$

(26)

Since $\hat{\Lambda}$ has a codimension equal to one, we need the following equation to find $\lambda_1, \ldots, \lambda_L$ uniquely:

$$\lambda_1 + \ldots + \lambda_L = 1.$$

(27)

Hence, we state the following theorem:
Theorem 4. Consider an GRN with $L$ genes with the master equation (14). Writing (22), let $\tilde{\pi}_1, \ldots, \tilde{\pi}_L$ be the extended conditional stationary distributions of $\Lambda_1, \ldots, \Lambda_L$ as in (25). Then the stationary distribution satisfies:

$$\tilde{\pi} := \lim_{\varepsilon \to 0^+} \pi_\varepsilon = \sum_{d=0}^{L-1} \lambda_d \tilde{\pi}_d,$$

where $\lambda_1, \ldots, \lambda_L$ are solutions of (26), (27).

Remark 4. The result characterizes the stationary solution of (14) which is a joint distribution in $X$ and $D$. However, we are particularly interested in the marginal distribution of protein process $X$, since these distributions are typically experimentally observable. Therefore, we can utilize (25) to write the marginal stationary distribution of $X$ as:

$$\pi(x) := \lim_{t \to \infty, \varepsilon \to 0^+} \Pr[X(t) = x] = \sum_{d=0}^{L-1} \lambda_d \pi_d(x),$$

(28)

where the explicit expressions for $\pi_d, d = 0, \ldots, L - 1$ are given in (19).

Hence, the stationary distribution always takes the form of a mixture of $L$ Poisson distributions. If a mode is defined as a local maximum of a stationary distribution, then this does not necessarily imply that the stationary distribution has $L$ modes since the peak values of two Poisson distributions can be very close to each other. In the remaining of the paper we will call each Poisson distribution in the mixture as a “mode” in the sense that it represents a component in the mixture distribution.

The number of local maxima of a distribution can be found easily given the above expression.

Remark 5. The marginal distribution in (28) is a joint distribution in the protein and multimerized protein processes. Using Remark 2 we can similarly write the marginal distribution of the non-multimerized protein process as:

$$\pi^{(m)}(x) = \sum_{d=0}^{L-1} \lambda_d \pi^{(m)}_d(x) = \sum_{d=0}^{L-1} \lambda_d \prod_{i=1}^{N} P(x_i; \frac{k_{id}}{k_{-i}}).$$

(29)

The Reduced Finite Markov Chain

The matrix $\Lambda_r$ in (26) is an $L \times L$ matrix that can be interpreted as the infinitesimal generator of an $L$-dimensional Markov chain. Since the structure of the GRN and the form of the conditional distribution (19) are known, then we give a more explicit form of Theorem 4 for the class of GRNs we are considering. Write the expression of the matrix $\Lambda_r$ using (25) as follows:

$$\Lambda_r = \begin{bmatrix}
1^T \hat{\Lambda}_{11} \pi_1 & \ldots & 1^T \hat{\Lambda}_{1L} \pi_L \\
\vdots & \ddots & \vdots \\
1^T \hat{\Lambda}_{L1} \pi_1 & \ldots & 1^T \hat{\Lambda}_{LL} \pi_L
\end{bmatrix}.$$

(30)

The $(d', d)$ entry represents the probability of transition from the binding configuration $d'$ to binding configuration $d$, and it can be interpreted as a weighted conditional expectation of $\pi_d$. 

14
Fix a binding configuration $d$. Then, the maximum number of possible reactions, and hence transitions, is $\frac{1}{2} \sum_{i=1}^{N} |B_i|$. Hence, $\Lambda_r$ is a sparse matrix. So instead of computing the infinite matrices and matrix product in (23) we provide an algorithm for computing the nonzero entries in $\Lambda_r$.

This can be achieved by considering all the possible transitions from a binding configuration $d = (d_1, .., d_N)$. The chain can transit from $d$ to $d'$ by a gene reaction modifying a single gene binding configuration. For instance consider $D_{d_i}$. Then for a constitutive or single TF gene binding/unbinding, there can be only one transition starting $D_{d_i}$. This transition is either the forward or reverse reactions in (3) or (4), respectively. For the case of two TFs, there can be two reactions among (5)-(8).

The algorithm can be described as follows:

**Proposition 5.** The matrix $\Lambda_r$ in (26) can be computed via the algorithm below.

- For each $d \in \{0, .., L-1\}$ write $d = (d_1, .., d_N) \in \prod_{i=1}^{N} B_i$. Then:
  - For each $i \in \{1, .., N\}$:
    * List the gene reactions that have $D_{d_i}$ as a reactant. By A3, there can be a maximum of two of them. For each such reaction:
      1. Let $D_{d_i'}$ be the $i^{th}$ binding configuration species of the reaction, and let $d'' = (d_1, .., d', .., d_N)$ be the next binding configuration of the network. Let $\alpha$ be the kinetic constant of the reaction under consideration. If the considered reaction is a binding reaction, then let $X_i, X_{i2}$ denote the TF or the dimerized TF.
      2. Then, the $(d', d)$ entry of $\Lambda_r$ can be written as:

        $$ [\Lambda_r]_{d'd} = \begin{cases} 
        \alpha, & \text{if the reaction is monomolecular} \\
        \alpha \frac{k_{d_i}}{n_i! k_1 \beta_i}, & \text{if the reaction is bimolecular}
        \end{cases} $$

      (31)

    - Set $[\Lambda_r]_{dd} = - \sum_{i' \neq i} 1^T \hat{\Lambda}_{d_i'} d_i \pi_d$. (32)

  - Set the rest of the entries of $\Lambda_r$ to zero.

**Remark 6.** The algorithm above gives an intuitive way to interpret Theorem 4 since it implies that each binding reaction of the form:

$$ TF + D_{d_i}^i \xrightarrow{\alpha} D_{d_i'}^i, $$

gives the rate $\alpha E[TF|D = d]$, and hence it corresponds to a reaction of the form:

$$ D_{d_i}^i \xrightarrow{\alpha E[TF|D = d]} D_{d_i'}^i. $$

(33)

This can be used to draw state diagrams for the reduced system immediately.
Basic Example

We start with the simplest form of network, which is the constitutive gene binding/unbinding model which has been verified as a model for transcriptional bursting \[18\] and has been studied using time-scale separation also by \[29, 53\]. Consider:

\[
D_0 \xrightarrow{\varepsilon \alpha} D_1
\]

\[
D_1 \xrightarrow{k} X + D_1,
\]

\[
X \xrightarrow{k_-} 0.
\]

Using the former framework we identify a single gene block with two states. Using (19), the conditional stationary distributions are:

\[
\pi_0(x) = \delta(x)
\]

\[
\pi_1(x) = \frac{(k_-)^x}{x!} e^{-k/k_-}
\]

The reduced Markov chain is a binary Bernoulli process with a rate of \(\frac{\alpha}{\alpha + \alpha_-}\).

Then the marginal distribution of \(X\) can be written using (28) as:

\[
\pi(x) = \frac{\alpha_-}{\alpha + \alpha_-} \delta(x) + \frac{\alpha}{\alpha + \alpha_-} \frac{1}{x!} \left(\frac{k_-}{k} \right)^x e^{-k/k_-},
\]

which is a bimodal distribution with peaks at 0 and \(k/k_-\). Note that same model with reverse time-scale separation is a Poisson with mean \(\frac{\alpha}{\alpha + \alpha_-} k\), which is the same as the deterministic equilibrium if we used the conservation law \(D_1(t) + D_0(t) = 1\) for the above model. Finally, note that the model with slow gene binding has the same mean with the fast switching model but this stationary distribution differs drastically.

The Role of Cooperativity

A transcription factor is said to be cooperative if it acts only after it forms a dimer or a higher-order \(n\)-mer that binds to the promoter region in the gene \[54\]. In standard deterministic modelling, a cooperative activation replaces the Michaelis-Menten dependence with a Hill dependence which enables the network to have additional multiple equilibria in some kinetic parameter ranges. For example, a non-cooperative self-regulating gene can only be mono-stable while the cooperative counterpart can be multi-stable for some parameters.

Results in previous sections show that the role of cooperativity with slow promoter kinetics are different from a deterministic setting. Recall that Theorem 4 establishes that the stationary distribution is a mixture of \(L\) Poisson processes independent of whether the activations are dimerized, \[21\] show that the location of the modes are just ratios of production to decay and are independent of the dimerization rates. This is in contrast to the case of a deterministic model where the number of stable equilibria is affected by dimerization.

Nevertheless, the multi-merization constants still appear in the weighting coefficients in \[31\]. If a TF binds non-cooperatively, the weighting factor in \[31\] is
the ratio of the rates of production to decay. Hence if we need to make a certain mode more probable we need to change either the location of the mode or association ratio (the ratio of the binding to unbinding kinetic constants). On the other hand, a dimerized TF gives an extra tuning parameter; namely the ratios of the rates of dimerization reaction and its reverse. Hence, a certain mode can be more probable by modifying the dimerization reactions rate constants with changing the location of the peaks or the association ratio. In addition, the weighing formula includes the association ratio raised to the cooperativity index. Hence, the cooperativity index is another tuning parameters that amplifies the weight of the association ratio.

In order to illustrate the above idea, we analyze a self-regulating gene with slow gene binding with and without cooperativity. The non-cooperative model has been studied by deriving a closed form in [28] and by time-scale approximation in [29].

### A Self-Regulating Gene

Consider a non-cooperative self-activating gene:

\[
\begin{align*}
X + D_0 \xrightarrow{\varepsilon \alpha} & D_1 \\
D_0 \xrightarrow{k_0} & \tilde{D} + X, \\
D_1 \xrightarrow{k_1} & D + X, \\
X \xrightarrow{k} & 0.
\end{align*}
\]

(34)

This is a single-gene block with two states. At the limit of slow gene binding, the gene binding/unbinding reaction can be written as in (33) which defines a binary Bernoulli process with the rate \(\alpha k_0/(\alpha k_- + \alpha k_0)\). Using (28) the stationary distribution is a mixture of two Poisson distributions and can be written as:

\[
\pi_{nc}(x) = \frac{\alpha_1}{\alpha_- + \alpha_1} \left(\frac{k}{x!}\right)^x e^{-k_1/k_-} + \frac{\alpha_-}{\alpha_- + \alpha_1} \left(\frac{k}{x!}\right)^x e^{-k_0/k_0},
\]

(35)

where

\[
\rho_1 = \mathbb{E}[X_2|D = 0] = k_0/k_-.
\]

Next, consider the same reaction network but now with cooperativity:

\[
\begin{align*}
X_2 + D_0 \xrightarrow{\varepsilon \alpha} & D_1 \\
D_0 \xrightarrow{k_0} & \tilde{D} + X, \\
D_1 \xrightarrow{k_1} & D + X, \\
X \xrightarrow{k} & 0 \\
2X \xrightarrow{\beta} & X_2.
\end{align*}
\]

(36)
Then the gene process is still a Bernoulli process with the a new rate to be given below. The marginal distribution for $X$ can be written as:

$$\pi_e(x) = \frac{\alpha \rho_2}{\alpha + \alpha \rho_2} \frac{(k_1)^x}{x!} e^{-k_1/k_{-1}} + \frac{\alpha_-}{\alpha + \alpha \rho_2} \frac{(k_2)^x}{x!} e^{-k_2/k_{-1}},$$

where

$$\rho_2 = \mathbb{E}[X_2|D = 0] = \frac{k_2^2 \beta}{2k_0 \beta_\alpha}.$$

Both distributions (35), (37) have modes at $\frac{k_1}{k_{-1}}$ and $\frac{k_0}{k_{-1}}$. The height of the first mode is proportional to $\rho_1, \rho_2$, respectively. The network is activating if $k_{21} > k_{20}$, and repressing otherwise.

Comparing (35) and (37) note that if we want increase the weight of the mode corresponding to the bound state then the mode location needs to be changed in the non-cooperative case, while in (37) the dimerization rates in the factor $\rho_2$ can be used in order to tune the weights freely while keeping the modes and the binding to unbinding kinetic constants ratio unchanged. For instance, we can make the distribution effectively uni-modal with a sufficiently high dimerization ratio.

Special cases:
The model above treats a network with possibly non-zero production rates for both binding configurations. We may also the special cases of pure activator or repressor:

1. **Pure Self-Activation**, i.e. $k_0 = 0$. Then the first peak disappears and we get only one peak at 0 as it forms an absorbing state. This is a manifestation of the Kaizer’s paradox. One way to circumvent this is to allow for a small transcriptional “leak”. This amounts to assigning $k_0 \ll k_1$. Note that in the non-cooperative case the choice of this kinetic rate determines completely the relative weight of the modes as in (35). Cooperativity allows us to tune the weight of the mode corresponding the bound state without changing the location as mentioned before.

2. **Pure Self-Repression**, i.e. $k_1 = 0$. Then we get two modes: one at 0 and the other at $k_0/k_{-1}$.

Comparison with Fast Switching:
In order to demonstrate that slow switching is responsible for the emergence of new modes, consider the non-cooperative network with fast switching:

$$X + D_0 \xrightarrow{\alpha/\varepsilon} D_1$$

$$D_0 \xrightarrow{k_{-1}} \bar{D} + X,$$

$$D_1 \xrightarrow{k_{-1}} D + X,$$

$$X \xrightarrow{k_{-1}} 0.$$

We state the following proposition:
Table 1: Number of stable equilibria/modes. S stands for stationary and QS stands for quasi-stationary.

### Proposition 6.
The marginal stationary distribution of the network above is given by:

\[
\Pr[X = m] = \frac{\alpha_m w_m + \alpha_{0} w_{m+1}}{\alpha_m + \alpha_-},
\]

where \( w_m \) satisfies the following recurrence relation:

\[
w_{m+1} = \frac{\alpha(k_1 m + \alpha_- k_0)(m + \alpha_- + 1)}{k_-(m + 1)(\alpha m + \alpha_-)^2} w_m, \quad m \geq 0.
\]

Since the ratio \( w_{m+1}/w_m \) is a ratio of two polynomials and the denominator has a higher degree than \( w_m \) is uni-modal, while slow TF-gene binding/unbinding was shown to give a bi-modal distribution.

### Comparison with the deterministic model:
Comparing this with the equilibria of the cooperative deterministic model, there is no correlation with the number of peaks for the probability distribution, since there always exists two modes if \( k_{20} \neq 0 \), while the deterministic system can have only one stable equilibrium. For the case pure self-activation, there exists only one mode for the probability distribution at 0, while the deterministic system can admit three equilibria.

Table 1 compares the number of stable equilibria in the deterministic model with the number of modes in the stochastic model in the case of a single gene copy.

### The Toggle Switch
So far we have considered examples of GRNs consisting of a single gene. A toggle switch is one of the most basic GRNs that exhibit deterministic multi-stability. It has two stable steady states and it can switch between them with an external input or via noise. The basic design is a pair of two mutually repressing genes. The ideal behavior is that only one gene is “on” at any moment in time. We now study the network with the slow gene binding. Consider the following network with

|                  | Non-Cooperative | Cooperative |
|------------------|----------------|-------------|
|                  | Leaky Non-Leaky | Leaky Non-Leaky |
| S QS*            | 2 1 1           | 2 1 1       |
| Slow gene binding | (at 0)         | (at 0)      |
|                  | 1 1 1           | 1* 1 1      |
| Fast gene binding | (at 0)         | (at 0)      |
| Deterministic    | 1 1 1-2         | 1-2 1-2     |

*Based on Monte-Carlo simulation. Theoretical analysis is an undergoing task.
cooperativity indices \( n, m \):

\[
Y_m + D_0^X \xrightarrow{\alpha_1 \alpha_2} D_1^X
\]

\[
D_0 \xrightarrow{k_{10}} D_0^X + X,
\]

\[
X \xrightarrow{k_{-1}} 0
\]

\[
X_n + D_0^Y \xrightarrow{\alpha_2 \alpha_2} D_1^Y
\]

\[
D_0^Y \xrightarrow{k_{20}} D_0^Y + Y,
\]

\[
Y \xrightarrow{k_{-2}} 0
\]

\[
nX \xrightarrow{\beta_1 \beta_1} X_n
\]

\[
mX \xrightarrow{\beta_2 \beta_2} Y_m.
\]

For the case \( n, m = 1 \) we choose \( \beta_1 = \beta_{-1}, \beta_2 = \beta_{-2} = 1 \).

The network has four binding configurations \((D_X, D_Y) \in \{(0,0), (0,1), (1,0), (1,1)\}\). Using Theorem 4 we expect to have a stationary distribution with four modes \((k_{10}/k_{-1}, k_{20}/k_{-2}), (k_{10}/k_{-1}, 0), (0, k_{20}/k_{-2}), (0, 0)\). The reduced-order Markov chain infinitesimal generator is:

\[
\Lambda_r = \begin{bmatrix}
- \alpha_1 \rho_1 - \alpha_2 \rho_1 & \alpha_2 & \alpha_1 & 0 \\
\alpha_2 \rho_1 & - \alpha_2 & 0 & \alpha_1 \\
\alpha_1 \rho_2 & - \alpha_2 & - \alpha_1 & \alpha_2 \\
0 & 0 & 0 & - \alpha_2 - \alpha_1
\end{bmatrix}, \tag{39}
\]

where

\[
\rho_1 = \left( \frac{k_{10}}{k_{-1}} \right)^n \frac{\beta_1}{n! \beta_{-1}}, \quad \rho_2 = \left( \frac{k_{20}}{k_{-2}} \right)^m \frac{\beta_2}{m! \beta_{-2}}. \tag{40}
\]

We can notice immediately that the transition rates towards the binding configuration \((1,1)\) are zero, which implies that its weight in the stationary distribution is zero. Hence, we have three modes only. The weighting vector of the stationary distribution can be found as the principal eigenvector of \(\Lambda_r\). Hence the marginal stationary distribution for \(X, Y\) is:

\[
\pi(x, y) = \frac{1}{\alpha_1 \rho_1 + \alpha_2 \rho_1 + 1} \left( P(y; k_{20}/k_{-2}) P(x; k_{10}/k_{-1}) + \frac{\alpha_1}{\alpha_{-1}} \rho_2 P(y; k_{20}/k_{-2}) \delta(x) + \frac{\alpha_2}{\alpha_{-2}} \rho_1 P(x; k_{10}/k_{-1}) \delta(y) \right).
\]

The desired switching behaviour can be achieved by minimizing the weight of the first mode at \( (h_{10}/h_{-1}, h_{20}/h_{-2}) \). If we fix \( \alpha_1/\alpha_{-1}, \alpha_2/\alpha_{-2} \), then this can be satisfied by tuning \( n, m, \beta_{\pm 1}, \beta_{\pm 2} \) to maximize \( \rho_1, \rho_2 \) in (40). Choosing higher cooperativity indices, subject to \( n < k_{10}/k_{-1}, m < k_{20}/k_{-2} \), achieves this. For instance, a standard design by Gardner et al. [55] uses \( n = 2, m = 3 \). Figure 2 depicts the effect of cooperativity on getting the desired behaviour with the same association and production ratios, and dimerization ratios equal to one. The weights of the modes for the non-cooperative case are 0.38462, 0.30769, 0.30769, 0, while for the
cooperative network the weights are 0.015385, 0.49231, 0.49231. Notice that cooperativity minimized the weight of the mode corresponding to both proteins at high concentrations.

The toggle switch has three modes regardless of the cooperativity index. This is unlike the deterministic model where only one positive stable state is realizable with non-cooperative binding. However, unlike a self-regulating gene, the toggle switch with fast switching can admit three modes in some parameter ranges, although the third stable state is the one with both low concentrations \([56]\).

Quorum Sensing synchronization of toggle switches

We consider an interconnection of \(N\) identical toggle switches

\[
\begin{align*}
Y_{ic} + D_0^{zi} & \xrightarrow{\varepsilon \alpha_x} D_1^{zi} & X_{ic} + D_0^{yi} & \xrightarrow{\varepsilon \alpha_y} D_1^{yi} \\
D_0^{zi} & \xrightarrow{k_x} X_i + D_0^{zi}, & D_0^{yi} & \xrightarrow{k_y} Y_i + D_0^{yi}, \\
X_i & \xrightarrow{k_x} 0, & Y_i & \xrightarrow{k_y} 0, \\
nX_i & \xrightarrow{\beta_x} X_{ic}, & nY_i & \xrightarrow{\beta_y} Y_{ic},
\end{align*}
\]

where \(i = 1, \ldots, N\).

The mechanism for quorum sensing is the diffusion of the protein molecules \(X_i, Y_i\) between cells. We model this as reversible reactions with a diffusion coefficient \(D\):

\[
X_i \xrightarrow{D} X_j, Y_i \xrightarrow{D} Y_j, \quad i \neq j, i, j = 1, \ldots, N. \tag{41}
\]

For a deterministic model, there exists a parameter range for which all toggle
switches will synchronize into bistability for sufficiently high diffusion coefficient \[42\].

Our aim is to analyze the stochastic model at the limit of slow promotor kinetics and compare it with the deterministic model. The network has \(4^N\) binding configurations. Consider a binding configuration

\[d = (d^X, d^Y) := (d^X_1, ..., d^X_N, d^Y_1, ..., d^Y_N) \in \{0, ..., 4^{N-1}\},\]

where \(d^X_i, d^Y_j \in \{0, 1\}\). Then the conditional GRN with \(D(t) = d\) can be written as:

\[
\begin{align*}
\emptyset & \xrightleftharpoons[k_{-x}]{d^X_i \kappa_x} X_i, & \emptyset & \xrightleftharpoons[k_{-y}]{d^Y_i \kappa_y} X_i, \\
nX_i & \xrightleftharpoons[\beta_x]{\beta_{-x}} X_{ic}, & nY_i & \xrightleftharpoons[\beta_y]{\beta_{-y}} Y_{ic}, \\
X_i & \xrightleftharpoons[D]{D} X_j, Y_i & \xrightleftharpoons[D]{D} Y_j, & i, j = 1, .., N.
\end{align*}
\]

Note this conditional network is not in the form of conditional GRNs that arise from the class of GRNs defined previously as in Figure 1. Nevertheless, it can be observed that it is a reversible zero-deficiency network. We show in the SI that our results can be generalized to networks that admit weakly reversible deficiency zero conditional GRNs.

The set of modes can be listed as follows:

\[
\left\{ \left( \frac{k_x}{k_{-x}} \omega^{X}_{d_1}, \ldots, \frac{k_x}{k_{-x}} \omega^{X}_{d_N}, \frac{k_y}{k_{-y}} \omega^{Y}_{d_1}, \ldots, \frac{k_y}{k_{-y}} \omega^{Y}_{d_N} \right) : d = (d^X, d^Y) \in \{0, ..., 4^{N-1}\} \right\},
\]

where

\[
\omega^{X}_{d_i} = \frac{\sum_{i=1}^{N} \bar{d}^X_i \omega^{X}_{d_i} + \bar{d}^X \kappa_{-x} / D}{N + \kappa_{-x} / D}, \quad i = 1, .., N,
\]

(43)

where \(\bar{d}^X_i = 1 - d^X_i\). \(\omega^{X}_{d_i}\) is defined similarly.

Hence, the stationary distribution is a mixture distribution of \(4^N\) Poisson distributions. The weights can be found by finding the principal eigenvector of for a Markov chain described as follows. As before, the Markov states are the binding configurations \(d = (d^X, d^Y) \in \{0, ..., 4^{N-1}\}\). Assume that \(D(t) = d\). Then, the state transitions are given by the following reactions:

\[
D_0^{x_i} \xrightleftharpoons[\alpha_x \rho_{d_i}^{X}]{\alpha_{-x}} D_1^{x_i}, D_0^{y_i} \xrightleftharpoons[\alpha_y \rho_{d_i}^{Y}]{\alpha_{-y}} D_1^{y_i},
\]

(44)

where

\[
\rho_{d_i}^{X} = \mathbb{E}[X_{ic}|D(t) = d] = \frac{1}{n!} \frac{\beta_x}{\beta_{-x}} \left( \frac{k_x}{k_{-x}} \omega^{X}_{d_i} \right)^n,
\]

(45)

and \(\rho_{d_i}^{Y}\) is defined analogously.

It can be shown that the origin has zero weight, hence the network has \(4^N - 1\) modes.

The dynamics of the network seem overly complicated. Nevertheless, the picture simplifies considerably with high diffusion coefficient \(D\). Note that from
that as $D \to \infty$, $X_1, \ldots, X_N$ will synchronize. We define synchronization in the sense that the marginal joint distribution of $X_1, \ldots, X_N$ is symmetric with respect to all permutations of the random variables. This implies that the marginal distributions $p_X, i = 1, \ldots, N$ are identical. Hence, for sufficiently large $D$, i.e. $D \gg \max\{k_x, k_y\}$, the probability mass is concentrated around the region for which $X_1, \ldots, X_N$ are close to each other.

Consequently, for large $D$ we can replace the population of toggle switches with a single toggle switch with the synchronized protein processes $X(t), Y(t)$. Next, we describe the stationary distribution of $X(t), Y(t)$.

From [13] it can be seen that $\omega_{ij}^X$ does not depend on $d_i$ for large $D$. Instead it depends only on $\sum_{i=1}^N d_i^N$, which is the total number of unbound promoter sites in the genes producing $X_1, \ldots, X_N$. The same holds for $\omega_{ij}^Y$. Hence, the number of modes will drop from $4^N - 1$ to $(N + 1)^2 - 1$. Hence the joint distribution of $X, Y$ is a mixture of Poisson distributions with following modes:

$$\left\{ \left( \frac{ik_x}{Nk_{-x}}, \frac{jk_y}{Nk_{-y}} \right): i, j = 0, \ldots, N, (i, j) \neq (0, 0) \right\}.$$

Note that similar to the single toggle switch, there are modes which have both $X, Y$ with non-zero copy number. In the other hand, there are many additional modes. Furthermore, similar to the single toggle switch, the undesired modes can be suppressed by tuning the multi-merization ratio. It can be shown that as the multi-merization ratio $\beta_x/\beta_{-x}, \beta_y/\beta_{-y} \to \infty$, the weights of modes in the interior of the positive orthant $\mathbb{R}^2_+$ approach zero.

In conclusion, for sufficiently high $D$ and sufficiently high multimerization ratio the population behaves as a multimodal switch with $2N$ modes which are:

$$\left\{ \left( \frac{ik_x}{Nk_{-x}}, 0 \right), \left( 0, \frac{ik_y}{Nk_{-y}} \right): i = 1, \ldots, N \right\}.$$  

**Numerical Example:**

In order to illustrate the previous results, consider a pair of toggle switches with $k_x = k_y = 150, k_{-x} = k_{-y} = 1, \beta_x = \beta_y = 1, \beta_{-x} = \beta_{-y} = 1, \alpha_x = \alpha_y = 0.3\varepsilon, \alpha_{-x} = \alpha_{-y} = \varepsilon$ and $n = 2$.

We need to choose $D$ such that the protein processes synchronize within one copy number. In other words, we require the maximum distance between the modes in [12] to be less than 1, which can be written as:

$$\max_{d \in \{0, \ldots, 4^{n-1}\}} \max_{i,j=1,\ldots,N} \left\{ \left| \frac{k_x}{k_{-x}} \omega_{di}^X - \frac{k_x}{k_{-x}} \omega_{dj}^X \right|, \left| \frac{k_y}{k_{-y}} \omega_{di}^Y - \frac{k_y}{k_{-y}} \omega_{dj}^Y \right| \right\} < 1$$

The minimal $D$ that satisfies this condition is $D = 75$. The stationary distribution is depicted in Figure 3. Although the network has 8 modes, the four which are the interior are suppressed. Comparing with the deterministic model, it bifurcates into synchronization with $D < 0.1$. The stable equilibria of the network are:

$$(X_1, Y_1, X_2, Y_2) \in\{(0.02, 149.98, 0.02, 149.98), (0.02, 149.98, 0.02, 149.98)\}.$$ 

Note that the stochastic model with slow promoter kinetics adds two additional modes at $(0, 75), (75, 0)$. This can be interpreted in the following manner. In
the stochastic model, the protein processes synchronize while the gene binding configurations do not. The high states (150, 0), (0, 150) correspond to the case when all the binding sites are empty. In the case when one binding site is empty, the first gene is producing while the second is not. Due to diffusion, both protein processes equilibrate to a distribution with a mode which is half of the mode corresponding to both being unbound.

3 Methods

In this section we include proofs of the results.

Proof of Proposition [1]
By the time-scale separation assumption, the gene reactions are slow and the pro-
tein reactions are fast. Then (46) can be written as:

$$\lambda_{z\bar{z}} := \begin{cases} 
\frac{1}{\varepsilon} R_j^{(f)}(z) + R_j^{(s)}(z) & \text{if } \exists j \text{ such that } \bar{z} = z - \gamma_j \\
- \sum_{\bar{x} \neq x} \lambda_{z\bar{x}} = - \sum_{j=1}^{n+1} \frac{1}{\varepsilon} R_j^{(f)}(z) + R_j^{(s)}(z) & \text{if } \bar{z} = z \\
0 & \text{otherwise}
\end{cases}$$

(46)

where (f), (s) denotes fast and slow, respectively.

Hence, the summation in (10) can decomposed into two terms. This implies that the system matrix can be written as a sum of a fast matrix $\tilde{\Lambda}$ and a slow matrix $\hat{\Lambda}$ as in (14).

We now show that (15) holds. Assume $\exists j$ such that $\bar{z} = z - \gamma_j$. Then we need to show that $\lambda_{z\bar{z}} = R_j^{(s)}(z)$ if $z = (x, d)$, $\bar{z} = (\bar{x}, \bar{d})$ and $\bar{d} \neq d$. In other words, there is no fast reaction component. As can be seen in Figure 1, the previous claim follows from the observation that the transitions due to protein reactions do not change $d$.

**Proof of Proposition 2**

As mentioned before, the joint stationary distribution the product of the marginal distributions since the underlying conditional stochastic processes are independent. If $X_i$ does not form a multimer then it is known the stationary distribution of the reaction network (17) is Poisson with mean $k_{id}/k_{-i}$ as in (19).

Assume that $X_i$ forms a multimer. In order to simplify notations consider $\emptyset \xrightarrow{k} X_i \xrightarrow{\beta} nX_j \xrightarrow{\gamma} X_n$. Let $x_1, x_2$ denote the molecular counts of $X_j, X_n$. Then, the master equation is

$$\dot{p}_{x_1, x_2} = (kp_{x_1-1, x_2} - k_- x_1 p_{x_1, x_2}) + (k_- (x_1 + 1) p_{x_1+1, x_2} - kp_{x_1, x_2})$$

$$+ \left( \beta_- x_2 (x_2 + 1) p_{x_1-2, x_2+1} - \frac{1}{n!} \beta \prod_{k=0}^{n-1} (x_1 - k) p_{x_1, x_2} \right)$$

$$+ \left( \frac{1}{n!} \beta \prod_{k=1}^{n} (x_1 + k) p_{x_1+n, x_2-1} - \beta_- x_2 p_{x_1, x_2} \right).$$

In order to solve this equation, we assume detailed balance and solve the recurrence equation, and then verify that it satisfies the above equation. This procedure gives the joint Poisson distribution in (19).

**Proof of Proposition 3**

Consider the state $0 \in \mathbb{Z}$. We first show that for all $z \in \mathbb{Z}$, $z$ leads to 0. This follows from the fact that all reverse reactions are monomolecular. So let $z = (x_1, d_1, ..., x_N, d_N)$. We list the set of reactions, i.e transitions, that will lead to 0. Consider $d_i \neq 0$, if $d_i = 1$ then we apply either the reaction (3) or (4). If $d_i = 10$, then we apply (5) and if $d_i = 01$ we apply (6). If $d_i = 11$, then we apply (5) and (5). Hence, $z$ leads to $z' = (x_1, 0, x_2, 0, ..., x_N, 0)$. Similarly, we can apply the decay reactions (9) and the reverse dimerization until $z'' = 0$.

Now we show that there exists a closed communicating class. If 0 does not lead to any state then \{0\} is a closed communicating class. Otherwise, let $U$ be
the smallest communicating class containing 0. Note that it is closed since if there exists \( z \in U \) that leads \( w \), then \( w \) leads \( z \) by using a reverse reaction.

In order to show that \( U \) is unique, assume that there exists another closed communicating class \( U' \). But this contradicts with the fact that all \( z \in U' \) lead to \( 0 \in U \).

Finally, we have shown that for all \( z \in Z \), \( z \) leads to 0. Hence \( z \) leads to \( U \). ■

Proof of Proposition 5

Recall that in (14), the matrix \( \hat{\Lambda} \) represents the slow matrix, which corresponds to the gene binding reactions. Hence, \( \hat{\Lambda}_{d'd} \) represents the matrix corresponding to the transition between states of the form \((x, d)\) and \((x, d')\). Assume that \( D(t) = d = (d_1, \ldots, d_N) \). Consider the \( i \)th block, then it has one or two reactions that can fire. Specifically, there are \( \frac{1}{2} |B_i| \) gene reactions that can fire. Assume that such a reaction is in one of the forms:

\[
D_{d_i}^i \xrightarrow{\alpha} D_{d_i'}^i, \quad D_{d_i}^i \xrightarrow{\alpha} D_{d_i'}^i + X_i,
\]

where \( X_i \) is a TF for that block. Then,

\[
[\Lambda_r]_{d'd} = 1^T \Lambda_{d'd} \pi_d = \alpha 1 \pi_d = \alpha.
\]

Now consider that the reaction is of the form:

\[
X_i + D_{d_i}^i \xrightarrow{\alpha} D_{d_i'}^i,
\]

then,

\[
[\Lambda_r]_{d'd} = 1^T \Lambda_{d'd} \pi_d = \alpha \sum_{x_i=0}^{\infty} \sum_{x_j \neq \tilde{i}} x_i \pi_{d_i} \sum_{x_j \neq \tilde{i}} x_j \pi_{d_j} = \alpha \mathbb{E}[X_{\tilde{i}}(t)|D(t) = d] = \alpha \frac{k_{m\tilde{i}}^{n\tilde{i}} \beta_{\tilde{i}}}{n! m! \prod_{j \neq \tilde{i}} k_{-jd_j}^{n_j} \beta_{-j}},
\]

where \( -\tilde{i} \) denotes the all the remaining indices excluding \( \tilde{i} \). The last equality holds since by evaluating the mean value of the Poisson distribution in (20).

Finally, (32) holds since \( 1^T \Lambda_r = 0 \), which follows from \( 1^T \hat{\Lambda} = 0 \). ■

Proof of Proposition 6

A decomposition dual to (14) can be written and it can be noted that the fast matrix is block-diagonal with respect to \( X + D_1 \) which is the slow variable, while the fast variable is \( D \).

Writing the stationary distribution as asymptotic expansion in terms of \( \varepsilon \) and take the limit as \( \varepsilon \) goes to zero we can find the distribution for the slow variable as follows:

\[
\text{Pr}[X + D = m] = w_m,
\]

where \( w_m \) satisfies the following recurrence relation:

\[
w_{m+1} = \frac{\alpha (k_1 m + \alpha_- k_0)(m + \alpha_- + 1)}{k_-(m+1)(\alpha m + \alpha_-)^2} w_m, \quad m \geq 0.
\]
The joint distribution can be given as:

\[
\begin{align*}
\Pr[X + D = m, D = 0] &= w_m \frac{\alpha - \alpha m}{am + \alpha}, \\
\Pr[X + D = m, D = 1] &= w_m \frac{\alpha m}{am + \alpha}.
\end{align*}
\]

Hence we can compute the marginal density of \(X\) as follows:

\[
\Pr[X = m] = \Pr[X + D = m, D = 0] + \Pr[X + D = m + 1, D = 1] = \frac{\alpha - w_m + \alpha m w_{m+1}}{am + \alpha}.
\]

4 Discussion

There is as yet no general theoretical framework to characterize when stochastic and deterministic models of the same network lead to different steady-state behaviors. Often in this context, an appeal is made to a fundamental mathematical result relating stochastic and deterministic modeling of biochemical networks due to Kurtz [57]. Kurtz’s Theorem asserts that sample paths of the stochastic model are close, with high probability, to the solution of the associated deterministic model, and the approximation improves as the reaction volume increases. However, the theorem only makes this assertion on finite time intervals and therefore it does not make any predictions regarding steady state behavior. Indeed, even with arbitrarily high reaction volumes, the steady-state behaviors of deterministic and stochastic models of the same network can be drastically different, a phenomenon known as the “Keizer’s paradox” [58].

References

[1] Uri Alon. An introduction to systems biology: design principles of biological circuits. CRC press, 2006.

[2] E Davidson. Gene regulatory networks special feature: Gene regulatory networks. 102:4935–4935, 2005.

[3] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology, 9(10):770–780, 2008.

[4] Domitilla Del Vecchio and Richard M Murray. Biomolecular feedback systems. Princeton University Press, 2015.

[5] Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. Science, 336(6078):183–187, 2012.

[6] Joseph X Zhou and Sui Huang. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. Trends in Genetics, 27(2):55–62, 2011.
[7] Adam Arkin, John Ross, and Harley H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998.

[8] Michail Stamatakis and Nikos V Mantzaris. Comparison of deterministic and stochastic models of the lac operon genetic network. *Biophysical journal*, 96(3):887–906, 2009.

[9] Thomas B Kepler and Timothy C Elston. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical journal*, 81(6):3116–3136, 2001.

[10] Mads Kærn, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, 2005.

[11] Sayuri K Hahl and Andreas Kremling. A comparison of deterministic and stochastic modeling approaches for biochemical reaction systems: On fixed points, means, and modes. *Frontiers in Genetics*, 7, 2016.

[12] James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

[13] Crispin W Gardiner et al. *Handbook of stochastic methods*, volume 3. Springer Berlin, 1985.

[14] Avigdor Eldar and Michael B Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, 2010.

[15] Orsolya Symmons and Arjun Raj. What’s luck got to do with it: Single cells, multiple fates, and biological nondeterminism. *Molecular cell*, 62(5):788–802, 2016.

[16] Paul C Bressloff. Stochastic switching in biology: from genotype to phenotype. *Journal of Physics A: Mathematical and Theoretical*, 50(13):133001, 2017.

[17] Jason R Pirone and Timothy C Elston. Fluctuations in transcription factor binding can explain the graded and binary responses observed in inducible gene expression. *Journal of Theoretical Biology*, 226(1):111–121, 2004.

[18] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS Biol*, 4(10):e309, 2006.

[19] Rajesh Karmakar and Indrani Bose. Graded and binary responses in stochastic gene expression. *Physical biology*, 1(4):197, 2004.

[20] Vahid Shahrezaei and Peter S Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–17261, 2008.

[21] A Paldi. Stochastic gene expression during cell differentiation: order from disorder? *Cellular and molecular life sciences*, 60(9):1775–1778, 2003.
[22] Kathryn Miller-Jensen, Siddharth S Dey, David V Schaffer, and Adam P Arkin. Varying virulence: epigenetic control of expression noise and disease processes. *Trends in biotechnology*, 29(10):517–525, 2011.

[23] Ty C Voss and Gordon L Hager. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics*, 15(2):69–81, 2014.

[24] Lei Yuan, Gary C Chan, David Beeler, Lauren Janes, Katherine C Spokes, Harita Dharaneeswaran, Anahita Mojiri, William J Adams, Tracey Scuito, Guillermo García-Cardeña, et al. A role of stochastic phenotype switching in generating mosaic endothelial cell heterogeneity. *Nature communications*, 7, 2016.

[25] Tsz-Leung To and Narendra Maheshri. Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science*, 327(5969):1142–1145, 2010.

[26] Luca Mariani, Edda G Schulz, Maria H Lexberg, Caroline Helmstetter, Andreas Radbruch, Max Löhnig, and Thomas Höfer. Short-term memory in gene induction reveals the regulatory principle behind stochastic il-4 expression. *Molecular systems biology*, 6(1):359, 2010.

[27] W Ebeling and L Schimansky-Geier. Stochastic dynamics of a bistable reaction system. *Physica A: Statistical Mechanics and its Applications*, 98(3):587–600, 1979.

[28] JEM Hornos, D Schultz, GCP Innocentini, JAMW Wang, AM Walczak, JN Onuchic, and PG Wolynes. Self-regulating gene: an exact solution. *Physical Review E*, 72(5):051907, 2005.

[29] Hong Qian, Pei-Zhe Shi, and Jianhua Xing. Stochastic bifurcation, slow fluctuations, and bistability as an origin of biochemical complexity. *Physical Chemistry Chemical Physics*, 11(24):4861–4870, 2009.

[30] Srividya Iyer-Biswas, F Hayot, and C Jayaprakash. Stochasticity of gene products from transcriptional pulsing. *Physical Review E*, 79(3):031911, 2009.

[31] Philipp Thomas, Nikola Popović, and Ramon Grima. Phenotypic switching in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 111(19):6994–6999, 2014.

[32] Davit A Potoyan and Peter G Wolynes. Dichotomous noise models of gene switches. *The Journal of chemical physics*, 143(19):195101, 2015.

[33] Peijiang Liu, Zhanjiang Yuan, Lifang Huang, and Tianshou Zhou. Roles of factorial noise in inducing bimodal gene expression. *Physical Review E*, 91(6):062706, 2015.

[34] H Feng and J Wang. A new mechanism of stem cell differentiation through slow binding/unbinding of regulators to genes. *Scientific reports*, 2(2):550, 2012.
[35] Cong Chen and Jin Wang. A physical mechanism of cancer heterogeneity. *Scientific reports*, 6, 2016.

[36] Stephen L Campbell and Nicholas J Rose. Singular perturbation of autonomous linear systems. *SIAM Journal on Mathematical Analysis*, 10(3): 542–551, 1979.

[37] Anna Ochab-Marcinek and Marcin Tabaka. Bimodal gene expression in non-cooperative regulatory systems. *Proceedings of the National Academy of Sciences*, 107(51):22096–22101, 2010.

[38] Azi Lipshtat, Adiel Loinger, Nathalie Q Balaban, and Ofer Biham. Genetic toggle switch without cooperative binding. *Physical review letters*, 96(18): 188101, 2006.

[39] Melissa B Miller and Bonnie L Bassler. Quorum sensing in bacteria. *Annual Reviews in Microbiology*, 55(1):165–199, 2001.

[40] Jordi Garcia-Ojalvo, Michael B Elowitz, and Steven H Strogatz. Modeling a synthetic multicellular clock: repressilators coupled by quorum sensing. *Proceedings of the National Academy of Sciences of the United States of America*, 101(30):10955–10960, 2004.

[41] Hideki Kobayashi, Mads Kærn, Michihiro Araki, Kristy Chung, Timothy S Gardner, Charles R Cantor, and James J Collins. Programmable cells: interfacing natural and engineered gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8414–8419, 2004.

[42] Evgeni V Nikolaev and Eduardo D Sontag. Quorum-sensing synchronization of synthetic toggle switches: A design based on monotone dynamical systems theory. *PLoS Comput Biol*, 12(4):e1004881, 2016.

[43] Pétier Érdi and János Tóth. *Mathematical models of chemical reactions: theory and applications of deterministic and stochastic models*. Manchester University Press, 1989.

[44] Harvey Lodish, Arnold Berk, C. A. Kaider, M. Krieger, A. Bretscher, H. Ploegh, A. Amon, and K. C. Martin. *Molecular cell biology*. 2016.

[45] David F Anderson and Thomas G Kurtz. *Stochastic analysis of biochemical systems*. Springer, 2015.

[46] Christopher V Rao and Adam P Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the gillespie algorithm. *The Journal of chemical physics*, 118(11):4999–5010, 2003.

[47] Slaven Peleš, Brian Munsky, and Mustafa Khammash. Reduction and solution of the chemical master equation using time scale separation and finite state projection. *The Journal of chemical physics*, 125(20):204104, 2006.
[48] R Phillips and P Kokotovic. A singular perturbation approach to modeling and control of markov chains. *IEEE Transactions on Automatic Control*, 26 (5):1087–1094, 1981.

[49] RZ Khasminskii, G Yin, and Q Zhang. Constructing asymptotic series for probability distributions of markov chains with weak and strong interactions. *Quarterly of Applied Mathematics*, 55(1):177–200, 1997.

[50] Sean P Meyn and Richard L Tweedie. Stability of markovian processes iii: Foster-lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, pages 518–548, 1993.

[51] Eitan Altman, Konstantin E Avrachenkov, and Rudesindo Núñez-Queija. Perturbation analysis for denumerable markov chains with application to queueing models. *Advances in Applied Probability*, pages 839–853, 2004.

[52] D. Del Vecchio M. Ali Al-Radhawi, E.D Sontag. Irreducibility of gene regulatory networks. *To Appear*, 2017.

[53] Andrew Duncan, Shuohao Liao, Tomáš Vejchodský, Radek Erban, and Ramon Grima. Noise-induced multistability in chemical systems: Discrete versus continuum modeling. *Physical Review E*, 91(4):042111, 2015.

[54] James E Ferrell. Q&a: Cooperativity. *Journal of Biology*, 8(6):53, 2009.

[55] Timothy S Gardner, Charles R Cantor, and James J Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403(6767):339–342, 2000.

[56] Rui Ma, Jichao Wang, Zhonghuai Hou, and Haiyan Liu. Small-number effects: a third stable state in a genetic bistable toggle switch. *Physical review letters*, 109(24):248107, 2012.

[57] Thomas G Kurtz. The relationship between stochastic and deterministic models for chemical reactions. *The Journal of Chemical Physics*, 57(7):2976–2978, 1972.

[58] Melissa Vellela and Hong Qian. A quasistationary analysis of a stochastic chemical reaction: Keizers paradox. *Bulletin of mathematical biology*, 69(5):1727–1746, 2007.