LEARNING AND PREDICTION WITH DYNAMICAL SYSTEM MODELS OF GENE REGULATION

A DISSERTATION
SUBMITTED TO THE INSTITUTE FOR COMPUTATIONAL
AND MATHEMATICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Arwen Vanice Meister
November 2013

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Wing H. Wong) Principal Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Walter Murray) Principal Co-Advisor

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Chiara Sabatti)

Approved for the University Committee on Graduate Studies

Abstract

Biological structure and function depend on complex regulatory interactions between many genes. A wealth of gene expression data is available from high-throughput genome-wide measurement and single-cell measurement technologies, but systematic gene regulation modeling strategies and effective inference methods are still needed. This thesis focuses on biophysics-based dynamical system models of gene regulation that capture the mechanisms of transcriptional regulation at various degrees of detail. Deterministic modeling is fairly well-established, but algorithms for inferring the structure of novel gene regulatory systems are still lacking. We propose a method for learning the parameters of a standard nonlinear deterministc model from experimental data, in which we transform the nonlinear fitting problem into a convex optimization problem by restricting attention to steady-states and using the lasso for parameter selection. Stochastic modeling is much less mature. The Master equation model captures the mechanisms of gene regulation in full molecular detail, but it is intractable for all but the simplest systems, so simulation and approximations are essential. To help clarify the often-confusing situation, we present a simulation study to demonstrate the qualitative behavior of multistable systems and compare the performance of the van Kampen expansion, Gillespie algorithm, and Langevin simulation.

Acknowledgements

I'm grateful to my advisors, Wing H. Wong and Walter Murray, for their brilliance and guidance, to my collaborators, Chao Du, Henry Li and Bokyung Choi, with whom it's been an absolute pleasure to work, and to the faculty, staff, and students who make the Wong lab and ICME communities so warm and vibrant.

I would also like to thank my husband, Stefan, my parents, Donna and Mitch, my sisters, Mirabel and Tatiana, and my wonderful friends at Stanford and in Portland, for their love and support throughout my time in graduate school.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation

Complex interactions between many genes give rise to biological structure and function that sustain life. The Central Dogma [JM61, Cri70] provides a qualitative description of how these processes occur, but precise quantitative modeling is still needed [TCN03, Ros11]. Research into the detailed mechanisms of gene expression over the past few decades has shown that expression is regulated by a complex system of gene interactions. Recently, microarray and sequencing technologies [DIB97, RRW$^+$00, RHB$^+$07, MWM$^+$08] have enabled high-throughput genome-wide expression level measurements, enabling detailed study of gene networks [HJW$^+$98, LRR$^+$02, TYHC03, SSR$^+$03, BJGL$^+$03, HKI07, ZCMW07]. Even newer single-cell technologies allow for unprecedented resolution that reveals the stochastic character of gene regulation [IKN06, BBO$^+$09, TBO$^+$08, OBB$^+$10]. Appropriate mathematical frameworks are essential for making sense of the data, and can even suggest promising experimental designs for future studies. Our overarching goal is to use mathematical models to understand how genes interact to give rise to the biochemical complexity that allows organisms to live, grow and reproduce.

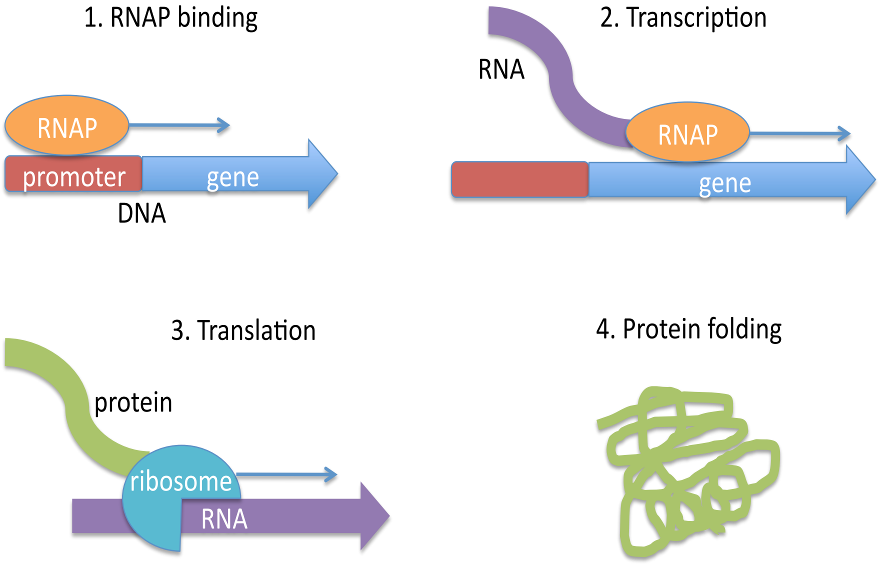


Figure 1.1: Basic steps of gene expression. 1. RNA polymerase (RNAP) binds to the gene promoter. 2. RNAP transcribes DNA sequence of gene to create an RNA copy. 3. Ribosome translates RNA transcript into a protein (amino acid sequence). 4. Protein folds into its functional configuration.

1.2 Introduction to gene regulation

For our purposes, a simplified view of the biology of gene regulation is sufficient. As Figure 1.1 shows, there are four basic steps involved in the expression of a particular gene (that is, production of the protein encoded by the gene). First, RNA polymerase (RNAP) binds to the gene promoter (a DNA region physically separate from the gene with a regulatory role). Next, the RNAP *transcribes* an RNA copy of the information encoded in the gene's DNA sequence. A ribosome then *translates* the RNA transcript to create a protein (consisting of a sequence of amino acids, each encoded by a sequence of four RNA 'letters', or base pairs). Finally, the protein folds into the final configuration that allows it to perform its function in the cell. Gene regulation (mostly) occurs during the RNAP binding step. As Figure 1.2 shows, regulatory proteins called *transcription factors* (TFs) can bind the gene promoter and modulate the binding energy of RNAP, thereby affecting the rate of gene expression. A TF

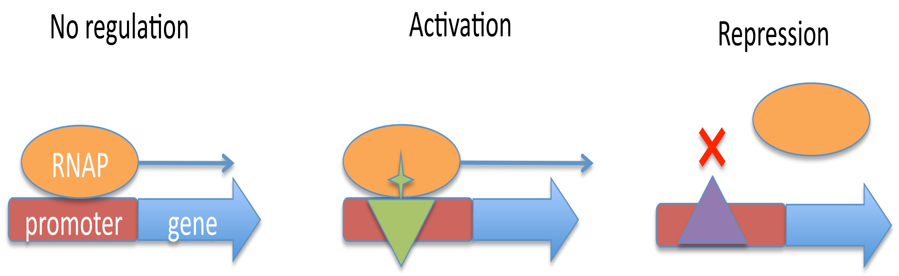


Figure 1.2: Gene expression is regulated by the binding of regulatory proteins (*transcription factors*, or TFs) to the gene promoter, which modulates RNAP binding energy. *Activators* encourage RNAP binding and accelerate gene expression, while *repressors* discourage or disallow RNAP binding, slowing or stopping gene expression. A particular gene may be regulated by zero, one, or several TFs.

is called an *activator* if, when bound to the promoter, it encourages RNAP binding and increases gene expression. A *repressor* makes RNAP binding more difficult or impossible, decreasing or stopping gene expression.

Each gene may be regulated by zero, one, or multiple TFs, each of which is a protein (or complex of multiple proteins) encoded by another gene (or in the case of self-regulation, by the gene itself). We say that gene A *regulates* gene B if the protein product of gene A is a TF that activates or represses gene B. Complex effects can arise when a relatively small collection of genes all regulate each other, giving rise to a *gene network*. A good example is the gene network responsible for pluripotency in embryonic stem cells, which we will discuss in much greater detail later on.

1.3 Literature review

The main contributions of this work will be an inference algorithm for learning a nonlinear deterministic dynamical system model of gene regulation, and a detailed simulation study comparing different approaches to stochastic modeling of gene regulation. In this section we will very briefly summarize some of the most important achievements in these two areas, on which our work builds. Additional literature review is provided in the chapters as appropriate.

1.3.1 Gene regulatory network inference

Gene expression measurements contain information useful for reconstructing the underlying interaction structure [DIB97, HJW$^+$98, HMJ$^+$00] because gene regulatory systems have a defined ordering [AW92], forming pathways that connect to form networks [Alo07, DSM10]. Many gene regulation pathways have been discovered over the past few decades [HHG$^+$10, AJL$^+$07]. At the turn of the century, researchers began applying statistical tools to genome-wide expression data to understand complex gene interactions. Eisen et al. (1998) showed that genes from the same pathways and with similar functions cluster together by expression pattern. Soon afterward, module-based network inference methods appeared, which group co-expressed genes into cellular function modules [SSR$^+$03, BJGL$^+$03]. Recently, methods based on descriptive but non-mechanistic mathematical models [GdBLC03, TYHC03, BBAIdB07, FHT$^+$07, Fri04] have gained prominence. These models describe gene regulation quantitatively and can be used to simulate and predict systems behaviors [Pal11, DESGS11]. However, more work is needed to develop effective model-based methods for inferring gene network structure from experimental data.

Existing inference methods typically rely either on heuristic approaches or on very simple, local models, like linear differential equation models in a neighborhood of a particular steady-state. Statistical corrrelation is a common method of establishing network connections [DESGS11] and can be very useful when hundreds or thousands of genes are monitored under specific, local cellular conditions (for instance, for grouping genes with similar functions). However, this approach works poorly when perturbations drive the network far from the original steady-state. Global nonlinear models are essential to account for complex global system behaviors, like the transformation of a normal cell into a cancerous cell via amplification of a particular gene.

1.3.2 Stochasticity

Experimental studies using single-cell biotechnologies have revealed that the biological mechanisms of gene regulation, including promoter activation and deactivation, transcription, translation, and degradation, are inherently stochastic [ELSS02, OTK$^+$02, BKCC03] [RWA02, KEBC05, RvO08, MNvO12] [HMM09]. Stochasticity can sometimes dramatically affect the behavior of gene regulatory networks [KS10, RvO08, PvO05] as stochasticity leads to different phase diagrams and can cause instability [KE01], and small molecular numbers can seriously impact system behavior [MWHL12]. These observations have important implications in synthetic biology, including the engineering of switches, feedback loops, and oscillatory systems [EL00, GCC00, HMC$^+$02, OTL$^+$04].

A number of stochastic models have been applied to gene regulation problem, since it is clear that additive noise (independent of expression level) is not sufficient in many cases [FCSI12]. Since gene regulation depends on a series of chemical reactions (including the binding of TFs and RNAP to the promoter, transcription, translation, and degradation), it can therefore be modeled with chemical equations [NT97, ARM98]. A number of ad-hoc approaches have also shown promise, include Poisson models (a birth-and-death model with constant rates has a Poisson steady-state distribution, but expression regulation or extrinsic noise often lead to non-constant rates, disrupting the Poisson character), fluctuation noise analysis in small systems [TvO01, OTK$^+$02, Tao04, RYA$^+$05, KSK$^+$07, MTK09], and structural inference on large networks based on noise correlation [DCL$^+$08, SOWES12]. The highest resolution is obtained from more sophisticated analysis based on the Master equation [VK07], including approximation methods [PMK06, HBS$^+$07, KSK$^+$07, MBBS08, SK12], and more accurate modeling via simulation or theoretical deduction [WWA10, LQ10, FCSI12, MWHL12, GMD12, KH08].

1.4 Objectives

We focus on dynamical system models of gene regulation since we require detailed, quantitative models that capture cells' ability to assume many different characters as they progress through their lifecycles and respond to stimuli. Dynamical systems models of gene regulation can be divided into two basic categories: deterministic and stochastic. In this thesis, we aim to advance the current state of knowledge in both categories by contributing a inference method for a standard deterministic model, and a simulation study comparing several different stochastic modeling approaches based on the Master equation.

1.4.1 Nonlinear inference

The deterministic domain is fairly well-established. A nonlinear deterministic model based from the thermodynamics of gene regulation [BBG$^+$05b, BBG$^+$05a] is considered standard and sufficiently complex for almost any application, and simplifications can easily be made for applications where simpler models are more appropriate. If we knew the model terms and parameters for a particular system, we would understand how it maintains and transitions between steady-states, and could predict its future behavior under a variety of conditions. Of course, the model is generally not known except for a few very well-studied systems, and identifying the terms and their coefficients directly is very difficult, if not impossible. We address this deficiency by proposing a systems-level inference method for selecting and fitting the parameters of the model, which brings the deterministic program full circle by enabling novel gene regulatory systems to be recovered directly from gene expression measurements.

1.4.2 Stochastic models

In the stochastic domain, not only are inference methods lacking, but researchers have not even reached a clear consensus on the correct approach to modeling. While the Master equation is generally accepted as the gold standard for modeling the processes of gene regulation in molecular detail, it is too complex to support network inference from experimental data. Approximations are needed to make it useful, but there is still disagreement and confusion over the proper way to carry them out. We attempt to shed light on this discussion by showing how N. G. van Kampen's theoretical expansion of the Master equation [VK07] applies to gene regulation, and by performing a detailed simulation study comparing several different approximation and simulation methods applied to synthetic systems with a variety of qualitative characteristics.

1.4.3 A hierarchy of gene regulation models

In this work, we discuss a number of different dynamical system models of gene regulation. To clarify the relationships between, we can organize them into a hierarchical family of models of decreasing complexity. At the apex of the hierarchy lies the Master equation model, which captures the full stochastic complexity of gene regulation. In this model, the RNA transcripts counts for each gene are stochastic random variables whose probability distributions evolve according to a system of differential equations called the Master equation. Many different stochastic approximations of the Master equation reduce the computational complexity at some cost to the accuracy.

The nonlinear deterministic model can be derived as an approximation of the Master equation (specifically, the master equation mean obeys the deterministic model to constant order) by appropriately truncating van Kampen's master equation expansion in the system size ([VK07]); other, higher-order approximations can be obtained in a similar principled manner. The great advantage of the deterministic model is a conceptually straightforward and experimentally feasible associated inference procedure that allows the underlying structure and parameters of a novel gene regulatory system to be recovered using perturbed steady-state data collected from a systematic set of experiments.

The nonlinear deterministic differential equations from the linear noise approximation can be linearized about any given steady-state to yield a linear system of deterministic differential equations that approximate the local behavior of the system near that steady-state. This model lends itself to a simple and robust inference procedure, although the recovered model's validity is of course limited to a neighborhood of the steady-state of interest. This approach is most appropriate for situations in which we are mostly interested in a particular steady-state as opposed to the global system behavior, or where data is either limited or very noisy.

1.4.4 Outline of the chapters

Our program in this dissertation will be to discuss gene regulation models and associated inference methods from simplest to most complex for clarity and a natural progression. The first few chapters focus on methods for learning linear and nonlinear deterministic dynamical systems models from experimental data, while the last two concentrate on stochastic modeling approaches to modeling.

In chapter 2, we will lay the groundwork by introducing linear and nonlinear deterministic dynamical system models of gene regulation, and discussing some of their quantitative and qualitative structural properties, including a review of Lyapunov stability theory. In chapter 3 we will show how the parameters of a deterministic linear dynamical system model can be robustly inferred from systematically-collected experimental data via convex optimization. In chapter 4 we will develop an experimental design and associated statistical inference method for learning this nonlinear model by transforming the fitting problem into a convex optimization problem by restricting attention to steady-states. We illustrate the method in detail by applying it to a synthetic six-gene network based on an embryonic stem cell subnetwork. In chapter 5, we discuss a nonlinear, stochastic Master equation model for gene regulation, which fully captures the mechanisms of the gene regulation in molecular detail. Since approximations and simulations of the Master equation are essential for studying large systems with multiple genes, we discuss van Kampen's expansion and multistable-system theory, the Gillespie algorithm, and the Langevin simulation, and show how each approach can be applied to the gene regulation problem. In chapter 6 we perform a detailed simulation study (including a new heuristic approach to constructing artificial multistable systems) to illustrate the behavior of systems with single versus multiple steady-states and compare the applicability and accuracy of the approximations and simulations described in chapter 5.

Chapter 2

Dynamical system models of gene regulation

Dynamical system models are ideal for quantitatively capturing cells' ability to assume different characters as they transition through their lifecycles and respond to stimuli. Their quantitative form means that they can be used to predict systems' future behavior, and they lend themselves to inference algorithms that allow the structure of novel gene regulatory systems to be learned from data. In the standard dynamical system model of gene regulation, the levels of RNA and protein evolve according to a system of differential equations. The basic assumptions is that each species of RNA is transcribed at a rate proportional the the probability of RNAP binding to the gene promoter (as a function of the expression levels of the TF proteins that regulate it) and degrades at a rate proportional to its current level, while the corresponding protein is translated at a rate proportional to the current RNA level and also degrades at a rate proportional to its own level.

While the translation and degradation rate constants are assumed to be fixed, the RNAP binding probability function is flexible both in its form and specific parameters. Depending on the level of detail required for a particular application, we may choose a linear form or one of several possible nonlinear forms. Choosing a linear binding probability function leads to a simple model and intuitive inference method (discussed in chapter 3), but has the drawback that the resulting system has only one steady-state, while the majority of interesting biological gene regulatory networks have multiple steady-states. Nonlinear forms allow us to model systems with multiple steady-states and thereby capture the range of states available to cells during their complex life-cycles. Many different nonlinear functional forms are possible, but a thermodynamic model due to Bintu et al is considered standard [BBG+05b, BBG+05a].

In this chapter, we first introduce general deterministic dynamical system models of gene regulation. Next we discuss the critical issue of steady-states and their stability, providing both a biological description and the necessary mathematical theory due to A. Lyapunov. Finally, we discuss standard linear and nonlinear forms of the RNAP binding probability function, which model the strongest mechanism of gene regulation. We focus primarily on the standard flexible biophysics-based nonlinear model due to Bintu et al, the basis of much of the work of the following chapters.

2.1 Dynamical system model

We can model gene expression regulation as a dynamical system by letting $x \in \mathbb{R}^n$ represent RNA concentrations and $y \in \mathbb{R}^n$ represent protein concentrations corresponding to a set of n genes. We assume that the production rate of the RNA transcript x_i of gene i is proportional to the probability $f(y)$ that RNA polymerase (RNAP) is bound to the promoter. That is, RNA transcription occurs at a rate τ_i whenever RNAP is bound to promoter. We model the probability that RNAP is bound to promoter as a nonlinear function f of y, since RNAP binding is regulated by a set of TFs. Further, we assume that the production of protein product y_i of gene i is proportional to the concentration of the RNA transcript x_i with rate r_i, and that both the RNA transcript and protein products of gene i degrade at fixed rates (γ_i^{r}, γ_i^{p}). Figures 1.1 and 1.2 may clarify the situation for the non-biologist. The resulting system of differential equations is:

$$\begin{aligned} \frac{dx_i}{dt} &= \tau_i f_i(y) - \gamma_i^{\mathrm{r}} x_i \\ \frac{dy_i}{dt} &= r_i x_i - \gamma_i^{\mathrm{p}} y_i. \end{aligned} \tag{2.1}$$

2.1.1 RNA-only simplification

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

$$\frac{dx_i}{dt} = \tau_i f_i(x) - \gamma_i x_i, \tag{2.2}$$

involving only the RNA concentrations x, which serve as a surrogate for the protein concentrations y. From a practical perspective, this model is often the only choice in applications involving experimental data, where protein data may not be available. While microarray and sequencing technologies enable fast genome-wide RNA expression level measurements [DvdHM+00, RHB+07, MWM+08], proteonomics has traditionally been much more difficult, although this is changing due to recent advances in mass spectrometry [VM12]. In certain situations, the model is not only the most practical but also the most accurate: for example, prokaryotes lack a distinct nucleus and perform transcription and translation simultaneously [Ral08].

The passage from (2.1) to (2.2) can be justified by assuming that the translation rate is fast relative to the time-scale of transcriptional regulation, so the protein reactions equilibriate quickly compared to the RNA reaction, that is $\frac{dy_i}{dt} \approx 0$. This implies that $y_i(t) = \frac{r_i}{\gamma_i^{\text{Protein}}} x_i(t)$, which means that the simplified equation holds by adjusting f to take the proportionality constant into account. The validity of this assumption is debatable, however, since transcription rates are still not very well understood. However, the RNA-only model is often the only practical option, since experimental technologies for measuring both mRNA and protein levels concurrently are not yet available.

2.2 Steady-states

One of the most important characteristics of gene regulatory networks is their ability to maintain multiple stable steady-states. Since steady-states and their stability are central to our problem, we pause to discuss their biological meaning and to review the mathematical tools needed for describing steady-states and determining stability.

Biologically, a steady-state of a gene regulatory system is one in which RNA and protein levels are constant: $\frac{dx_i}{dt} = \frac{dy_i}{dt} = 0$. Steady-states of the system correspond to cell states with roughly constant gene expression levels, like embyronic stem cell, skin cell or liver cell. In contrast, an embryonic stem cell in the process of differentiating is not in steady state. One of the most interesting features of certain gene regulatory networks is their ability to maintain multiple stable steady-states. For instance, the embryonic stem cell gene regulatory network can also maintain endoderm, trophectoderm, and differentiated stem cell steady-states.

2.2.1 Lyapunov stability theory

The classical theory for general dynamical systems due A. Lyapunov (1857-1918) provides all we need to characterize steady-states and their stability mathematially. We will outline the key definitions and theorems here; for a complete discussion, see a text like Walker's *Dynamical systems and evolution equations*, [Wal39]. The Lyapunov stability criterion will be very useful here, and again in chapters 5 and 6.

Consider a general nonlinear dynamical system of the form

$$\dot{x}(t) = f(x(t)), \tag{2.3}$$

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

Definition: The equilibrium x_e is said to be *stable* if for all $\epsilon > 0$ there exists $\delta > 0$ such that

$$\bar{x} \in \mathcal{B}(x_e, \delta) \implies \phi(t; \bar{x}) \in \mathcal{B}(x_e, \delta), \text{ for all } t \geq 0,$$

(where $\mathcal{B}(x, \epsilon)$ denotes the open ball of radius ϵ centered at x).
It is said to be *asymptotically stable* if it is stable, and for all $\epsilon > 0$ there exists $\delta > 0$ such that

$$\bar{x} \in \mathcal{B}(x_e, \delta) \implies \lim_{t \to \infty} \phi(t; \bar{x}) = x_e.$$

In essence, stability means that that there exists a neighborhood of the steady-state such that trajectories that start inside that neighborhood remain there for all time. Asymptotical stability means that in addition to this, nearby trajectories are attracted to the steady-state and eventually get infinitely close to it. The next theorem provides the essential stability criterion.
Theorem: Assume $f \in C^1(\mathbb{R}^n)$, and set $A = \frac{\partial f}{\partial x}(x_e)$ (the Jacobian matrix at x_e). If there exists a symmetric positive definite matrix P such that $A^T P + PA \prec 0$, then x_e is asymptotically stable.

2.3 Linear model

Gene regulatory systems' ability to maintain multiple stable steady-states is an important feature, and dynamical systems of the form (2.1) can only have multiple steady-states if the functions f_i modeling the RNA-promoter binding probability have a nonlinear form. However, researchers are sometimes mainly interested in the regulatory network responsible for maintaining a single steady-state of the system. The simplest and often most effective way to do this use a linear model, which represents the most basic effects of the regulators on their targets in a neighborhood of that steady-state. We can always derive such a model from a more complex one by linearizing the system about the steady-state of interest. As we will see in chapter 3, the linear model lends itself to a simple, robust and intuitive inference method based on perturbing each gene in turn and observing the system response.

The dynamical system model (2.1) is linear if the functions f_i are linear or affine:

$$f_i(y) = \sum_j b_{ij} y_j + c_i, \tag{2.4}$$

for some constants $b_{ij}, c_i \in \mathbb{R}$. The interpretation is simple: gene i has a baseline transcription rate τc_i, it is activated by y_j if $b_j > 0$, and it is repressed by y_j if $b_{ij} < 0$. The activation or repression is linear in the regulators y_j and does not "plateau" for large regulator concentrations (that is, there are no saturation effects). For this reason, the model is usually only physically realistic in a limited range.

Linearization about a steady-state

We can always derive a linear model from a nonlinear one by linearizing about a particular steady-state. Suppose we model the cell state as a time-varying vector $x(t) \in \mathbb{R}^n$ of gene expression levels that evolves according to

$$\frac{dx}{dt} = A(x(t)),$$

where $A : \mathbb{R}^n \to \mathbb{R}^n$ is a smooth nonlinear function. We can either use RNA levels as a surrogate for gene expression as in equation (2.2), in which case $A_i(x) = f_i(x) - \gamma_i x_i$, or model both RNA and protein levels as in in equation (2.1), in which case we set $z = (x, y) \in \mathbb{R}^{2n}$ and define

$$A_i(z) = \begin{cases} f_i(y) - \gamma_i^R x_i, & \text{if } 1 \leq i \leq n \\ r_i x_i - \gamma_i^P y_i, & \text{if } n+1 \leq i \leq 2n \end{cases}.$$

Steady-states μ such that $A(\mu) = 0$ correspond to basic cell types like embryonic stem cell or liver cell. Taylor expanding A about a steady-state μ yields:

$$\frac{dx}{dt} = A(x) \approx T(x - \mu) \implies x(t) - \mu \approx e^{tT}(x_0 - \mu), \tag{2.5}$$

where T is the $n \times n$ Jacobian matrix of A at μ and x is close to μ. The matrix T models the regulatory network at equilibrium: $T_{i,j} > 0$ if gene j up-regulates gene i; $T_{i,j} < 0$ corresponds to down-regulation. The diagonals of T reflect not only self-regulation, but also degradation of gene products. (We assume that gene degradation occurs at a known fixed rate γ.) The resulting model is reasonably accurate in a neighborhood of the steady-state μ. Furthermore, it is useful as a basis for learning the basic structure of the underlying gene network that maintains that particular steady-state.

2.4 Nonlinear models

Although linear models are appropriate in certain situations, their usefulness is limited by the fact that they cannot capture the ability of gene regulatory systems to maintain multiple stable steady-states. Since multistability is a key feature that often motivates the study of gene regulatory networks, we now turn our attention to nonlinear models. In this section we discuss some of the most popular choices.

Michaelis-Menten kinetics and the Hill equation are classical nonlinear model for activation or repression by a single factor, based on thermodynamic theory. Michaelis-Menten kinetics [MM13] can be applied to gene regulation by a single transcription factor by modeling transcription as the enzymatic reaction series

$$\begin{aligned} X + D_0 &\overset{k_{\pm 1}}{\longleftrightarrow} D_1 \\ D_1 + P &\xrightarrow{k_t} D_1 + P + Y, \end{aligned}$$

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

$$\frac{dD_1}{dt} = k_1 D_0 X - k_{-1} D_1 \tag{2.6}$$

$$\frac{dY}{dt} = k_t P D_1. \tag{2.7}$$

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

$$\frac{dD_1}{dt} = 0$$

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

$$\frac{dY}{dt} = k_t P \frac{D_T X}{K_1 + X}, \quad \text{where} \quad K_1 = \frac{k_{-1}}{k_1}, \quad D_T = D_0 + D_1.$$

A model for gene repression can be derived in a similar manner by replacing equation (2.7) with

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

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

$$\frac{dY}{dt} = k_t P \frac{D_T K_1}{K_1 + X}, \quad \text{where} \quad K_1 = \frac{k_{-1}}{k_1}, \quad D_T = D_0 + D_1.$$

The Hill equation [Hil13] is another classical model with a similar form, which models cooperative binding. For activation by a single transcription factor, for example, it has the form:

$$\frac{dY}{dt} = \frac{Z}{1 + Z}, \quad \text{where} \quad Z = {\frac{X}{K_1}}^{n},$$

where n is the Hill coefficient.

In order to account for multiple regulatory mechanisms in sufficient detail and generality, however, we require more complex approaches. We will focus primarily on a standard global nonlinear model: the quantitative, experimentally interpretable biophysics-based ordinary differential equation (ODE) gene regulation model of Bintu et al [BBG$^+$05b, BBG$^+$05a]. Many models of this type have been proposed, and the idea traces back to the beginning of systems biology in the biophysics field [AJS82, SA85, vHRGW74], but the Bintu model is widely accepted within the biophysics community [BBG$^+$05a, BBG$^+$05b]. The Bintu model is based on the thermodynamics of RNA transcription, the process at the core of gene expression regulation [HJW$^+$98, HKI07]. Transcription occurs when RNA polymerase (RNAP) binds the gene promoter; transcription factors (TFs) can modulate the RNAP binding energy to activate or repress transcription. RNA transcripts are then translated into protein.

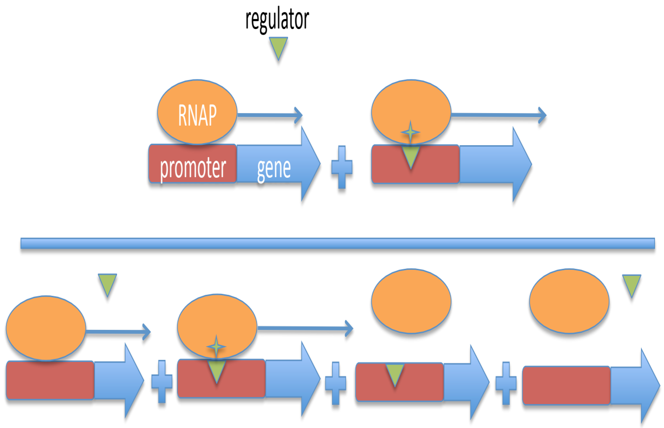


Figure 2.1: Illustration of the basic thermodynamic argument in the derivation of the Bintu et al form of RNAP binding probability function. The probability is equal to the weighted sum of system configurations in which RNAP is bound to the promoter divided by the weighted sum of all possible configurations.

Bintu models the mechanism of transcription in detail, using physically interpretable parameters. The form of the equations is rich and flexible enough to include the full range of gene regulatory behavior. Another notable biophysics-based model is that of the annual DREAM competition, but it has many biochemical assumptions and model parameters, like the Hill coefficient of transcription factor binding events, that cannot be estimated using gene expression measurements, so the network reconstruction requires ad hoc inference methods to learn the underlying gene interactions [YAYG10, PSdlF10, MPS$^+$10, SMF11]. Compared to the DREAM model, the Bintu model has the advantages of simplicity and interpretability, and better lends itself to principled inference.

2.5 Bintu model for binding probabilities

Based on the thermodynamics of RNAP and TF binding, one can deduce the following nonlinear form for the RNAP binding probability functions f_i that appear in equation

(2.1):

$$f_i(y) = \frac{b_{i0} + \sum_{j=1}^{m} b_{ij} \Pi_{k \in S_{ij}} y_k}{1 + \sum_{j=1}^{m} c_{ij} \Pi_{k \in S_{ij}} y_k}, \tag{2.8}$$

where S_{ij} lists the gene products that interact to form a regulatory complex, and b_{ij}, c_{ij} are nonnegative coefficients that must satisfy $c_{ij} \geq b_{ij} \geq 0$ [BBG$^+$05b, BBG$^+$05a]. (We assume that the concentration of each complex is proportional to the product of the concentrations of the constituent proteins, and absorb the proportionality constant into corresponding coefficients b_{ij}, c_{ij}). The coefficients b_{ij} and c_{ij} depend on the binding energies of regulator complexes to the promoter. b_{i0} and c_{i0} correspond to the case when the promoter is not bound by any regulator ($\Pi_{k \in S_{i0}} y_k = 1$), and the coefficients are normalized so that $c_{i0} = 1$. A detailed derivation is given in appendix A1, but Figure 2.1 illustrates the basic thermodynamic principle: the numerator of f represents the weighted sum of possible configurations in which RNAP is bound to the promoter (with or without a TF), and the denominator is the weighted sum of all possible configurations of the system (RNAP and TF bound or unbound).

The form of f_i allows us to model the full spectrum of regulatory behavior in quantitative detail. Terms that appear in the denominator only are repressors, and the degree of repression depends on the magnitude of the coefficient, while terms that appear in the numerator and denominator may act as either activators or repressors depending on the relative magnitudes of the coefficients and the current gene expression levels. Terms may represent either single genes or gene complexes. The model can even be extended to account for environmental factors that affect gene regulation, though we will not discuss it further here.

As an example, consider the simple two-gene network shown in Figure 2.2. Suppose that genes 1 and 2 have RNA concentrations x_1, x_2, and protein concentrations y_1, y_2, respectively, and that gene 1 is activated by protein 2 and repressed by its own product, while gene 2 is repressed by a complex formed by proteins 1 and 2. The

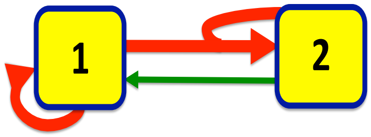


Figure 2.2: Simple two-gene network example described by equation 2.9 (with parameters $b_{11} = c_{11} = 0.1$ for activators; $c_{12} = 10$ for repressors; and $b_{10} = 0.01$ for constants in the numerator). Gene 1 is activated by the protein product of gene 2 and repressed by its own product (an example of self-regulation). Gene 2 is repressed by a complex formed by the product of gene 1 and its own product (synergistic self-regulation). In the diagram, the edge colors indicate activation (green) or repression (red) and the edge weights indicate coefficient sizes, illustrated above with typical sizes.

situation corresponds to the following equations:

$$\begin{aligned}
\frac{dx_1}{dt} &= \tau_1 \frac{b_{10} + b_{11} y_2}{1 + c_{11} y_2 + c_{12} y_1} - \gamma_1^{RNA} x_1, & \frac{dy_1}{dt} &= r_1 x_1 - \gamma_1^{Protein} y_1 \\
\frac{dx_2}{dt} &= \tau_2 \frac{b_{20}}{1 + c_{21} y_1 y_2} - \gamma_2^{RNA} x_2, & \frac{dy_2}{dt} &= r_2 x_2 - \gamma_2^{Protein} y_2.
\end{aligned} \tag{2.9}$$

In the notation above, we have $S_{11} = \{2\}, S_{12} = \{1\}, S_{21} = \{1, 2\}$. The parameters $b_{10}, b_{11}, c_{11}, \ldots$ determine the magnitude of the repression or activation. As this example shows, the model is flexible enough to capture a wide range of effects, including self-regulation (that is, regulation of a gene by its own protein product, most commonly as repression) and synergistic regulation by protein complexes (two or more proteins bound together to form a regulatory unit), in quantitative detail. Furthermore, the model is predictive: if we know or can infer the coefficients in the model, we can predict the future behavior of the system starting from any initial condition.

Chapter 3

Linear Inference

The simplest way to study the regulatory network responsible for maintaining a single steady-state of the system is based on a linear model (equation (2.1) with f of the form (2.4)), which represents the most basic effects of the regulators on their targets in a neighborhood of that steady-state. The linear model lends itself to a simple, intuitive and robust inference methods based on observing the system response to various perturbations. The perturbation-response approach is probably the most popular method of ODE-based gene network inference since it is so straightforward [FHT$^+$07], and has led to many successes [BBAIdB07, GdBLC03, dBTG$^+$05], but it has the drawbacks of being limited to a neighborhood of a single steady-state, and requiring timeseries data that can be rather challenging to obtain experimentally. Furthermore, the inevitably noisy data can lead to high variance and produce models which are not physically realistic. In this chapter, we explain the basic approach and discuss regularization techniques that help alleviate the issues related to noise. We analyze their effectiveness by applying them to a synthetic six-gene network.

3.1 Steady-state network inference

Consider a linear dynamical system of the form

$$\frac{dx}{dt} = T(x - \mu)$$

with a single steady-state μ. In the last chapter we derived this equation by linearizing a nonlinear system about the steady-state μ: T was the Jacobian of the original nonlinear rate function evaluated at μ. Alternatively, we may simply accept this form. The solution is given by:

$$x(t) - \mu \approx e^{tT}(x_0 - \mu),$$

but for model fitting purposes it is easier to work with the differential form.

It is straightforward to infer T at a particular steady-state μ using perturbation-response data. There are two basic approaches, depending on the nature of the perturbation. If the perturbation is permanent (that is, the system does not eventually return to its original steady-state) but small enough not to send the underlying system to a completely different steady-state, then

$$Tx + u = 0,$$

where u represents a perturbation. Given n linearly independent perturbations u and noiseless measurements of x, we could solve a linear system of equations for T exactly. Both Gardner and di Bernardo [GdBLC03, dBTG$^+$05] used this approach with different types of perturbations.

Bansal et al [BBAIdB07] used a slightly different approach, which will adopt for the remainder of chapter. If we measure the derivative shortly after a perturbation $x_0 = \mu + \epsilon$, then

$$\frac{\Delta x}{\Delta t}|_{x_0} \approx T\epsilon, \quad \text{where} \quad \frac{\Delta x}{\Delta t}|_{x_0} = \frac{x(t) - x_0}{t - t_0}.$$

In principle, with n linearly independent perturbations and exact derivative measurements we could recover T exactly. To be even more concrete, suppose we perturb one gene at a time, i.e. $x_j^0 = \mu + \epsilon e_j$, where $e_j(k) = \delta_{j,k}$. Then

$$\frac{\Delta x}{\Delta t}|_{x_j^0} \approx \epsilon t_j,$$

where t_j is the jth row of T. That is, the response to a single-gene perturbation determines the corresponding row of T (whose entries describe the regulatory effects of gene j on every other gene).

Of course, the data are likely to be very noisy, leading to parameter estimates with high variance and hence large error. Regularization based on structural knowledge can reduce error and also produce more physically-interpretable and biologically useful models. We know that the regulatory network is sparse, since each regulator has only a few targets. That is, $T + \gamma I$ should be sparse (taking the degradation rate γ into account on the diagonal). Furthermore, we know the equilibrium is stable, since the cell recovers from small perturbations [LB03]. Mathematically, μ is stable if there exists a Lyapunov matrix P such that $PT + T^T P \prec 0$ [Wal39], or equivalently, if the eigenvalues of T all have non-positive real parts.

To recover the network matrix T from noisy data $x(t)$ following a perturbation x_0, we can solve

$$\begin{array}{ll} \text{minimize} & \|\frac{x(t)-x_0}{t-t_0} - T(x_0-\mu)\|_2 + \lambda \sum_{i,j} |(T+\gamma I)_{i,j}| \\ \text{subject to} & PT + T^T P \prec 0 \end{array}$$

with variables $T \in \mathbb{R}^{n\times n}, P \in \mathbb{R}^{n\times n}$, and data $t, \gamma \in \mathbb{R}, \mu, x_0, x(t) \in \mathbb{R}^n$. The ℓ_1-regularization term encourages sparsity [Tib96]. The problem is not jointly convex in T and P, so we will use an iterative heuristic to solve it approximately and efficiently [ZJPP11]. We will alternately fix one variable and solve in the other, starting with $P = I$ fixed.

For simplicity, we gave the formulation for one perturbation x_0 and one measurement, while we actually need at least n perturbations to recover $T \in \mathbb{R}^{n\times n}$, and might have several measurements. Assuming N perturbations $x_0^{(j)}$ leading to trajectories $x^{(j)}(t)$, $j = 1, \ldots, N$ and m measurements per trajectory, the problem data are $\mu \in \mathbb{R}^n, \gamma \in R, t_i \in \mathbb{R}, x_0^{(j)}, x^{(j)}(t_i)$, $j = 1, \ldots, N, i = 1, \ldots, m$, and the complete problem is:

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{N} \sum_{i=1}^{m} \|\frac{x^{(j)}(t_i)-x_0^{(j)}}{t_i-t_0} - T(x_0^{(j)}-\mu)\|_2 + \lambda(\sum_{i\neq j} |(T+\gamma I)_{ij}| \\ \text{subject to} & PT + T^T P \prec 0 \end{array}$$

with variables T, P.

3.2 Knockdown data

One way to obtain perturbation data is from noisy genome-wide expression measurements shortly after a gene “knockdown,” in which the expression level of one gene is reduced to a fixed level. Since gene knockdown can often send gene regulatory systems to a completely different steady-state, this type of perturbation is usually too severe for approach (1). With this type of data, it is better to apply approach (2) to derivatives estimated from expression levels measured shortly after the perturbation, since we can at least be fairly confident that the system is still in a neighborhood of the steady-state of interest. Modeling a knockdown as a small perturbation and the subsequent evolution as an exponential trajectory is a rather poor approximation, but data fitting combined with regularization can still allow approximate network recovery. Recovering the diagonals of T is particularly challenging, since the knockdowns fix gene expression at a reduced level, thereby preventing direct detection of self-regulation. Multiple time points can yield indirect information, since perturbing a regulator at time t_0 leads to perturbed targets at time t_1, and we can observe the effects of target self-regulation at time t_2. However, the signal is very weak compared to the direct signal, so regularization is especially important for the diagonals.

3.3 Demonstration

We demonstrate the approach on a synthetic model of a six-gene subnetwork in embryonic stem cell, where the network matrix T_{true} is known [CP08]. The network and T_{true} are shown in Figure 3.1, and the system is discussed in much more detail in the next chapter. To generate data, we fix each gene in turn at 50% of equilibrium level and let the others evolve. We sample $x(t_1), x(t_2)$ for small t_1, t_2. To generate noisy versions of the data, we add 10% Gaussian noise to the signal.

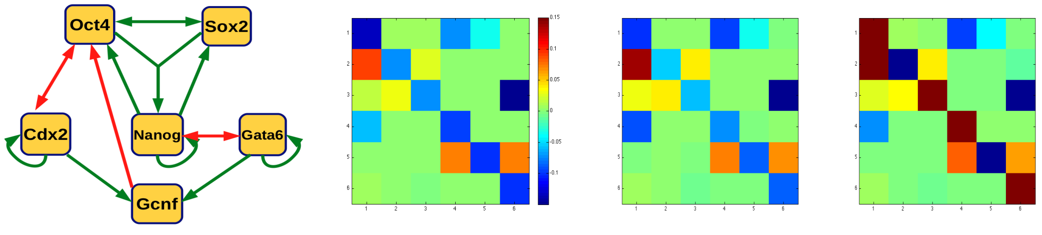


Figure 3.1: Network structure; T_{true}; basic clean recovery; basic noisy recovery (without enforcing sparsity or stability).

3.3.1 Basic Recovery

We first try basic recovery, minimizing $\|(x(t) - x_0) - tT(x_0 - \mu)\|_2$ without enforcing sparsity or stability (Figure 3.1). In the noiseless case, the recovery works well. The matrix is not quite sparse or stable, but it has many nearly-zero entries and only one small positive eigenvalue. With noisy data, $T_{\text{recovered}}$ is still nearly sparse, but the diagonals are not recovered and the matrix has large positive eigenvalues, violating the stability constraint.

3.3.2 Enforcing sparsity

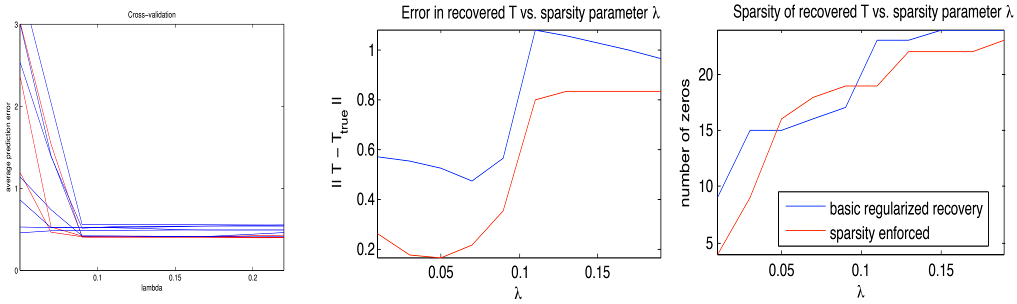


Figure 3.2: Sparsity parameter selection. Cross-validation for several noisy instances (left); absolute error in $T_{\text{recovered}}$ compared to T_{true} versus λ (center); sparsity of $T_{\text{recovered}}$ versus λ (right).

For noisy data, ℓ_1 regularization can improve both sparsity and diagonal recovery. We tune the sparsity parameter λ with leave-one-out cross validation, omitting each

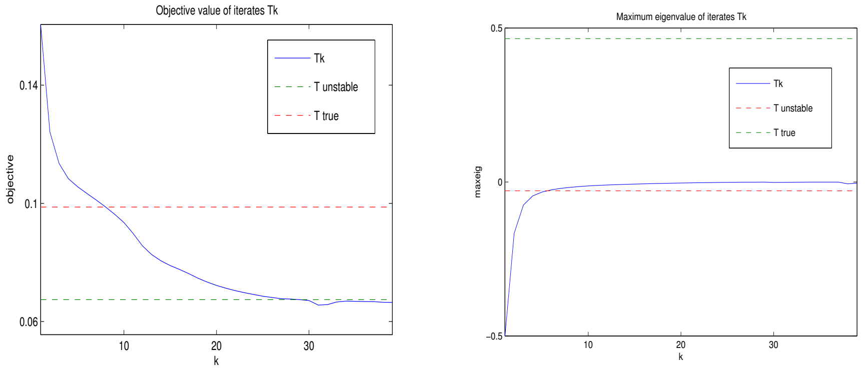


Figure 3.3: Stability iteration. Objective value of iterates T_k (left); maximum eigenvalue (real-part) of iterates T_k (right).

knockdown in turn, fitting on the other five, and testing on the omitted data. We then average the prediction error over all the test sets. Figure 3.2 (left) shows the results for several noisy data instances. The error drops sharply at around $\lambda = 0.09$; further increasing λ does not significantly change the error, but choosing λ too large makes the recovery too sparse. $\lambda = 0.1$ seems a reasonable choice. The plots of the absolute error and sparsity of T versus λ in Figure 3.2 indicate that $\lambda = 0.1$ provides a good tradeoff between accuracy and sparsity.

3.3.3 Enforcing stability

We enforce the stability constraint using an iterative heuristic in which we solve alternately in T and P, starting with $P = I$. The iterates are always feasible ($P_k T_k + T_k^T P_k \prec 0 \; \forall k$), but the iteration is not guaranteed to converge to the solution, nor are there non-heuristic stopping criteria. We terminate when $\|T_k - T_{k-1}\| \leq \epsilon$ for some tolerance ϵ.

```
$P = I$;    $k = 1$;
while $\|T_k - T_{k-1}\| \geq \text{tol}$
    $T_k = \underset{P_kT+T^TP_k \prec 0}{\text{argmin}} \; \|(x(t) - x_0) - tT(x_0 - \mu)\|_2 + \lambda \sum_{i,j} |(T + \gamma I)_{i,j}|$
    $P_k = \underset{PT_k+T_k^TP \prec 0}{\text{argmin}} \; (0)$
    $k = k + 1$.
```

We test on noisy data with $\lambda = 0.1$. Plots of the objective value and maximum eigenvalue of the iterates T_k are shown in Figure 3.3. The optimum objective value is unknown. The objective values of the iterates do appear to converge to the objective value of the matrix recovered from the same noisy data instance without enforcing stability. T_{true} has a higher objective value (since our model is only approximate, the recovered matrices fit it better than T_{true} does). Since stability is equivalent to $\text{Re}(\lambda_i(T)) \leq 0 \; \forall i$, the maximum eigenvalue of the T_k provides a measure of stability. The maximum eigenvalues of the iterates increase quickly to just below zero, so the stability condition is not unnecessarily strict in the end.

3.4 Conclusions

We can recover the network matrix T from noisy data quite successfully using the linear model with ℓ_1-regularization and iterative enforcement of the stability constraint. Figure 3.4 shows a matrix recovered from noisy data using this method. It has the right sparsity level, corresponds to a stable equilibrium at μ, and captures the off-diagonals of T_{true} very well and the diagonals reasonably well. The basic unregularized approach works reasonably well by itself, but the sparsity and stability constraints help by imparting desired qualitative properties to the network, and also by regularizing the solution, greatly improving the network recovery from noisy data. The basic perturbed-response inference approach and variants like this one are among the

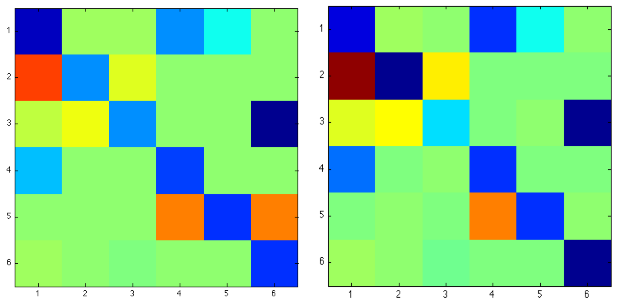


Figure 3.4: Successful recovery. T_{true} (left) (16 zeros and $Re(\lambda_{\max}) = 0.03$); T recovered from noisy data with sparsity and stability (right) (17 zeros and $Re(\lambda_{\max}) = 0.0006$).

most popular methods for learning gene networks, but their applicability is restricted to neighborhoods of a particular steady-state. In the next chapter we overcome this limitation by proposing a method for learning a nonlinear model of gene regulation.

Chapter 4

Nonlinear Inference

In this chapter, we propose an experimental design and associated statistical method for inferring an unknown gene network by fitting the standard global nonlinear Bintu model of gene regulation [MLCW13]. The required data is gene expression measurements at a set of perturbed steady-states induced by gene knockdown and overexpression [HGYI05]. We show how to design a sequence of experiments to collect the data and how to use it to fit the parameters of the Bintu model, leading to a set of ODEs that quantitatively characterize the regulatory network. Although the original fitting problem is nonlinear, we can transform it into a convex optimization problem by restricting our attention to steady-states. We use the lasso [Tib96] for parameter selection. As a proof of principle, we test the method on a simulated embryonic stem cell (ESC) transcription network [CP08] given by a system of ODEs based on the Bintu model. Here, we demonstrate that the inference algorithm is computationally efficient, accounts for synergistic regulation and self-regulation, and correctly recovers the parameters used to generate the data. Furthermore, the method requires only a set of steady-state gene expression measurements. Experimental researchers in the biological sciences can use this method to infer gene networks in a much more principled, detailed manner than earlier approaches allowed.

4.1 Inference problem

The model given by equations (2.1) and (2.8) fully describes the evolution of RNA and protein levels and provides a comprehensive, quantitative model of gene regulation, provided we know the parameters. Unfortunately, b_{ij}, c_{ij} are extremely difficult to measure, as they depend on binding energies of RNAP and TFs to the gene promoter. The sheer number of measurements required to characterize all possible TFs (both individual proteins and complexes) also makes this approach infeasible. Therefore, our goal is to use a systems level approach to fit the model using RNA expression data. Specifically, we will assume that $\tau_i, \lambda_i^{RNA}, \lambda_i^{Protein}$ are known or can be measured (if these these quantities are not available we can simply absorb them into the coefficients b_{ij}, c_{ij}, although more accurate rate estimates will likely improve the coefficient estimates). Our data will be measurements of the RNA concentrations x at many different cellular steady states (which correspond to steady-states of the dynamical system). The problem is to infer the values of the coefficients b_{ij}, c_{ij}.

4.2 Linear problem at steady-state

The key to solving this problem efficiently is to restrict our attention to steady-states, as proposed by Choi [Cho12]. This restriction allows us to transform a nonlinear ODE fitting problem into a linear regression problem. A steady-state of the system is one in which RNA and protein levels are constant: $\frac{dx_i}{dt} = \frac{dy_i}{dt} = 0$. As discussed in Chapter 2, steady-states of the system correspond to cell states with roughly constant gene expression levels, like embyronic stem cell, skin cell or liver cell; in contrast, an embryonic stem cell in the process of differentiating is not in steady state. Perturbed steady-states are particularly interesting. After a perturbation like gene knockdown, a cell's gene expression levels are in flux for some time while they adjust to the change. Eventually, if it is still viable, the cell may settle to a new steady-state [HGYI05]. These perturbed steady-states are especially helpful for understanding gene regulation.

In our model, the steady-state conditions $\frac{dx_i}{dt} = \frac{dy_i}{dt} = 0$ mean that:

$$0 = \tau_i f_i(y) - \lambda_i^{RNA} x_i, \qquad 0 = r_i x_i - \lambda_i^{Protein} y_i \implies y_i = \frac{r_i x_i}{\lambda_i^{Protein}}.$$

Defining $\tilde{f}_i(z) = f_i(\frac{r_i}{\lambda_i^{Protein}} z)$ yields

$$0 = \tau_i \tilde{f}_i(x) - \lambda_i^{RNA} x_i,$$

Absorbing the constants into the coefficients b_{ij}, c_{ij}, (so that $\tilde{b}_{ij} = b_{ij} \Pi_{k \in S_{ij}} \frac{r_k}{\lambda_k^{Protein}}$, $\tilde{c}_{ij} = c_{ij} \Pi_{k \in S_{ij}} \frac{r_k}{\lambda_k^{Protein}}$) we obtain the final equation

$$\tau_i \frac{b_{i0} + \sum_j b_{ij} \Pi_{k \in S_{ij}} x_k}{1 + \sum_j c_{ij} \Pi_{k \in S_{ij}} x_k} - \gamma_i x_i = 0,$$

or

$$\tau_i (b_{i0} + \sum_j b_{ij} \Pi_{k \in S_{ij}} x_k) - \gamma_i x_i (1 + \sum_j c_{ij} \Pi_{k \in S_{ij}} x_k) = 0,$$

(by multiplying both sides by the denominator). The last equation is linear in the coefficients b_{ij}, c_{ij}! In order to solve for b_{ij}, c_{ij}, we will need to collect many different expression measurements x at both naturally occurring and perturbed steady-states. Each steady-state measurement will lead to a different linear equation. These equations can be arranged into a linear system that we can solve for the coefficients.

4.3 Problem formulation

Our problem is to find b_{ij}, c_{ij} such that

$$0 = \tau_i (b_{i0} + \sum_j b_{ij} \Pi_{k \in S_{ij}} x_k^{(m)}) - \gamma_i x_i (1 + \sum_j c_{ij} \Pi_{k \in S_{ij}} x_k^{(m)}), \quad \forall m = 1, \ldots, M,$$

given RNA expression data $x^{(m)}$ at many different steady-state points $m = 1, \ldots, M$ and known translation and degradation rates $\tau_i, \lambda_i^{RNA}, \lambda_i^{Protein}$. (The experimental means of collecting the necessary steady-state expression data will be discussed in the next section.) We solve a separate problem for each gene i, since the coefficients b_{ij}, c_{ij} in the differential equation $dx_i/dt = \ldots$ for gene i are independent of the coefficients in the differential equations for other genes. Since we cannot know ahead of time which potential regulatory terms $\Pi_{k \in S_{ij}} x_k$ are actually involved, we include all possible terms up to second-order and look for sparse b_{ij}, c_{ij}, intepreting $c_{ij} = 0$ to mean that term $\Pi_{k \in S_{ij}} x_k$ is not a regulator of gene i.

Consider gene 2 in the two-gene example. Suppose we have expression measurements for a naturally occurring steady state (x_1^0, x_2^0), and a perturbed steady-state following gene 1-knockout $(0, x_2^1)$. We obtain two linear equations in the coefficients b_{20}, c_{21}:

$$\begin{aligned} \tau b_{20} - \lambda_2 x_2^0 (1 + c_{21} x_1^0 x_2^0) = 0, &\quad \text{(steady-state } (x_1^0, x_2^0)) \\ \tau b_{20} - \lambda_2 x_2^1 = 0, &\quad \text{(steady-state } (0, x_2^1)). \end{aligned}$$

If we knew a priori that complex $x_1 x_2$ was the only regulator of gene 2, these two equations would allow us to solve for the coefficients ($b_{20} = \frac{\lambda_2 x_2^1}{\tau}$, $c_{21} = \frac{x_2^0 - x_2^1}{(x_2^0)^2}$). Typically we do not know the regulators beforehand, however, and we need to use the data to identify them. That is, we include all possible terms (up to second-order) in the equations:

$$\tau(b_{20} + b_{21} x_1^{(m)} x_2^{(m)} + b_{22} x_1^{(m)} + b_{23} x_2^{(m)}) - \lambda_2 x_2^{(m)} (1 + c_{21} x_1^{(m)} x_2^{(m)} + c_{22} x_1^{(m)} + c_{22} x_2^{(m)}) = 0.$$

and estimate sparse coefficients b_{ij}, c_{ij} using several steady-state measurements $(x_1^{(m)}, x_2^{(m)})$. (We should find that the recovered coefficients $b_{21}, b_{22}, b_{23}, c_{22}, c_{23}$ are very close to zero, since the corresponding terms do not appear in the true equation.)

Temporarily suppressing the superscript m denoting the observation, we can compactly express the general system above by defining z_i as the vector with entries $z_i(j) = \Pi_{k \in S_{ij}} x_k$ (with the convention that $z_i(0) = 1$, $z_i(j) = x_j$ for $j = 1, \ldots, n$), which yields

$$0 = \tau_i b_i^T z_i - \gamma_i z_i(i) c_i^T z_i,$$

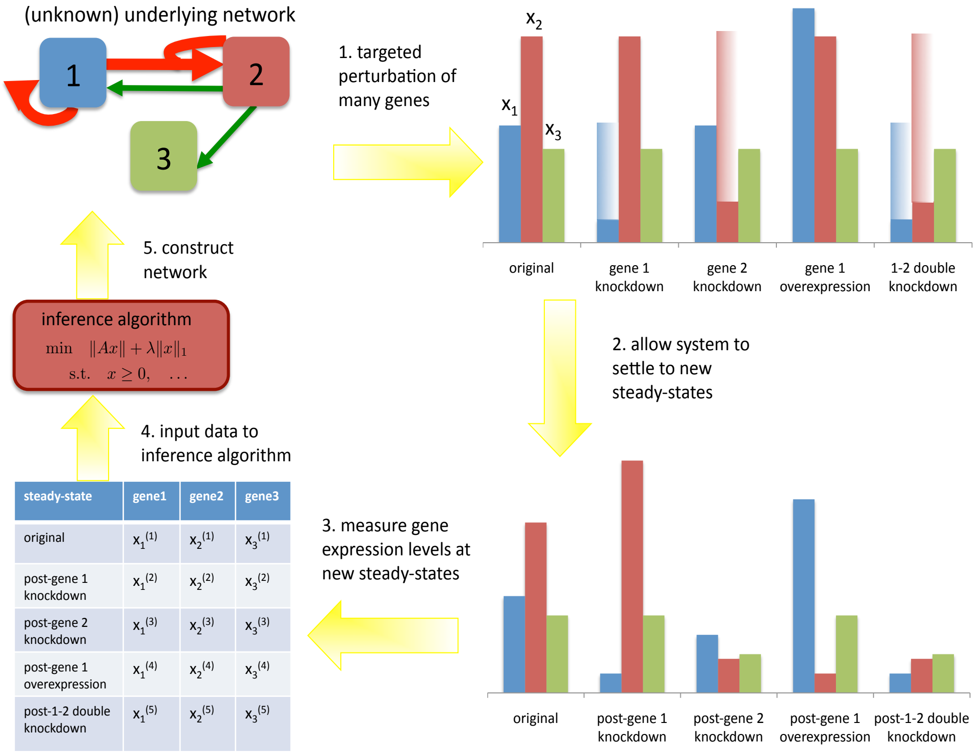
for each observation ($m = 1, \ldots, M$). If we form a matrix G_i by concatenating the row vectors $z_i^{(1)}, \dots, z_i^{(M)}$ and let D_i be a diagonal matrix with entries $z_i^{(m)}(i)$, $m = 1, \dots, M$, we can express this as

$$\begin{bmatrix} \tau_i G_i & -\gamma_i D_i G_i \end{bmatrix} \begin{bmatrix} b_i \\ c_i \end{bmatrix} = 0.$$

with the constraints $0 \leq b_i \leq c_i, \quad c_i(0) = 1$. Stating the problem in this form elucidates the required number of steady-state measurements, M. If the linear system above were dense and had no constraints on the coefficients b_{ij}, c_{ij}, and the steady-state expression vectors were (numerically) linearly independent, then we would require $M = 2T_{n,k}$, where $T_{n,k}$ is the number of the terms in the rational-form polynomial of degree k in n genes ($k = 2$ if we include up to second-order regulatory interactions). $T_{n,k}$ is equal to the number of subsets of $\{1, 2, \dots, n\}$ with k or fewer elements since each term represents an interaction between j distinct genes ($0 \leq j \leq k$), hence $T_{n,k} = \sum_{j=0}^{k} \binom{n}{j} \leq n^k$ for $k \leq n$ ($T_{n,2} \leq n^2$, for example). However, the constraints reduce the dimension of the solution space ($c_i(0) = 1$ reduces it by 1, while $0 \leq b_i \leq c_i$ reduces it by up to n), and our algorithm also uses ℓ_1-regression to search for sparse solutions, which may allow us to reconstruct the coeffcients from far fewer measurements than $2T_{n,k}$.

4.4 Experimental approach

The set of steady-state gene expression measurements needed to fit the model can be generated via a systematic sequence of gene perturbation experiments. Figure 4.1 summarizes the overall approach to finding the regulatory interactions among a set of genes comprising a (roughly) self-contained network of interest. First, molecular perturbations targeting each gene, or possibly pair of genes, in the network would be designed and applied one at a time. Following each perturbation, the cells would be allowed to settle down to a new steady-state, at which point the gene expression levels would be measured. The collection of gene expression measurements from different steady-states would be input to the inference algorithm described in the next section,


(unknown) underlying network
1
2
3
1. targeted perturbation of many genes
x_2
x_1
x_3
original
gene 1 knockdown
gene 2 knockdown
gene 1 overexpression
1-2 double knockdown
2. allow system to settle to new steady-states
3. measure gene expression levels at new steady-states
original
post-gene 1 knockdown
post-gene 2 knockdown
post-gene 1 overexpression
post-1-2 double knockdown
4. input data to inference algorithm
inference algorithm
min $\|Ax\| + \lambda\|x\|_1$
s.t. $x \geq 0, \quad \ldots$
5. construct network

steady-state	gene1	gene2	gene3
original	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$
post-gene 1 knockdown	$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$
post-gene 2 knockdown	$x_1^{(3)}$	$x_2^{(3)}$	$x_3^{(3)}$
post-gene 1 overexpression	$x_1^{(4)}$	$x_2^{(4)}$	$x_3^{(4)}$
post-1-2 double knockdown	$x_1^{(5)}$	$x_2^{(5)}$	$x_3^{(5)}$

Figure 4.1: Experimental approach for gene network inference. (1) Design and perform perturbation experiments targeting each gene (or possibly pair of genes) in the network: these may include overexpression, knockdowns, or knockouts. (2) Following each perturbation, allow the system to settle to a new steady-state. (3) Measure expression levels of all genes at each induced steady-state, and collect results in a data matrix. (4) Use steady-state expression data as input to inference algorithm. (5) Construct regulatory network from inference algorithm output.

which outputs a dynamical systems model of the gene network capable of predicting the behavior of the network following other perturbations. Perturbation data not used in the inference algorithm could be used to validate the recovered model.

The key experimental steps in this procedure, gene perturbations and gene expression measurements, are established technologies. Gene perturbations, including overexpression, knockdown, and knockout, are routinely used in biological studies to investigate gene function. These experiments can be performed for many laboratory organisms and cell lines both *in vitro* and *in vivo* [AJL$^+$07]. Overexpression experiments amplify a gene's expression level, usually by introducing an extra copy of the gene. Knockdown experiments typically use RNAi technology: the cell is transfected with a short DNA sequence, driven by a (possibly inducible) promoter element, that produces siRNA or shRNA that specifically binds the RNA transcripts of the gene of interest and triggers degradation. Morpholinos can also be used for gene knockdown. Gene knockout can be achieved by removing all or part of a gene to permanently disrupt transcription [AJL$^+$07]. Overexpression [RVL$^+$07], knockdown [RVL$^+$07, FCJ$^+$08], and knockout [LCH$^+$11] experiments have all been performed for the Oct4 gene, which helps maintain the stem cell steady-state. In some cases, much of the work is already done: for example, the *Saccharomyces* Genome Deletion Project has a nearly complete library of deletion mutants [WSA$^+$99].

Techniques for gene expression measurement are also well-established. Gene expression is usually measured at the transcript level: the RNA transcripts are extracted and reverse-transcribed into cDNA, which can be quantified with either RT-qPCR, microarray, or sequencing technologies [AJL$^+$07, MWM$^+$08]. Housekeeping gene expression measurements are used as controls to determine the expression levels of the genes of interest. The gene perturbations and subsequent expression measurements required to collect data for our inference algorithm may be time-consuming due to the large number of perturbations, but all the experimental techniques are quite standard and resources like deletion libraries can be extremely helpful.

4.5 Algorithm

We need to solve the linear system

$$\begin{bmatrix} \tau_i G_i & -\gamma_i D_i G_i \end{bmatrix} \begin{bmatrix} b_i \\ c_i \end{bmatrix} = 0.$$

for b_i, c_i, subject to the constraints $0 \leq b_i \leq c_i$, $c_i(0) = 1$. To account for measurement noise and encourage sparsity in b_i, c_i (since we know that each gene has only a few regulators), we will minimize the ℓ_2-norm error with ℓ_1 regularization [Tib96], which leads to the convex optimization problem

$$\begin{array}{ll} \text{minimize} & \left\| \begin{bmatrix} \tau_i G_i & -\gamma_i D_i G_i \end{bmatrix} \begin{bmatrix} b_i \\ c_i \end{bmatrix} \right\|_2^2 + \lambda \left(\|b_i\|_1 + \|c_i\|_1 \right) \\ \text{subject to} & 0 \leq b_i \leq c_i, \quad c_i(0) = 1, \end{array} \tag{4.1}$$

where λ is a parameter controlling sparsity that we can choose using cross validation. Since the problem is convex, it can be solved very efficiently even for large values of n and m.

Note that we use ℓ_1-regularization as a convex relaxation of a cardinality-constrained quadratic program, but the problem could also be stated as a mixed-integer quadratic program, which is feasible to solve for moderately sized systems. While our discussion focuses on the ℓ_1-regularized problem, the mixed-integer quadratic program approach is analogous: in particular, the same data is required, and we can choose the appropriate sparsity parameters using cross-validation.

4.6 Nonidentifiability

Our model's ability to capture self-regulation is very powerful, but it also leads to a particular form of nonidentifiability. For certain forms of the equation, given only steady-state measurements, it can be impossible to determine whether self-regulation is either completely absent or present in every term. Specifically, any valid equation of the form:

$$\frac{dx_i}{dt} = \frac{b_{i0} + \sum_{j=1}^{N} b_{ij} \Pi_{k \in S_{ij}} x_k}{1 + \sum_{j=1}^{N} c_{ij} \Pi_{k \in S_{ij}} x_k} - \gamma_i x_i, \quad b_{i0} < 1 \tag{4.2}$$

is indistinguishable at steady-state from any member of the following family of valid equations indexed by the constant w:

$$\frac{dx_i}{dt} = \frac{(wb_{i0} + \gamma_i)x_i + \sum_{j=1}^{N} wb_{ij} \Pi_{k \in S_{ij}} x_i x_k}{1 + wx_i + \sum_{j=1}^{N} wc_{ij} \Pi_{k \in S_{ij}} x_i x_k} - \gamma_i x_i, \quad w \geq \frac{\gamma}{1 - b_{i0}}. \tag{4.3}$$

We will refer to these as the 'simple' and 'higher-order' forms of the equation, respectively. The short proof of their equivalence is given in appendix A2. The condition $w \geq \frac{\gamma}{1-b_{i0}}$ guarantees that $w > 0$ and $0 \leq wb_{i0} + \gamma_i \leq w$ (since $0 \leq b_{i0} < 1$) and $0 \leq wb_{ij} \leq wc_{ij}$ (since $0 \leq b_{ij} \leq c_{ij}$).

We can distinguish between these two alternative forms by measuring the derivative of the concentration away from steady-state and comparing it to the derivative predicted by each form of the equation. This requires only a few extra thoughtfully-selected measurements. The details are in appendix A3.

4.7 Simulated six-gene subnetwork in mouse ESC

To demonstrate the inference approach, we apply our method to a synthetic six-gene system based on the Oct4, Sox2, Nanog, Cdx2, Gcnf, Gata6 subnetwork in mouse embryonic stem cell (ESC). Chickarmane and Peterson (2008) developed this system based on a synthesis of knowledge about ESC gene regulation accumulated over the past two decades [CP08]. The network structure is shown in Figure 4.3(a), and the

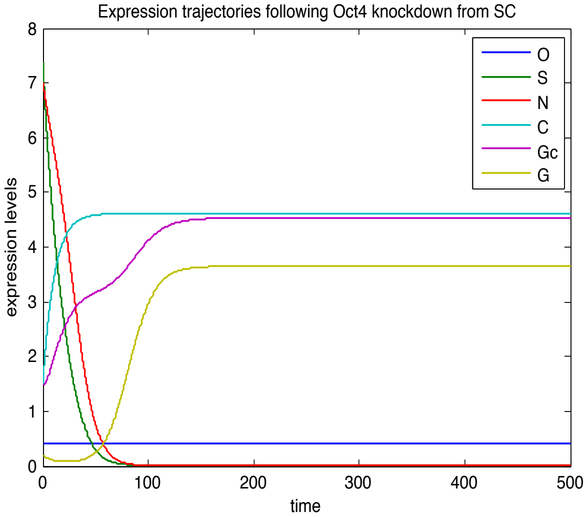


Figure 4.2: Gene expression trajectories during an Oct4 knockdown from SC steady-state. The expression of Oct4 is artificially reduced to 20% of its SC steady-state expression level and held there, causing the expression levels of the targets of Oct4 to change in response, which in turn impact their targets. The system eventually reaches a new steady-state different from SC. We measure the vector of expression levels at the new steady-state and use it as data in the inference algorithm. Since Oct4 is knocked down, this induced steady state does not provide useful information about the Oct4 equation, but it is useful for understanding the role of Oct4 and other genes in the equations of the remaining five genes.

detailed model is given by the following system of ODEs in the six genes:

$$
\begin{aligned}
\frac{d[O]}{dt} &= \frac{0.001 + [A] + 0.005[O][S] + 0.025[O][S][N]}{1 + [A] + 0.001[O] + 0.005[O][S] + 0.025[O][S][N] + 10[O][C] + 10[Gc]} \\
&\quad - 0.1[O] \\
\frac{d[S]}{dt} &= \frac{0.001 + 0.005[O][S] + 0.025[O][S][N]}{1 + 0.001[O] + 0.005[O][S] + 0.025[O][S][N]} - 0.1[S] \\
\frac{d[N]}{dt} &= \frac{0.001 + 0.1[O][S] + 0.1[O][S][N]}{1 + 0.001[O] + 0.1[O][S] + 0.1[O][S][N] + 10[O][G]} - 0.1[N] \\
\frac{d[C]}{dt} &= \frac{0.001 + 2[C]}{1 + 2[C] + 5[O][C]} - 0.1[C] \\
\frac{d[Gc]}{dt} &= \frac{0.001 + 0.1[C] + 0.1[G]}{1 + 0.1[C] + 0.1[G]} - 0.1[Gc] \\
\frac{d[G]}{dt} &= \frac{0.1 + [O] + 0.00025[G]}{1 + [O] + 0.00025[G] + 15[N]} - 0.1[G].
\end{aligned}
\tag{4.4}
$$

This model has many of the same qualitative characteristics as the biological mouse ESC network [CP08]. In particular, the system can support four different steady-states: embryonic stem cell (ESC), differentiated stem cell (DSC), endoderm and trophectoderm, and can switch from one to another when certain genes' expression levels are changed. In the Oct4 equation, A represents an external activating factor, whose concentration $[A]$ depends on the culture condition. Each of the four steady-states has a corresponding value of $[A]$: 10 for ESC and DSC, 25 for endoderm, and 1 for trophectoderm. For the remainder of this paper, we will regard $[A]$ as known. The explicit system of ODEs (4.4) allows us to generate data to fit our model and also to quantitatively compare our recovered solution to the ground truth. The qualitative similarity of this synthetic network to a real biological network gives us confidence that our results in this numerical experiment are likely to translate well to real biological networks. We observe that the Cdx2, Gcnf and Gata6 equations have alternative forms (provided we ignore the very small constant term in the $\frac{d[C]}{dt}$ equation and $[G]$ term in the $\frac{d[G]}{dt}$) . With the minimum possible value of w, the alternative forms are:

$$\begin{aligned}
\frac{d[C]}{dt} &= \frac{0.95}{1+2.5[O]} \quad (w=2) \\
\frac{d[Gc]}{dt} &= \frac{0.1001[Gc]+0.01[C][Gc]+0.01[Gc][G]}{1+0.1[Gc]+0.01[C][Gc]+0.01[Gc][G]} - 0.1[Gc] \quad (w=0.1) \\
\frac{d[G]}{dt} &= \frac{0.111[G]+0.111[O][G]}{1+0.111[G]+0.111[O][G]+1.67[N][G]} - 0.1[G] \quad (w=0.111).
\end{aligned} \tag{4.5}$$

To resolve the specific form, we will apply our method twice, once allowing self-regulation and again disallowing it. Then we will compare the two recovered forms of each equation and the quality of the fits to determine whether nonidentifiability exists in each case. If so, we will break the tie by examining derivatives.

The details of the simulation are given in appendix A4. We begin by testing the algorithm on noiseless data. We solve the optimization problem (4.1) once, then we solve it again with additional constraints prohibiting self-regulation. In each case, we use cross validation to select the sparsity parameter λ (Figure A1). The quality of the fit is comparable for the latter three equations whether we allow self-regulation or not,

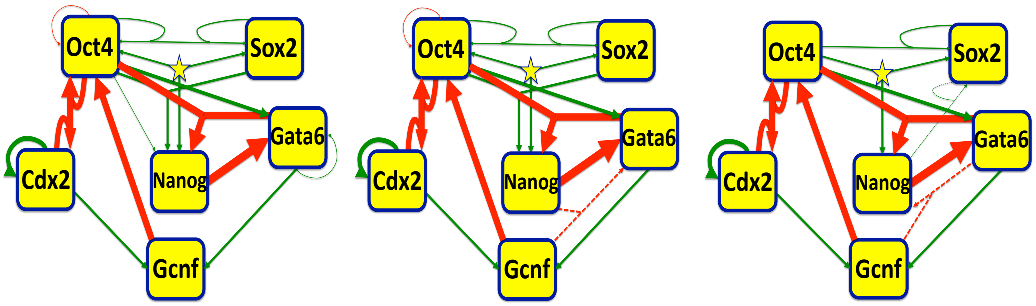


Figure 4.3: Recovery of a synthetic gene regulatory network based on the biological ESC network using our inference algorithm. The diagrams represent systems of ODEs that quantitatively model the gene interactions. Edge color indicates activation (green) or repression (red), and edge weights correspond to coefficient magnitudes. The arrows point from regulator to target, and self-loops indicate self-regulation. The yellow star represents the third-order complex OSN. (In addition to all possible first- and second-order terms, we allow this special third-order term with a free coefficient.) The left figure represents the original system of ODEs used to generate the data. The center figure shows the network recovered using our inference algorithm on noiseless data, and the right figure shows the recovery with 1% noise added. Both recovered networks reflect coefficient thresholding at 0.1% (noiseless case) or 1% (noisy case) of the largest recovered coefficent in each gene equation (with the exception of the noiseless-case Oct4 equation, thresholded at 0.01% to show the sucessful recovery of weak edges). The algorithm performs almost perfectly in the noiseless case, except for a false positive repressor on Gata6 and two very weak activation edges missing. In the noisy case, the algorithm recovers all of the strong edges, but misses some of the weaker ones and returns a few small false positives at our chosen thresholding level. Overall, the method captures the major network structure even in the noisy case.

while for the first three equations disallowing self-regulation has a significant negative impact on the fit (Table A1), indicating that the first three equations are unambiguous while the last three have two possible forms. To resolve the nonidentifiability in the latter three equations, we measure the derivatives of Cdx3, Gcnf and Gata6 immediately after some additional informative perturbations: Oct4, Cdx2 and Nanog knockouts, respectively (Figure A2). The test reveals that that Gcnf and Gata6 have the simple form, while Cdx2 has a higher-order form. In this example, the original

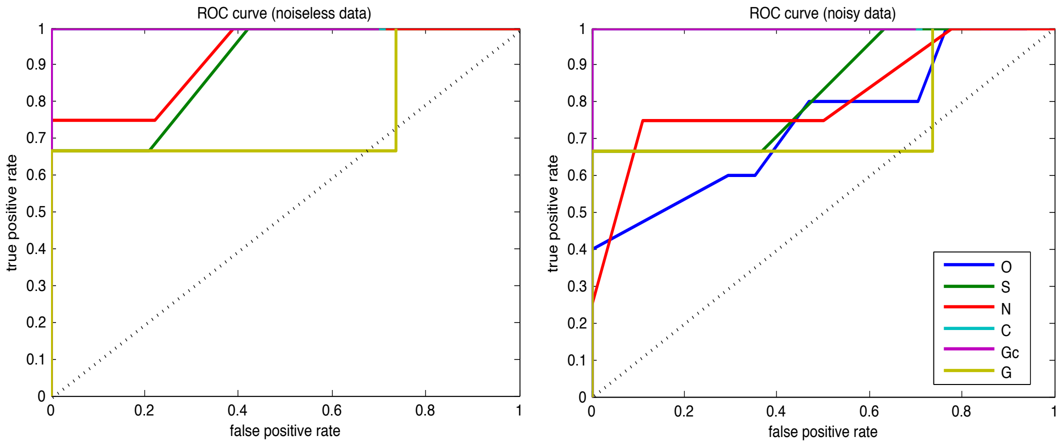


Figure 4.4: ROC curves for recovered networks from noiseless (left) and noisy (right) data showing the tradeoff between true positive rate (TPR) and false positive rate (FPR) for edge recovery. The ROC curves show the TPR and FPR that result from a range of coefficient threshold choices above which we consider an edge to have been recovered. For the equation $dx_i/dt = \ldots$ and threshold t, TPR is defined as the proportion of true edges j with $c_{ij}^{\text{recovered}} > t$ and FPR as the proportion of false edges with $c_{ij}^{\text{recovered}} > t$. For equations with two possible forms, we compare the simple forms of the true and recovered equations. Each gene equation has a different ROC curve as indicated by the legend. The dotted black line is the expected ROC curve for 'random guessing' algorithm, while the $(0, 1)$ point corresponds to a perfect algorithm (in fact, our algorithm performs perfectly for the Gcnf equation).

coefficients are recovered almost exactly:

$$\begin{aligned}
\frac{d[O]}{dt} &= \frac{0.001 + [A] + (0.005[O][S] + 0.025[O][S][N])}{1 + [A] + (0.001[O] + 0.005[O][S] + 0.025[O][S][N]) + 10[O][C] + 10[Gc]} \\
&\quad - 0.1[O] \\
\frac{d[S]}{dt} &= \frac{0.001 + 0.005[O][S] + 0.025[O][S][N]}{1 + 0.005[O][S] + 0.025[O][S][N]} - 0.1[S] \\
\frac{d[N]}{dt} &= \frac{0.1[O][S] + 0.1[O][S][N]}{1 + 0.1[O][S] + 0.1[O][S][N] + 10[O][G]} - 0.1[N] \\
\frac{d[C]}{dt} &= \frac{2[C]}{1 + 2[C] + 5[O][C]} - 0.1[C] \\
\frac{d[Gc]}{dt} &= \frac{0.001 + 0.1[C] + 0.1[G]}{1 + 0.1[C] + 0.1[G]} - 0.1[Gc] \\
\frac{d[G]}{dt} &= \frac{0.1 + [O]}{1 + [O] + 0.03[N][Gc] + 15[N]} - 0.1[G]. \qquad (4.6)
\end{aligned}$$

Next we add zero-mean Gaussian noise to each measurement, with standard deviation 1% of the measurement magnitude. We use the same steady-states as in the noiseless case, plus overexpression-knockdown of each pair of genes starting from ESC and DSC. Using a similar approach (detailed in appendix A4), we recover:

$$\begin{aligned}
\frac{d[O]}{dt} &= \frac{[A]}{1+[A]+9.9[Gc]+9.9[O][C]} - 0.1[O] \\
\frac{d[S]}{dt} &= \frac{0.001[O][S]+0.0005[S][N]+0.025[O][S][N]}{1+0.001[O][S]+0.0005[S][N]+0.025[O][S][N]} - 0.1[S] \\
\frac{d[N]}{dt} &= \frac{0.09[O][S][N]}{1+0.1[G][Gc]+0.09[O][S][N]+9.1[O][G]} - 0.1[N] \\
\frac{d[C]}{dt} &= \frac{2[C]}{1+2[C]+5[O][C]} - 0.1[C] \\
\frac{d[Gc]}{dt} &= \frac{0.1[C]+0.1[G]}{1+0.1[C]+0.1[G]} - 0.1[Gc] \\
\frac{d[G]}{dt} &= \frac{0.1+0.9[O]}{1+0.9[O]+14.2[N]} - 0.1[G].
\end{aligned} \tag{4.7}$$

In order to produce clean equations and network diagrams, we choose appropriate thresholds for each equation below which we zero the coefficients. (In practice, choosing thresholds is a judgment call based on the expected number of regulators, the noise level of the data, and the level of detail appropriate for the application.) We set the thresholds at 0.1% (noiseless case) or 1% (noisy case) of the largest coefficient recovered for each equation. For example, the largest recovered coefficient in the $d[G]/dt$ equation is roughly 15 in either case, so we zero the coefficients that fall below 0.015 (noiseless case) or 0.15 (noisy case). The recovered systems of equations shown above reflect these choices. In the noiseless case, relaxing the threshold on the Oct4 equation to 0.01% leads to the recovery of more correct terms, listed in parentheses. For completeness, we also provide receiver operating characteristic (ROC) curves in Figure 4.4 to show the tradeoff between true positives and false positives at other thresholds. The network diagrams in Figure 4.3(b,c) include an edge if the corresponding coefficient is above the threshold, with weights reflecting the size of the coefficients. These diagrams show that the recovery is nearly perfect in the noiseless case: using the gentler threshold for the Oct4 equation we recover all the true edges except for three very weak ones, and return just one small false positive repressor in the Gata6 equation. In the noisy case, we recover all the large coefficients correctly, although there are a few small false positives and the we miss several of the weakest edges. Overall, the method is able to capture the major network structure.

4.8 Discussion

Our experiment on the synthetic ESC system demonstrates that our algorithm can be used to infer a complex dynamical systems model of gene regulation and that the method can tolerate low levels of noise. Term selection from among all possible single gene and gene-complex regulators (up to second-degree interactions, plus the third-degree interaction OSN) was successful. The inferred equations are easy to interpret in terms of gene networks, and the detailed quantitative information allows for prediction of future expression trajectories from any starting point.

The approach is also scalable. Since we have formulated our problem as the convex optimization problem (4.1), it can be solved efficiently even for large systems using prepackaged software. Furthermore, it is trivially parallelizable, since we need to solve a version of (4.1) to infer the differential equation coefficients b_{ij}, c_{ij} for each gene i. Parallelization is even more helpful for the cross validation step, where we need to solve (4.1) for each gene and a sequence of choices sparsity parameter λ. We tested the scalability by running the algorithm with the parallelization discussed above on a simulated 100-gene system. The algorithm ran correctly in a reasonable time frame (a few hours) on a computing cluster.

The high resolution of our model is one of its most valuable features, but it means that accurate term selection may require much data, especially in the presence of noise. In our experiment, when we added 1% Gaussian noise, we needed extra data (knockdown/overexpression pairs) in order to accurately select terms. When we tried 5% noise, the algorithm consistently selected the large terms in five of the six equations, but we had to add even more data in order to correctly identify the major repressor in the Nanog equation. The Nanog equation is subtle in that Oct4 acts as both an activator in complexes with Sox2 and Nanog and a repressor in a complex with Gata6, so the algorithm tends to select different Gata6 complexes (or the Gata6 singleton) as the major repressor when the data is insufficient. In the 5% noise case, we needed additional data on the role of Gata6 (double-knockdowns and double-overexpression of pairs including Gata6 from ESC and DSC) in order to select Oct4-Gata6 as the major repressor of Nanog fairly consistently. As discussed earlier, another difficulty is the nonidentifiability that arises from accounting for self-regulation while restricting data to steady-states. Distinguishing between the two possible forms of nonidentifiable equations requires extra derivative data (which can be collected experimentally, although it is more difficult and time-consuming) and extra steps in the algorithm. The constraints on the convex optimization problem (4.1), which arise from thermodynamic considerations, are sufficient to prevent further nonidentifiability, but in certain cases, certain problems can suffer from near-nonidentifiability of other forms, which may contribute to the challenge of term-selection with noisy or limited data. We ensure accurate term selection by making sure we include enough diverse, high-quality steady-state measurements.

Finally, it is important to note that the processes of gene transcription and translation are inherently stochastic, since TF and RNAP binding result from chance collisions between molecules in the cell. We will study stochasticity in detail in the following chapters. Although the model discussed in this chapter does not explicitly account for intrinsic noise, we will show that is the first-order approximation of a Markovian model that captures the stochasticity of gene regulation in full molecular detail. We will also see that stochasticity in systems with multiple steady-states leads to multimodal steady-state expression distributions. Therefore, our algorithm is valid as long as we collect gene expression data by measuring the location of gene expression peaks, rather than measuring mean expression levels.

4.9 Conclusions

The model we use is based on the detailed thermodynamics of gene transcription, and quantitatively captures the full spectrum of regulatory phenomena in a detailed, physically interpretable, predictive manner. Since we can formulate the model fitting problem as a convex optimization problem, we can solve it efficiently and scalably using prepackaged software. ℓ_1-regularization allows for term-selection while maintaining the problem convexity. The experiments required to collect the necessary steady-state gene expression data are straightforward to perform, as technologies for knockdowns and overexpression are well-established and measuring gene expression is relatively simple. The model accounts for activation and repression by single-protein TFs and synergistic complexes as well as self-regulation, and describes the magnitude of each type of regulation in quantitative detail. Furthermore, the model can be extended to account for environmental effects and auxiliary proteins involved in regulation, including enhancers and chromatin remodelers. The fitted model can predict the evolution of the system from any starting point. Given a set of steady-states gene expression measurements, our algorithm can be used to fit a model which not only predicts further steady-states of the system, but also fully describes the transitions between them. Finally, beside the study of gene regulation, our approach will be useful in many other application areas where it is necessary to infer a nonlinear dynamical system by suitable experimentation and statistical analysis.

Chapter 5

Intrinsic Noise

Like any physical quantity, gene expression level measurements are subject to noise. In fact, the experimental techniques for measuring gene expression levels and biological quantities in general tend to be much noisier than measurements in other scientific disciplines. Fortunately, the extrinsic noise arising from measurement error can be modeled with a Gaussian distribution, so taking the mean of enough experimental replicates yields an accurate estimate of the true quantity. However, in addition to straightforward extrinsic noise, gene expression is also characterized by intrinsic noise arising from the fundamental stochasticity of the underlying processes, which cannot be simply averaged away [SES02, Pau04].

Intrinsic noise arises from the inherent stochasticity of gene transcription, translation, and degradation [ELSS02, RO05, OTK$^+$02, NT97, ARM98]. The probability that RNAP binds a gene promoter depends on the presence of regulators, which also bind the promoter with probability proportional to their concentrations. Degradation of each RNA transcript, and translation of RNA transcripts into proteins, are also stochastic processes.

In this chapter, we describe a model for gene regulation that properly accounts for the stochastic nature of the processes involved, namely, the Master equation. Since we cannot hope to solve the Master equation exactly except in simple cases, we discuss theoretical approximation based on an expansion method due to N. G. van Kampen and R. Kubo, and simulation algorithms due to Gillespie and Langevin.

The van Kampen expansion shows that the stochastic model reduces to the deterministic model (2.1) of the previous chapters in a first-order approximation, provided the system has a single stable steady-state. Stochastic systems with multiple stable steady-states are much more complex and alternate theory (also due to van Kampen) applies. The simulation studies of chapter 6 will provide further insight into the behavior of these systems.

5.1 The Master Equation

The basic approach to treating the stochasticity of gene regulation is to model gene expression level as a Markov process, whose future state depends probabilistically only on the current state. This is the most appropriate description for most processes in physics and chemisty ([VK07]), and this case is no exception: the mechanisms of transcription, translation, and degradation mean that the probability of each of these events depends only on current quantity of each of the species involved in these processes, including RNAP, transcription factors, and ribosomes (each of which is the product of one or more genes, and can therefore be accounted for our formulation). In this section we introduce the Master equation, which follows directly from the Markov property ([VK07]). The Master equation is the natural model for gene regulation under the Markov assumption.

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

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

Hence a Markov process is completely determined by the functions $P(y_1, t_1)$ and the transition probabilities $P(y_2, t_2 | y_1, t_1)$. The Master Equation holds for any Markov process: appendix A7 contains a complete derivation of the Master Equation as an equivalent form of the Chapman-Kolmogorov equation, which is a direct consequence of the Markov property (adapted from chapters IV and X of van Kampen's *Stochastic Processes in Physics and Chemistry* [VK07]). For a general Markov process Y, the

Master Equation reads:

$$\frac{\partial P(y,t)}{\partial t} = \int \{W(y|y')P(y',t) - W(y'|y)P(y,t)\}dy', \tag{5.1}$$

where $W(y'|y) \geq 0$ is the transition probability per unit time from y to y'. If the range of Y is a discrete set of states labeled by n, then (5.1) reduces to:

$$\frac{dp_n(t)}{dt} = \sum_n \left(W_{n,n'}p_{n'}(t) - W_{n',n}p_n(t)\right). \tag{5.2}$$

Birth-and-death processes

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

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

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

$$\dot{p}_n = r_{n+1}p_{n+1} + g_{n-1}p_{n-1} - (r_n + g_n)p_n. \tag{5.3}$$

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

$$\dot{p}_0 = r_1p_1 - g_0p_0,$$

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

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

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

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

Multivariable birth-and-death processes

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

$$\frac{dP(\mathbf{k},t)}{dt} = \sum_{j=1}^{n} \left[g_j(\mathbb{E}_j^- \mathbf{k}) P(\mathbb{E}_j^- \mathbf{k}, t) + r_j(\mathbb{E}_j^+ \mathbf{k}) P(\mathbb{E}_j^+ \mathbf{k}, t) - (g_j(\mathbf{k}) + r_j(\mathbf{k})) P(\mathbf{k}, t) \right]$$

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

The quasi-steady-state-assumption (QSSA)

It is often necessary or expedient to make the simplifying assumption that a certain species involved in a process is approximately in steady-state relative to other species. This is typically justified by separation-of-time-scales: for example, for a two-step reaction involving a fast reaction between a primary and intermediate species followed by a slower conversion of the intermediate into a final product (especially an irreversible reaction), we may assume that the net rate of formation of the intermediate species is approximately zero, since the first reaction equilibrates so quickly relative to the second. This is called the quasi-steady-state-assumption (QSSA). In the example, the QSSA lets us assume that the ratio of primary to intermediate species is approximately equal to its thermodynamic steady-state value. Appendix A6 follows Rao and Arkin in showing how to use the QSSA to eliminate an intermediate species from a multivariate Master equation [RA03].

5.1.1 The Master equation for gene regulation

We now wish to develop a stochastic model for gene regulation. We will start simply, considering a system with a single gene, and temporarily ignoring the distinction between RNA and protein. Let $X(t)$ be a discrete random variable representing the number of RNA transcripts present in the cell at time t. $X(t)$ has a time-dependent probability distribution given by $P(k,t) \equiv \mathbb{P}\{X(t) = k\}$. We can model $X(t)$ as a birth-and-death process with birth rate $\tau F(k)$ and death rate λk, where F models the RNAP-promoter binding probability as a function of the current number of transcripts (in the single-gene case, we can only account for self-regulation). If there are initially k RNA transcripts, then over an infinitesimal timestep Δt either a degradation event may occur with probability $\lambda k \Delta t$, a RNAP-promoter binding event may occur followed by RNA transcription with probability $\tau F(k)\Delta t$, or neither may occur. (It is highly unlikely ($O(\Delta t^2)$) that two or more of these events occur within a single timestep.) Hence, as Figure 5.1 shows, the probability $P(k,t)$ increases by $P(k-1)$ times the probability transcription plus $P(k-1)$ times the probability degradation, and decreases by $P(k)$ times the probability of transcription plus the probability of degradation. The Master Equation governing the evolution of $P(k,t)$ over time is therefore:

$$\frac{dP(k,t)}{dt} = \tau F(k-1)P(k-1,t) + \lambda(k+1)P(k+1,t) - (\tau F(k) + \lambda k)P(k,t). \tag{5.4}$$

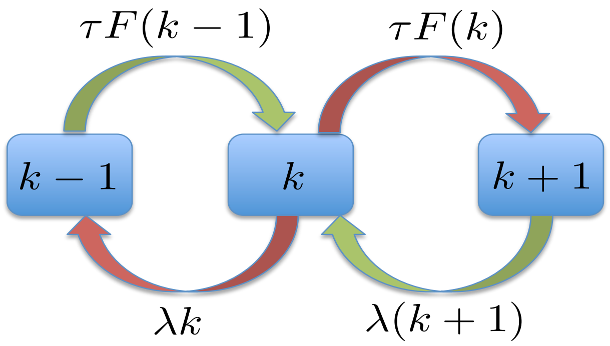
Explicit steady-state solution for one gene systems

A general single-species birth-and-death process governed by the Master equation

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

has an explicit steady-state probability distribution given by

$$p_k^s = \frac{g_0 g_1 \cdots g_{k-1}}{r_1 r_2 \ldots r_k} p_0, \tag{5.5}$$



$$\begin{aligned}\frac{dP(k)}{dt} = & \tau F(k-1) P(k-1) + \lambda(k+1) P(k+1) \\ & -(\tau F(k) + \lambda k) P(k)\end{aligned}$$

Figure 5.1: Informal derivation of the Master equation for gene regulation. In a infinitesimal timestep, $P(k,t)$, the probability of k RNA transcripts, increases by $P(k-1,t)$ times the probability, $F(k-1)$, of a transcription event (number of transcripts increases by one) plus $P(k+1,t)$ times the probability, $\lambda(k+1)$, of a degradation event (number of transcripts decreases by one). It decreases by $P(k)$ times the probability of transcription plus $P(k)$ time the probability of degradation.

(van Kampen VI.3.8 [VK07]). The proof is by induction. Applied to a single-gene system, this formula becomes

$$P^s(k) = P^s(0)\frac{(\Omega\tau/\lambda)^k}{k!}\prod_{j=0}^{k-1} f(\frac{j}{\Omega}). \tag{5.6}$$

This formula, which Chao Du independently derived and brought to my attention in October 2012, is very useful for studying one-gene systems with little computation. For example, it can be used to directly compute the steady-state mean and variance of a single-gene system.

Multiple genes

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

$$\frac{dP(\mathbf{k},t)}{dt} = \sum_{j=1}^{n} \left[\tau_j F_j(\mathbb{E}_j^- \mathbf{k}) P(\mathbb{E}_j^- \mathbf{k}, t) + \lambda_j (k_j + 1) P(\mathbb{E}_j^+ \mathbf{k}, t) - (\tau_j F_j(\mathbf{k}) + \lambda_j k_j) P(\mathbf{k}, t)\right]$$

$$\text{where } \mathbb{E}_j^{\pm}\mathbf{k} = \Big[k_1, \ldots, k_{j-1}, k_j \pm 1, k_{j+1}, \ldots, k_n\Big]^T. \tag{5.7}$$

RNA and protein

Initially, we simplified the discussion by ignoring protein translation and focusing only on the number of RNA transcripts of each gene. The same multivariate Master equation that allowed us to handle multiple genes also allows us to model the stochasticity of protein translation. If we introduce another a discrete random vector $\mathbf{Y}(t) \in \mathbb{Z}^n$, where $Y_j(t)$ denotes the number of protein translates of gene j, and define $P(\mathbf{k}^r, \mathbf{k}^p, t) \equiv \mathbb{P}(\mathbf{X}(t) = \mathbf{k}^r, \mathbf{Y}(t) = \mathbf{k}^p)$ the Master equation corresponding to the deterministic model (2.1) is:

$$\begin{aligned}\frac{dP(\mathbf{k}^r, \mathbf{k}^p, t)}{dt} = \sum_{j=1}^{n} \{ & \tau_j F_j(\mathbb{E}_j^- \mathbf{k}^p) P(k^r, \mathbb{E}_j^- \mathbf{k}^p, t) + \lambda_j^r (k_j + 1) P(\mathbb{E}_j^+ \mathbf{k}^r, \mathbf{k}^p, t) \\ & + r_j (k_j^r - 1) P(\mathbb{E}_j^- \mathbf{k}^r, \mathbf{k}^p, t) + \lambda_j^p (k_j^p + 1) P(\mathbf{k}^r, \mathbb{E}_j^+ \mathbf{k}^p, t) \\ & - (\tau_j F_j(\mathbf{k}^p) + \lambda_j^r k_j^r + r_j k_j^r + \lambda_j^p k_j^p) P(\mathbf{k}^r, k^p, t)\}.\end{aligned}$$

If we apply the QSSA to the translation step as described in appendix A6, that is, we assume that protein levels are approximately proportional to RNA levels at all times (which may or may not be biologically accurate, as discussed in chapter 2), then this equation reduces to the form (5.7).

5.1.2 Implicit assumptions and simplifications

Any modeling effort is necessarily a compromise between accuracy and tractability, and this case is no exception. Since the biological mechanisms of gene transcription are extraordinary complex and not completely understood, our model relies on a number of simplifying assumptions, both biological and physical in nature. One of the most explicit is the assumption that the rates of degradation, translation, and transcription (when RNAP is bound) are constant for each gene. In reality, the rates are affected by many other processes including chromatin remodeling, translational regulation, and protein folding. However, especially since it can be quite difficult to measure the rates accurately, the hope is that average rates suffice and the nonlinear form of f is flexible enough to capture much of the complexity. There are also a few key assumptions implicit in the form of f. One is that RNAP levels are approximately constant on the time scale of interest. A second is that nonspecific binding energy is equal for all non-promoter locations the RNAP could occupy. Finally, the model omits several reversible intermediate reactions such as binding and unbinding of RNAP and TFs. Since these reactions typically occur very quickly relative to the transcription time-scale, we can reasonably assume that the quantities involved are in thermodynamic steady-state, and apply Rao and Arkin's QSSA argument (appendix A6) to eliminate the reversible reactions from the model [RA03].

5.2 Expansion and simulation methods

The Master equation cannot be solved explicitly except in the simplest cases. For a one-gene system, we have an explicit formula for the steady-state distribution (equation 5.6), but no such formula exist for multiple genes. Therefore, in order to make further progress we will need approximations of the Master equation and efficient simulation methods. Fortunately, much of the work has already been done by physicists studying the Master equation, beginning in the 1970's. N. G. van Kampen ([VK07] and Ryogo Kubo ([KMK73]) developed a systematic expansion method for approximating the Master equation at any level of detail. Gillespie created a stochastic simulation algorithm to generate statistically correct trajectories of the Master equation; another simulation method based on the Langevin equation is less accurate but more efficient. We will summarize their findings in this section and show how they can be applied to the gene regulation problem. In the next chapter we will perform simulation studies on simple synthetic gene regulatory systems to illustrate the application of these methods and understand their strengths and weaknesses.

5.2.1 The Gillespie algorithm

The Gillespie algorithm enables numerical simulation of statistically correct trajectories of a system governed by the Master equation. The iterative Monte Carlo procedure randomly chooses the next event that will occur and the intervening time interval, then updates the molecular numbers of each species and the trajectory time [Gil77]. If the simulated system is in state $X(t)$ at time t, the waiting time τ before its next jump is drawn from an exponential distribution, and the probability of jumping to state $X^{(\mu)}$ is $w_\mu = W(X^{(\mu)}|X)$ (the Master equation transition probability for $X \to X^{(\mu)}$ per unit time). The basic steps of the algorithm are:

1. Initialize the molecular numbers of each species, $X_1, \ldots, X_n$, and set $t = 0$.

2. Randomly choose the next event to occur, and an exponential waiting time τ, by generating uniform random numbers u, v from Unif(0,1), and setting

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

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

$$t \leftarrow t + \tau, \qquad X(t) \leftarrow X_\mu.$$

4. Repeat steps 2-3 until the simulation time limit is reached ($t > T_{\text{sim}}$).

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

5.2.2 The van Kampen expansion

N. G. van Kampen provides a systematic approximation method involving an expansion in the powers of small parameter inversely related to system size [VK07]. The Master equation can be approximated at any level of detail by truncating the expansion to omit the higher-order terms. Ryogo Kubo, a contemporary of van Kampen, arrived at an equivalent formulation by a slightly different approach ([KMK73]). We will follow van Kampen's development here since it is more transparent.

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

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

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

$$\begin{aligned} W(X+1|X) = F(X) \equiv \Omega f(\frac{X}{\Omega}) &\iff \Phi_0(x;+1) = f(x) \\ W(X-1|X) = \lambda X \iff \Phi_0(x;-1) = \lambda x. \end{aligned}$$

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

$$X(t) = \Omega\phi(t) + \Omega^{\frac{1}{2}}\xi. \tag{5.8}$$

The motivation for the Ansatz is the observation that the relative fluctuation effects in chemical systems tend to scale as the inverse square root of the system size [Gil00]. It is justified *a posteriori* by the fact that $P(x,t)$, expressed in terms of ξ, turns out to be independent of Ω to first approximation. As part of the expansion procedure, ϕ is chosen to track the peak, and turns out be exactly the deterministic solution.

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

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

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

$$\alpha_\nu(x) = \int r^\nu \Phi(x; r) dr. \tag{5.9}$$

The first jump moment corresponds to the deterministic equation:

$$\frac{dy}{dt} = \alpha_1(y) = \frac{1}{\Omega}\int rW_\Omega(Y + r|Y)dr.$$

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

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

in our case $\alpha_1(y) = f(y) - \lambda y$, $\alpha_2(y) = f(y) + \lambda y$.

The complete calculation (adapted from chapter 10 of van Kampen) is provided in appendix A7. A crucial step in the expansion is the cancellation of terms of order $\Omega^{\frac{1}{2}}$, which cannot belong to a proper expansion for large Ω. The cancellation is made possible by choosing $\phi(t)$ (the macroscopic part of X) such that

$$\frac{d\phi}{dt} = \alpha_1(\phi).$$

That is, ϕ exactly satisfies the deterministic equation!

The final result (to order Ω^{-1}) is that

$$\begin{aligned}\frac{\partial \Pi}{\partial t} &= -\alpha_1'(\phi)\frac{\partial \xi \Pi}{\partial \xi} + \frac{1}{2}\alpha_2(\phi)\frac{\partial^2 \Pi}{\partial \xi^2} + \frac{1}{2}\Omega^{-\frac{1}{2}}(\alpha_2'(\phi)\frac{\partial^2 \xi \Pi}{\partial \xi^2} \\ &- \alpha_1''(\phi)\frac{\partial \xi^2 \Pi}{\partial \xi} - \frac{1}{3!}\alpha_3(\phi)\frac{\partial^3 \Pi}{\partial \xi^3}) + O(\Omega^{-1}) \qquad (5.10)\end{aligned}$$

with jump moments α_ν defined by (5.9).

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

5.2.3 The linear noise approximation

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

$$\frac{\partial \Pi}{\partial t} = -\alpha_1'(\phi)\frac{\partial \xi \Pi}{\partial \xi} + \frac{1}{2}\alpha_2(\phi)\frac{\partial^2 \Pi}{\partial \xi^2} + O(\Omega^{-\frac{1}{2}}). \qquad (5.11)$$

This is a linear Fokker-Planck equation, and the solution turns out to be a Gaussian (see van Kampen VIII.6 [VK07]). Hence it is completely characterized by the first and second moments of ξ, which are of the most interest to us anyway. Multiplying equation 5.11 by ξ and ξ^2, respectively, yields

$$\frac{\partial}{\partial t}\langle\xi\rangle = \alpha_1'(\phi)\langle\xi\rangle \tag{5.12}$$

$$\frac{\partial}{\partial t}\langle\langle\xi\rangle\rangle = 2\alpha_1'(\phi)\langle\langle\xi\rangle\rangle + \alpha_2(\phi). \tag{5.13}$$

After solving for $\langle\xi\rangle, \langle\langle\xi\rangle\rangle$ and solving the deterministic equation for ϕ, we can use the Ansatz (5.8) to find the mean and variance of X:

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

The initial condition $P(X,0) = \delta(X - X_0)$ implies $\phi_0 = x_0$ and $\langle\xi\rangle_0 = \langle\langle\xi\rangle\rangle_0 = 0$, hence $\langle\xi\rangle_t \equiv 0$. (Even if ξ has a nonzero initial distribution, if $\alpha_1'(\phi) < -\epsilon < 0$ we still will have $\langle\xi\rangle \leq e^{-\epsilon t} \to 0$.) Hence the mean of the solution to the Master equation with initial distribution δ_{x_0} approximately satisfies the deterministic equation:

$$\frac{\partial}{\partial t}\langle x\rangle = \alpha_1(\langle x\rangle) + O(\Omega^{-1}). \tag{5.14}$$

5.2.4 Connection to the nonlinear deterministic model

Equation 5.14 provides the link between the stochastic Master equation and the nonlinear deterministic dynamical system model of chapter 4. It shows that the deterministic equation is an approximate model for the evolution of the mean expression of the stochastic process. That is,

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

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

5.2.5 The Fokker-Planck and Langevin equations

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

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

$$\frac{\partial P(y,t)}{\partial t} = -\frac{\partial}{\partial y}\alpha_1(y)P + \frac{1}{2}\frac{\partial^2 \Pi}{\partial y^2}\alpha_2(y)P. \tag{5.15}$$

In the general form of the equation, α_1, α_2 are any real differentiable functions with $\alpha_2 > 0$, but in Planck's derivation of the equation as an approximation to the Master equation [PG58], they are exactly the first and second jump moments (5.9). Since the Fokker-Planck equation is always linear in P, we follow van Kampen in appropriating the term *linear* to mean that α_1 is linear and α_2 constant.

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

$$dy = \alpha_1(y)dt + \sqrt{\alpha_2(y)}dW, \tag{5.16}$$

where $W(t)$ is a Wiener process, or Brownian motion. (Again, $\alpha_1, \alpha_2 > 0$ may be any C^1 functions in general, but in the case of interest to us, they represent the jump moments.) Equations 5.15 and 5.16 are mathematically equivalent using the Ito intepretation of 5.16 (see van Kampen IX.4 [VK07] for the proof).

These equations are very appealing for modeling physical processes since they are easy to derive and interpret. For both equations, α_1, α_2 (thought of for now as general functions, not as the jump moments) can be inferred without even knowing the underlying Master equation, using only the macroscopic law and fluctuations around the steady-state solution (known from statistical mechanics). The approach works very well in situations where the macroscopic law α_1 is linear [PG58, Ray91, Ein06, VS06]. However, confusion can arise when α_1 is nonlinear, since effects on the order of the fluctuations are invisible macroscopically [VK65]. One of the major motivations for van Kampen's systematic expansion was the need to resolve disagreements between authors who had developed different but equally plausible characterizations of the noise in nonlinear systems using this approach.

For systems with linear deterministic equations, the van Kampen approximation agrees exactly with the Fokker-Planck model, since the linear noise approximation yields a linear Fokker-Planck equation. However, discrepancies may arise for nonlinear systems, and we should consider the van Kampen theory definitive in such cases. The error in the nonlinear Fokker-Planck model is that it retains the full functional dependence on the nonlinear functions α_1, α_2 (in effect, keeping infinitely many terms of their Taylor expansions) while cutting off their third-order and higher derivatives in the expansion about the deterministic path $\phi(t)$. In contrast, the truncated van Kampen expansion replaces α_1, α_2 by their Taylor polynomials at a level of detail consistent with the order of the approximation. The van Kampen expansion provides a completely consistent approximation of the Master equation to any desired order of accuracy, while the Fokker-Planck model is a slightly inconsistent second-order approximation only. Nevertheless, the discrepancy between the Fokker-Planck and van Kampen approximations is often not too serious (and a second-order approximation is typically good enough), so the model are still very useful in many cases.

The Langevin equation, in particular, lends itself to efficient simulation [KH08, Gil00]. Simulation provides insight into the behavior of individual trajectories as well as moment information, and applies directly to multistable systems (while van Kampen requires alternative theory since the expansion method only applies to systems with one stable steady-state). However, simulation can be very expensive. The exact Gillespie algorithm and other direct simulation methods are only computationally feasible for very small systems. Fortunately, the Langevin simulation works well for large systems with many genes, since trajectories of the Langevin equation can be simulated by evolving a small system of stochastic differential equations, rather than accounting for every single reaction like the Gillespie algorithm. Hence the Langevin simulation is appropriate for large systems with complex qualitative structures.

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

$$\alpha_1(y) = f(y) - \lambda y, \qquad \alpha_2(y) = f(y) + \lambda y,$$

the Langevin equation is:

$$dx = (f(x) - \lambda x)dt + \sqrt{f(x)}dW_1 + \sqrt{\lambda x}dW_2, \tag{5.17}$$

where W_1, W_2 are independent Wiener processes.

5.3 Systems with multiple stable steady-states

We have alluded several times to the fact that stochasticity can lead to unexpected results for systems with multiple stable steady-states. The basic reason is that random fluctuations can send stochastic trajectories out of the domain of attraction of one deterministic steady-state and into the domain of another. van Kampen treats these issues in detail in chapter XIII of his book [VK07]. In this section, we will summarize the points that are most relevant to our topic. In the next chapter, simulation studies will illustrate these points and provide additional insight.

For simplicity, consider a birth-and-death process with two distinct stable steady-states, $\phi_a < \phi_c$ and an unstable steady-state ϕ_b ($\phi_a < \phi_b < \phi_c$). By this we mean that the corresponding deterministic equation $d\phi/dt = \alpha_1(\phi)$ has those properties, i.e.

$$\alpha_1(\phi_a) = \alpha_1(\phi_b) = \alpha_1(\phi_c) = 0, \tag{5.18}$$

$$\alpha_1'(\phi_a) < 0, \quad \alpha_1'(\phi_b) > 0, \quad \alpha_1'(\phi_c) < 0. \tag{5.19}$$

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

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

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

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

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

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

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

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

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

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

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

We can identify the escape time τ_{ca} with the mean first-passage time from ϕ_a to ϕ_c. For the one-dimensional process defined by equation 5.3, the mean first-passage

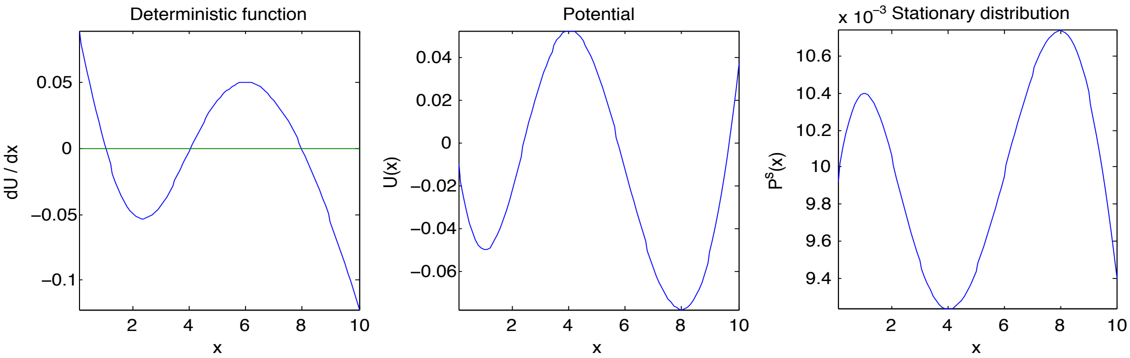


Figure 5.2: Bistability in a stochastic system modeled by a Fokker-Planck equation of the form (5.21), corresponding to deterministic equation $\frac{dx}{dt} = \frac{dU}{dt}$. The deterministic function $\frac{dU}{dt}$ (left) has zeros at the three steady-states $\phi_a \approx 1, \phi_b \approx 4, \phi_c \approx 8$. The points ϕ_a and ϕ_c are stable, while ϕ_b is unstable. The potential $U(x)$ (center) has minima at ϕ_a and ϕ_b and a maximum at ϕ_c, corresponding to low energy (favorable) at the two steady-states and high energy (unfavorable) at the unstable state. ϕ_b is more stable than ϕ_a since its potential well is deeper and wider. The stationary distribution (right), to which the stochastic system will eventually converge, is bimodal with peaks at ϕ_a and ϕ_c. The peak at ϕ_c is higher since ϕ_c is more stable than ϕ_a.

time from ϕ_a to ϕ_c is given by

$$\tau_{ca} = \sum_{k=a}^{c-1} \frac{1}{g_k p_k^s} \sum_{j=0}^{k} p_j^s, \quad \text{where } p^s \text{ is the stationary distribution (5.5),} \tag{5.20}$$

as shown in appendix A8. The escape rate is $O(p_b^s)$, the height of stationary distribution at the unstable point b, so the escape time scales exponentially in the system size [BHK98].

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

$$\frac{\partial P(x,t)}{\partial t} = \frac{\partial}{\partial x} U'(x) P + \theta \frac{\partial^2 P}{\partial x^2}. \tag{5.21}$$

Although this model is not even approximately appropriate for the gene regulation problem since the diffusion coefficient is constant, while in the gene regulation problem it is a function of x, it helps clarify some important issues. To that end, suppose the derivative of U satisfies the bistability conditions (5.19), so that dU/dx and U have the shapes shown in Figure 5.2. dU/dx has zeros at the steady-states ϕ_a, ϕ_b, ϕ_c, and U has minima at the stable points ϕ_a, ϕ_c and a maximum at the unstable point ϕ_b.

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

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

and for small θ we can approximate

$$\begin{aligned}
&C^{-1} \approx e^{-U(a)/\theta}\sqrt{\frac{2\pi\theta}{U''(a)}} + e^{-U(c)/\theta}\sqrt{\frac{2\pi\theta}{U''(c)}} \\
&\pi_a^s \approx \int_{-\infty}^{b} P^s(x)dx = C\sqrt{\frac{2\pi\theta}{U''(a)}}, \quad \pi_c^s \approx \int_{b}^{\infty} P^s(x)dx = C\sqrt{\frac{2\pi\theta}{U''(c)}} \\
&\frac{\pi_a^s}{\pi_c^s} \approx e^{-(U(a)-U(c))/\theta}\sqrt{\frac{U''(c)}{U''(a)}} \qquad \text{[van Kampen XIII.1.10 – 1.11]}.
\end{aligned}$$

Hence the relative stability of the two stable steady-states depends on both the depths of the potential energy wells ($U(a)$ and $U(c)$) and their widths ($U''(a)$ and $U''(c)$). In Figure 5.2, ϕ_c is more stable than ϕ_a, since its potential energy is lower and energy well is wider. The relative stability in this example is about $\frac{\pi_a}{\pi_c} = 0.76$, meaning that at stochastic steady-state, about 43% of trajectories will be near ϕ_a and 57% will be near ϕ_c at a given time (as shown in Figure 5.2, right pane).

Similarly, we can approximate the escape time (mean first-passage time) by

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

Hence the escape time depends on the height of the energy barrier $U(b)$ and energy well $U(a)$, and the widths $U''(a), U''(b)$ of the well and barrier. Since the potential energy difference is $O(\Omega)$, we see again that the escape time scales exponentially with the system size.

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

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

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

$$P^s(y) = \frac{C}{\alpha_2(y)}\exp\left(2\int_0^y \frac{\alpha_1(t)}{\alpha_2(t)}dt\right). \qquad \text{[van Kampen VIII.1.4]} \tag{5.22}$$

We can define an "effective potential" by

$$U_{\text{effective}} = -2\int_0^y \frac{\alpha_1(t)}{\alpha_2(t)}dt + \log(\alpha_2(y). \tag{5.23}$$

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

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

For the escape time, van Kampen (XII.3.7) gives:

$$\tau_{ca} = \int_a^c e^{\Phi(y')}dy' \int_a^{y'} e^{-\Phi(y'')}\frac{2dy''}{\alpha_2(y'')}, \quad \Phi(y) = -\int_a^y \frac{2\alpha_1(y')}{\alpha_2(y')}.$$

In the one-dimensional case, we could use the exact steady-state solution of the Master equation (5.6) and the escape time (5.20) instead, although the Fokker-Planck stationary solution may be more convenient. In multivariate case, we must the the Fokker-Planck approach, since there is no general approach to finding stationary solutions of multivariate Master equations. In contrast, finding the stationary solution (if it exists) of a Fokker-Planck equation amounts to solving a second-order partial differential equation, which is straightforward numerically and sometimes even explicitly.

5.4 Summary

Nonlinear Master equation models capture the stochastic mechanisms of gene regulation in full molecular detail. The Master equation can rarely be solved explicitly for multiple gene systems, but theoretical approximations and simulation algorithms can give insight into these systems. The Gillespie algorithm allows us to numerically simulate exact trajectories of the Master equation, although the computational cost becomes prohibitive for large systems with many genes. The van Kampen expansion method allows us to rigorously approximate the Master equation at any level of detail we desire (the deterministic model (2.1) being the simplest), provided the system has only one stable steady-state, and van Kampen provides alternative theory for analyzing systems with multiple stable steady-states. The Langevin equation (equivalent to the Fokker-Planck equation) is an inexact approximation to the Master equation and is the basis of a highly efficient simulation method that is well-suited for large multiple-gene systems. In the next chapter, we will perform simulation studies on simple synthetic gene regulatory systems to illustrate the application of each of these methods and evaluate their performance. As one might expect, the behavior of systems with multiple stable steady-states is particular interesting.

Chapter 6

Stochastic simulation studies

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

6.1 One gene system with one stable steady-state

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

$$\frac{dy}{dt} = f(y) - \gamma y, \quad f(y) = \frac{2\gamma}{1+y}, \quad \gamma = 0.1. \tag{6.1}$$

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

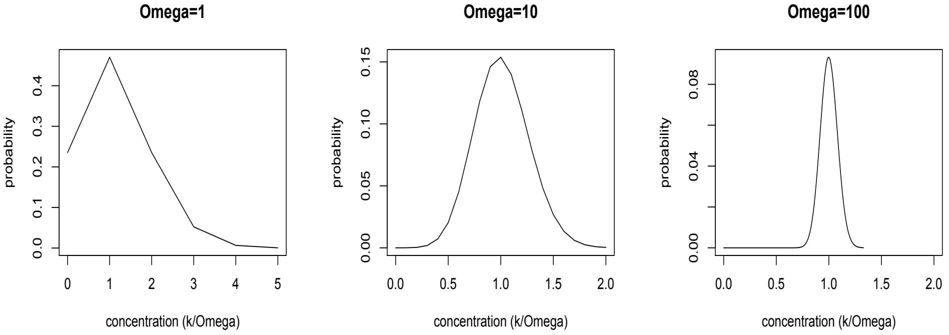


Figure 6.1: Steady-state probability distributions of a one gene system with one steady-state (6.1) for increasing system sizes $\Omega = 1, 10, 100$. The distribution always peaks at the deterministic steady-state solution ($y = 1$), and the variance decreases as Ω increases. For smaller values of Ω, it's clear that the mean lies slightly above the deterministic solution, but as Ω increases, the distribution becomes quite symmetric.

The corresponding Master equation is

$$\frac{dP(k)}{dt} = F(k-1)P(k-1) + \gamma(k+1)P(k+1) - (F(k) + \gamma k)P(k), \tag{6.2}$$

where $F(k) = \Omega f(k/\Omega)$. Numerical evolution of the Master equation by iteratively updating a vector of probabilities according to (6.2) is feasible in this case because the system is so simple. The Master equation also has the explicit steady-state solution :

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

Figure 6.1 shows the stationary probability distributions for a range of values of Ω, revealing that as Ω increases, the distribution is increasingly sharply peaked at $y^s = 1$. That is, the mean approaches $y^s = 1$, and the variance goes to zero as Ω increases. To second order, the van Kampen expansion gives:

$$\begin{aligned}\frac{d\phi}{dt} &= \alpha_1(\phi) = f(\phi) - \gamma\phi \\ \frac{d\langle\xi\rangle}{dt} &= \alpha_1'(\phi)\langle\xi\rangle + \frac{1}{2}\Omega^{-\frac{1}{2}}\alpha_1''(\phi)\langle\xi^2\rangle = (f'(\phi)-\gamma)\langle\xi\rangle + \frac{1}{2}\Omega^{-\frac{1}{2}}f''(\phi)\langle\xi^2\rangle \\ \frac{d\langle\xi^2\rangle}{dt} &= 2\alpha_1'(\phi)\langle\xi^2\rangle + \alpha_2(\phi) = 2(f'(\phi)-\gamma)\langle\xi^2\rangle + (f(\phi)+\gamma\phi).\end{aligned}$$

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

$$f(\phi^s) = 0.2(1+\phi^s)^{-1} = 0.1; \quad f'(\phi^s) = -0.05; \quad f''(\phi^s) = 0.05$$

and plugging into the last two equations yields

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

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

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

The Langevin model for this system is given by the SDE

$$dX = (F(X) - \gamma X)dt + \sqrt{F(X)}dW_1 + \sqrt{\gamma X}dW_2,$$

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

Figure 6.2 compares the exact Master equation, second-order van Kampen approximation, Gillespie simulation, and Langevin simulation for this system with initial condition $y^s = 1$ (the steady-state value) and three different values of Ω. As Ω increases, the agreement improves as the mean approaches the deterministic trajectory (that is, the steady-state value $y^s = 1$), and the variance decreases. The discrepancy between the stochastic mean and the deterministic trajectory and the variance are both $O(\Omega^{-1})$ (as predicted by the van Kampen expansion).

The Master equation governs the evolution of the probability distribution; Figure 6.3 shows the final probability distributions for each value of Ω. In each case, the initial probability is a delta-distribution centered at Ωy^s, and the probability spreads out over time to reach a steady-state distribution, which is extremely close to a Gaussian for $\Omega >> 1$. For larger values of Ω, the final probability distribution remains sharply peaked around y^s.

6.2 Two gene system with one stable steady-state

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

$$\frac{dy_1}{dt} = f_1(y) - \gamma y_1, \quad f_1(y) = \frac{0.1 + 0.1y_2}{1 + y} \tag{6.3}$$

$$\frac{dy_2}{dt} = f_2(y) - \gamma y_2, \quad f_2(y) = \frac{0.4 + 0.1y_1y_2}{1 + y_1y_2}, \quad \gamma = 0.1. \tag{6.4}$$

It has a single deterministic steady-state at $y_1 = 1, y_2 = 2$. With two genes, directly evolving the Master equation is very expensive for moderately sized systems, as each probability distribution is now two-dimensional (in general, the computational cost of evolving the Master equation with system size Ω is $O(\Omega^n)$ per timestep), so we omit this method and focus on the van Kampen approximation and the Gillespie and Langevin simulations. Figure 6.4 shows that the situation is qualitatively very similar to the one gene case we just discussed. The approximation and simulation means differ from the deterministic trajectory by $O(\Omega^{-1})$, and the variance is also $O(\Omega^{-1})$. For $\Omega = 1$, the y_2 variance and mean discrepancy of the Langevin simulation and the van Kampen approximation are slightly lower than those of the exact Gillespie trajectories. This inaccuracy arises from zero-boundary effects and the non-Gaussianity of the probability distribution at small system sizes (on the order of a single molecule).

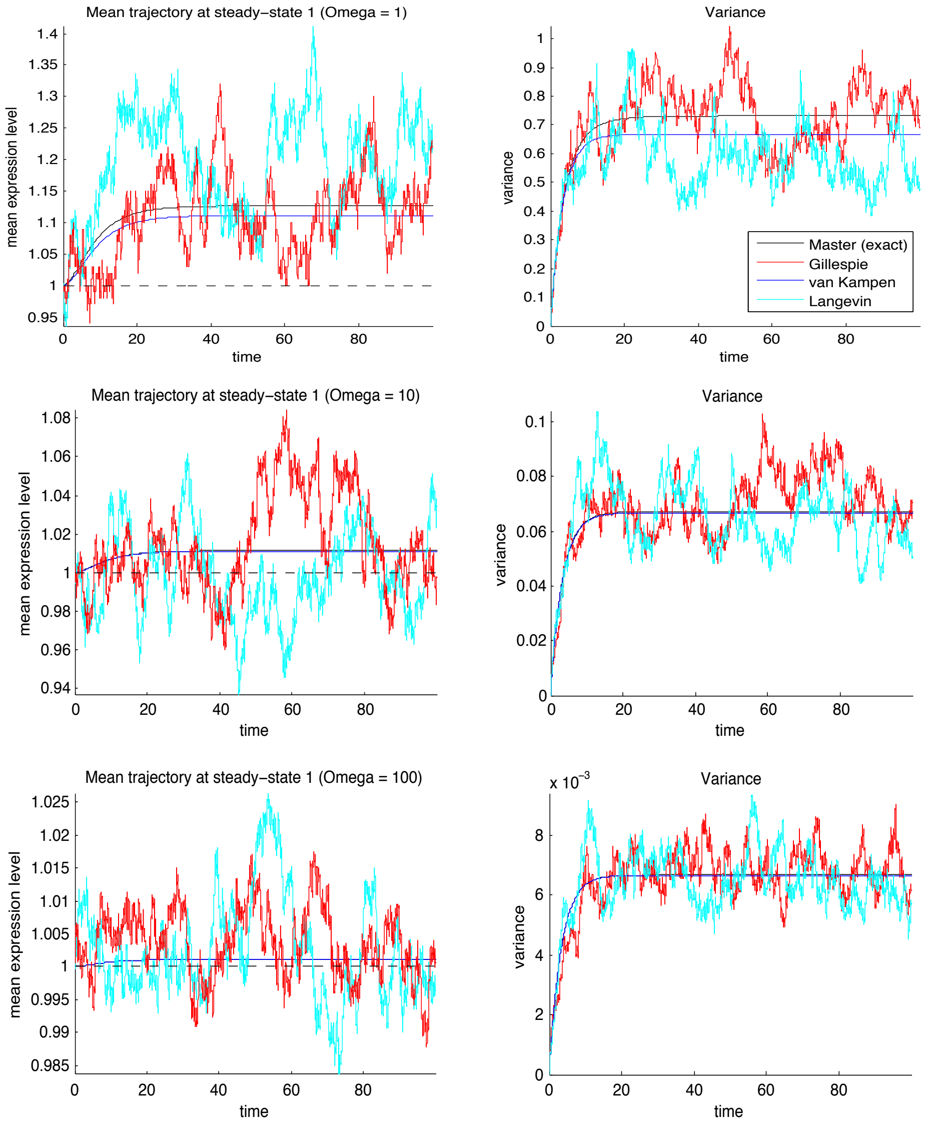


Figure 6.2: One gene system with one steady-state (6.1). Mean (left) and variance (right) trajectories via Master equation (black), van Kampen approximation (blue), and average of 100 trajectories of the Gillespie (red) and Langevin simulation (cyan) with $\Omega = 1, 10, 100$ (top to bottom, respectively). Excellent agreement between simulations, van Kampen approximation, and exact Master equation for both mean and variance. Discrepancy between the stochastic mean and deterministic trajectory and magnitude of the variance are both $O(\Omega^{-1})$.

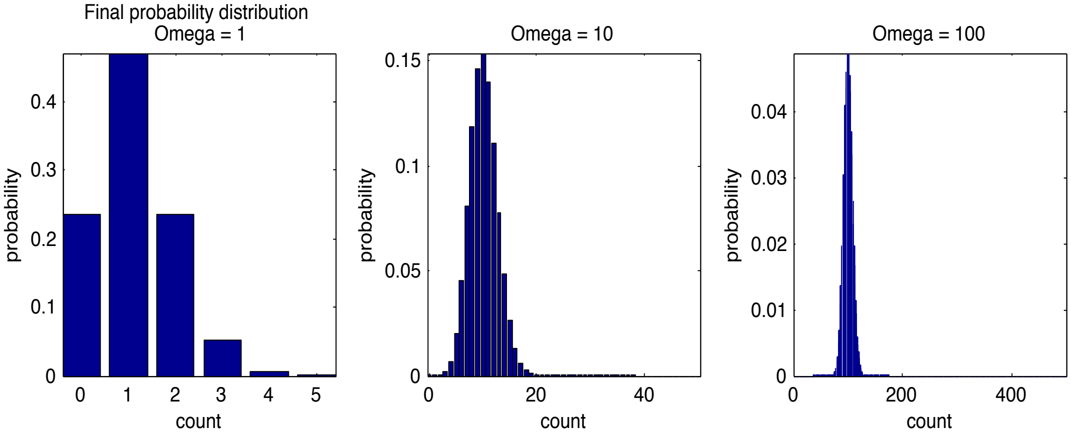


Figure 6.3: Final probability distributions of the exact Master equation for a one gene system with one steady-state (6.1), with $\Omega = 1, 10, 100$. The probabilities converge to approximately Gaussian steady-state distributions peaked near the deterministic steady-state. For larger system sizes, the distribution is more Gaussian and the peak is sharper.

6.3 Constructing multistable systems

Gene regulatory systems with multiple stable steady-states are ubiquitous in nature as this property plays a key role in cellular lifecycles and responses to external stimuli. However, constructing synthetic systems with multiple stable steady-states with our chosen functional form (2.8) can be challenging. One approach, which Chickarmane et al used to develop their ESC-inspired system ([CP08]), is to start with a well-understood biological network with multiple steady-states and use experimental data and knowledge of qualitative behavior to suggest the appropriate terms and parameter values. This can be an interesting and useful program, especially as the synthetic network may later be useful for gaining further insight into the behavior of the original biological network; however, there are very few biological networks well-understood enough to lend themselves to this type of modeling. Furthermore, it is limiting in the sense that it relies on existing known networks, and provides little insight into methods for generating original networks. Ideally, we would like to be able to create novel networks from scratch with specific properties of our own choosing. In this

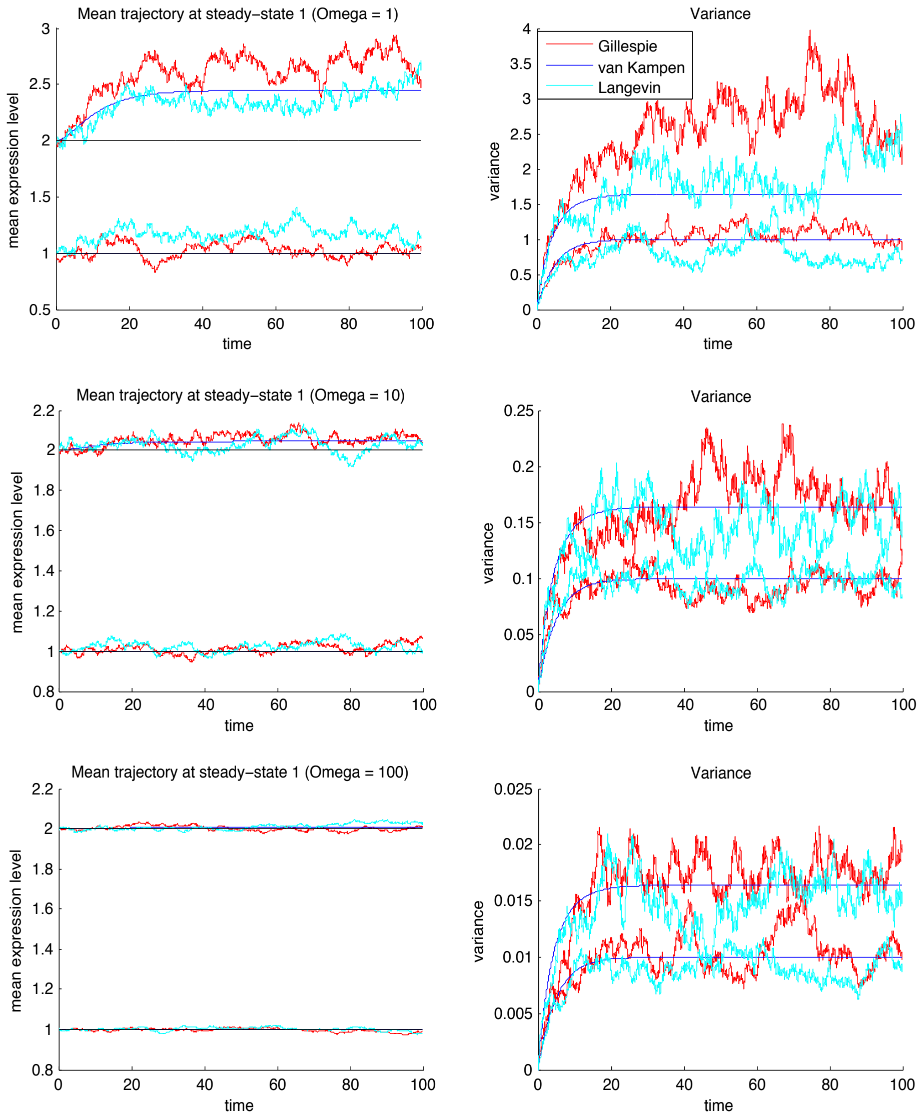


Figure 6.4: Two gene system with one stable steady-state (6.4). Mean (left) and variance (right) trajectories of van Kampen approximation (blue) and average of 100 trajectories of Gillespie (red) and Langevin simulation (cyan) with $\Omega = 1, 10, 100$ (top to bottom, respectively). As for the one-gene system, agreement between the simulations and the van Kampen approximation is excellent, and both the variance and the discrepancy between the mean and deterministic trajectory are $O(\Omega^{-1})$. The only exception is for $\Omega = 1$, where slight inaccuracy of the Langevin simulation and van Kampen expansion arises from the non-Gaussianity of the probability distribution.

section we will discuss our efforts toward this end. Although we have not fully solved this problem by any means and would encourage further work in this direction, we have developed a heuristic algorithm that, together with some trial-and-error, allowed us to generate the two multistable synthetic gene networks we study later in this chapter.

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

$$f_i(y) = \frac{b_{i0} + \sum_{j=1}^{m} b_{ij} \Pi_{k \in S_{ij}} y_k}{1 + \sum_{j=1}^{m} c_{ij} \Pi_{k \in S_{ij}} y_k} \implies f_i(e_j) - \gamma e_{1j,i} = 0, \quad 1 \le i \le n, \quad 1 \le j \le m.$$

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

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

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

Hence, we wish to find b_{ij}, c_{ij} such that $f_i(e_j) = \gamma e_{j,i}$ while satisfying the Jacobian condition and the other constraints. That is, we want to solve the feasibility problem:

$$\begin{aligned} \text{find} \quad & b_i, c_i \\ \text{subject to} \quad & f_i(e_j) = \gamma e_{j,i} \\ & 0 \le b_i \le c_i, \quad c_i(0) = 1, \\ & \exists P_j : J_f(e_j)^T P + P J_f(e_j) \prec -\epsilon, \quad 1 \le j \le m. \end{aligned} \tag{6.5}$$

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

The problem is nonconvex due to the rational form of f, so we can either use a nonconvex solver, or use heuristics and trial-and-error and solve with a convex solver. Specifically, we can we can use the iterative approach described in chapter 2 to enforce the stability constraint, and simply replace the denominator of each f

with a constant value and add a constraint forcing the denominator to be equal to that constant. Of course, not all constant values lead to feasible problems, so if we use the heuristic approach, we must guess-and-check the denominator values as well as the steady-state locations.

6.4 One gene system with two stable steady-states

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

$$\frac{dy}{dt} = f(y) - \gamma y, \quad f(x) = \frac{0.1 + x + 0.1x^4}{1 + 10x + 0.5x^2 + 0.1x^4}, \quad \gamma = 0.1 \tag{6.6}$$

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

At the end of the last chapter, we discussed methods from chapter XIII of van Kampen's book for analyzing the equilibrium behavior of systems with multiple stable steady-states. These tools provide a great deal of insight into long-term system behavior with minimal computation, since they only require the stationary probability distribution (which can be computed directly in the single-gene case using (5.6), or approximated using the Fokker-Planck equation in general). These tools will allow us to predict some of the basic behavior of system (6.6) with very little effort. Simulations will confirm and complete the picture.

Let us first examine the most basic properties of the system with $\Omega = 1$. As Figure 6.5 shows, the deterministic system is bistable. The deterministic function $\alpha_1(x) = f(x) - \gamma x$, has three zeros corresponding to the three deterministic steady-states. The derivative of the deterministic function is negative ($d\alpha_1/dt < 0$) at the stable steady-states, and positive at the unstable steady-state. The stationary distribution has a strong peak at the more stable steady-state, e_1, and a weaker one at the less-stable point e_2. The system is three times as likely to be in the domain of e_1 than in the domain of e_2. We can use the relative stability of the two stable points

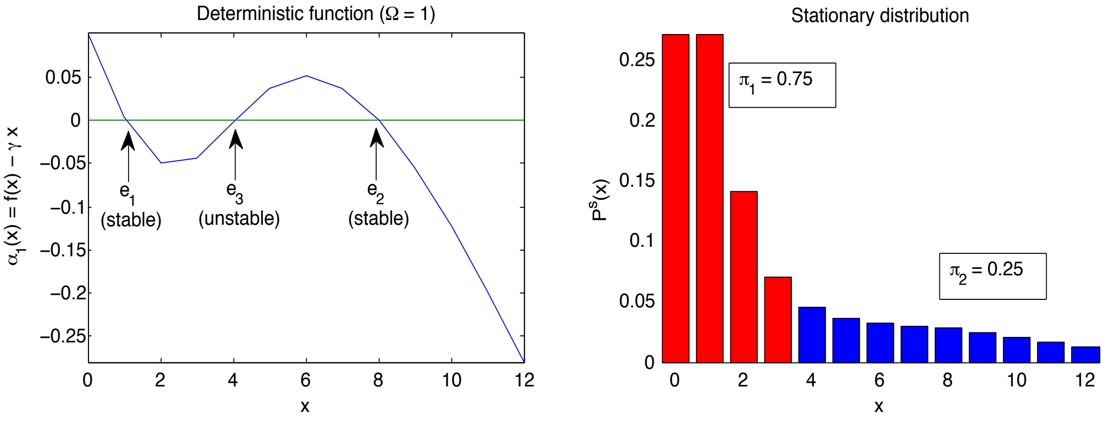


Figure 6.5: The deterministic function $\alpha_1(x) = f(x) - \gamma x$ for the system (6.6) with $\Omega = 1$, has three zeros corresponding to the three deterministic steady-states, e_1, e_2, e_3. The derivative of the deterministic function is negative $(d\alpha_1/dt < 0)$ at the stable steady-states e_1, e_2, and positive at the unstable steady-state e_3. The stationary distribution (computed with equation (5.6)) has a strong peak at e_1 and a weaker one at e_2. The system is much more likely to be in the domain of e_1 $(x < e_3)$ than in the domain of e_2 $(x > e_3)$: specifically, $\pi_1 \approx 0.75$, and $\pi_2 \approx 0.25$. The steady-state mean is given by $\pi_1 e_1 + \pi_2 e_2 \approx 2.78$.

to estimate the steady-state mean: $\pi_1 e_1 + \pi_2 e_2 \approx 2.78$, which will be confirmed by our simulation study.

Next, let us examine the system with $\Omega = 10$. Figure 6.6 shows the deterministic function, the effective potential, and the (approximate) stationary distribution, computed using the Fokker-Planck approach. The deterministic function and stationary distribution have the same qualitative properties as they did for $\Omega = 1$, except the e_1 peak of the stationary distribution is now even higher relative to e_2 $(\pi_1 \approx 97\%, \pi_2 \approx 3\%)$ and the steady-state mean, 1.24 is therefore closer to e_1. The effective potential has minima at the stable steady-states, but the "energy" of the more stable state, e_1, is much lower.

Simulations reveal how the mean, variance, and probability distribution of the system actually evolve. Figures 6.7, 6.10 and 6.11 compare the exact Master equation, second-order van Kampen approximation, and Master equation and Langevin

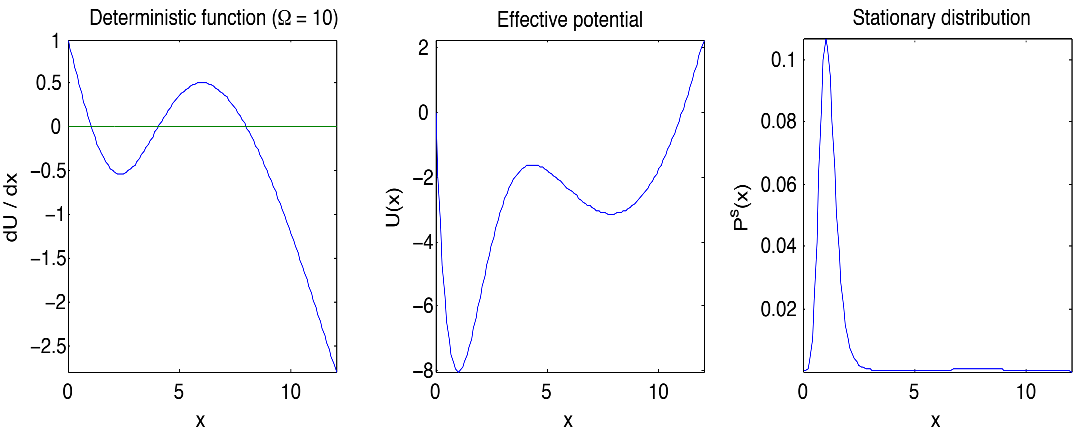


Figure 6.6: The deterministic function, effective potential, and (approximate) stationary distribution for system (6.6) with $\Omega = 10$, computed with the Fokker-Planck approximation and equations (5.22, 5.23). (The result is nearly identical to what we would have obtained with the explicit equation (5.6)). The deterministic function and stationary distribution have the same qualitative properties as they did with $\Omega = 1$, except the e_1 peak in the stationary distribution is now even higher relative to e_2 ($\pi_1 \approx 97\%, \pi_2 \approx 3\%$), and the steady-state mean is shifted toward e_1: $\pi_1 e_1 + \pi_2 e_2 \approx 1.24$. The effective potential has minima at the two stable steady-states e_1, e_2, and a maximum at e_3. The more stable steady-state, e_1, has lower "energy".

simulations for $\Omega = 1, 10$, and all but the exact Master equation for $\Omega = 100$ (due to instability), respectively. Unlike for the one gene system described by equation (6.1), the exact Master equation and both simulations deviate dramatically from both the van Kampen approximation and the deterministic trajectory, at least for $\Omega = 1, 10$. The reason for this is the bistability of the system. Especially when Ω is fairly small (hence the variance is relatively large) each stochastic trajectory starting from steady-state e_1 has a reasonably large probability of escaping from the domain of attraction of e_1 and being attracted to e_2, and vice versa. In the long run, the system settles to a bimodal steady-state distribution, in which both stable steady-states are represented proportional to their relative stability. Therefore, the steady-state mean regardless of the starting point converges to the roughly the weighted average of the two deterministic stable steady-states predicted by the basic stability analysis described in Figure 6.5. The second-order van Kampen expansion centered at either of the two steady-states does not account for this blending effect and therefore underestimates both the variance and the deviation of the mean trajectory from the deterministic trajectory. In reality, the second-order expansion should never have been applied in this case since it is only valid for systems with a single stable steady-state, as van Kampen explains in chapter X of his book ([VK07]).

Figure 6.8 shows how the probability distribution evolves from two different initial conditions, peaked at e_1 and e_2, respectively. Regardless of the starting point, the probability distributions eventually converge to identical steady-state distributions, with a strong, sharp peak near e_1 and a weaker peak centered near e_2. When the initial condition is a peak at e_1, the probability spreads out over time and shifts some of its weight toward e_2, and vice-versa, although much more weight is shifted from e_2 to e_1 than the other direction.

The system behavior with $\Omega = 10$ is qualitatively similar, as Figure 6.10 shows, but the bimodal steady-state probability distribution is even more sharply peaked at e_1 and the stochastic mean converges to an an average closer to e_1, in agreement with the analysis of Figure 6.6.

The situation appears to be different for $\Omega = 100$, as shown in Figure 6.11. In fact, the system seems to behave much more like a single stable steady-state system, in that the stochastic mean remains close to the initial steady-state, the van Kampen approximation agrees well with the simulation results, and the variance and the difference between the mean and deterministic trajectory are both on the order of $O(\Omega^{-1})$. The explanation is that for very large systems, the probability of a jump between e_1 and e_2 is extremely small, so the escape time is much longer than the length of the simulation. Figure 6.4 confirms that the escape time scales exponentially with the system size, as discussed in chapter 5 and appendix A8. Therefore, for a large system like this one, the stochastic trajectories are highly unlikely to diverge from the deterministic steady-state where they originated for the duration of the simulation. If the simulation ran long enough, some trajectories would eventually escape from their initial domains of attraction, and the same blending of the two steady-states that we observed in the smaller systems would occur. The large system size means that the initial time period in which the two stable steady-states operate independently of each other takes up the entire simulation, however, so we never observe this blending.

6.5 Two gene system with two stable steady-states

We constructed a two-gene system with two stable steady-states using the heuristic approach described in section 6.3. The two steady-states are:

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

The deterministic model is

$$\begin{aligned} \frac{dy_i}{dt} &= f_i(y) - \gamma y_i, \quad f_i(y) = \frac{b^{(i)T} z(x)}{c^{(i)T} z(x)}, \quad i = 1, 2 \\ z(x) &= \begin{bmatrix} 1 & x_1 & x_2 & x_1 x_2 & x_1^2 & x_2^2 \end{bmatrix}^T, \quad \gamma = 0.1, \end{aligned} \tag{6.7}$$

with b_i, c_i, $i = 1, 2$, given in appendix A9.

Figure 6.12 compares the second-order van Kampen approximation, and the Gillespie and Langevin simulations for $\Omega = 10$ and $\Omega = 1000$. The qualitative behavior of this system is exactly the same as that of the bistable one gene system. For small system sizes, each stochastic trajectory has a reasonable probability of escaping from the domain of attraction of one stable steady-state and being attracted to the other, so in the long run, the system settles to a bimodal steady-state distribution. Hence, regardless of the initial condition, the steady-state mean converges to a weighted average of the two deterministic stable steady-states. The second-order van Kampen approximation centered at either of the two steady-states does not properly apply, and would seriously underestimate both the variance and the deviation of the mean trajectory from the deterministic trajectory. For very large systems, in contrast, the probability of a giant fluctuation between e_1 and e_2 is very small. Since the escape time scales exponentially with the system size, it can far exceed the length of the simulation for large systems. Therefore, the stochastic trajectories remain close to the deterministic steady-state where they originated for the duration of the simulation,

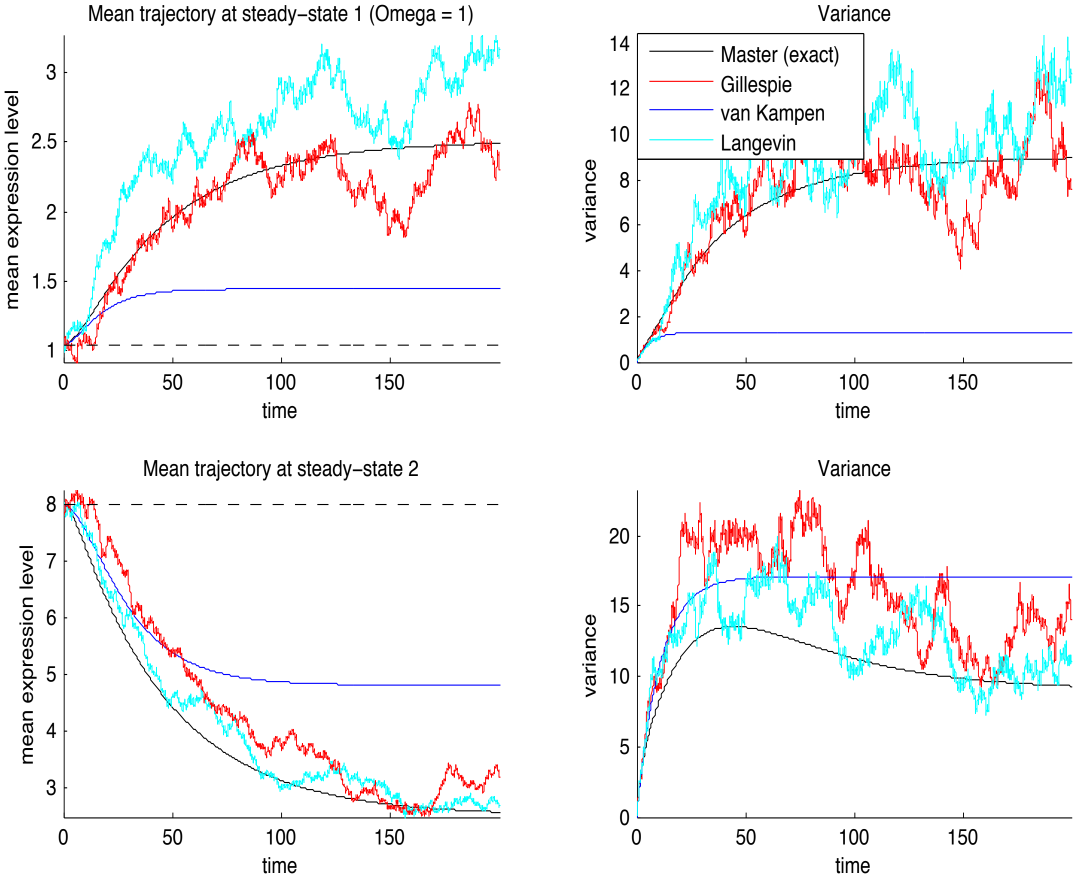


Figure 6.7: One gene system with two stable deterministic steady-states (6.6), $\Omega = 1$. Mean (left) and variance (right) trajectories via the Master equation (black), the (improperly applied) van Kampen expansion (blue), and the average of 100 trajectories of the Gillespie (red) and Langevin simulation (cyan). Regardless of the starting point, the stochastic mean trajectory eventually converges to the weighted average of the two deterministic stable steady-states predicted by the analysis of Figure 6.5: $\pi_1 e_1 + \pi_2 e_2 \approx 2.78$. The (improperly applied) van Kampen expansion seriously underestimates the discrepancy between the mean and the deterministic trajectory since, as an expansion about e_1, it effectively ignores e_2, and vice versa; van Kampen's stability analysis is therefore the correct theoretical approach in this case.

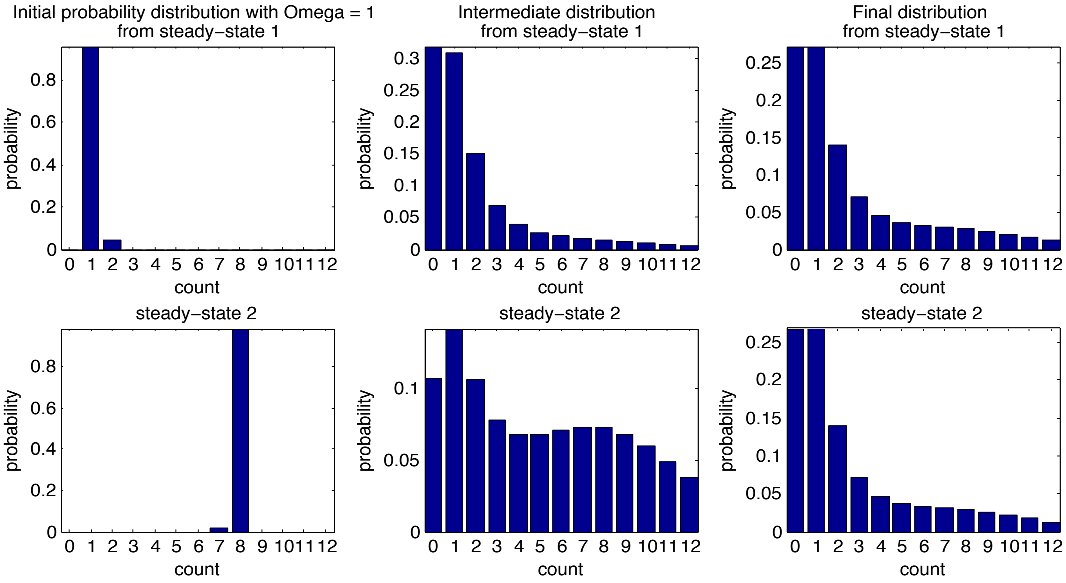


Figure 6.8: Initial (left), intermediate (center), and final (right) probability distributions of the exact Master equation for the one gene system with two stable steady-states (6.6), starting from e_1 (top) or e_2 (bottom), with $\Omega = 1$. The probability distributions start out peaked at their respective initial conditions. Over time, some of the probability begins to flow from one deterministic steady-state to the other. Regardless of the initial condition, the system eventually reaches a single bimodal stochastic steady-state (the same distribution shown in Figure 6.5), with a stronger peak at e_1 (the more stable of the two points) and a weaker peak at e_2.

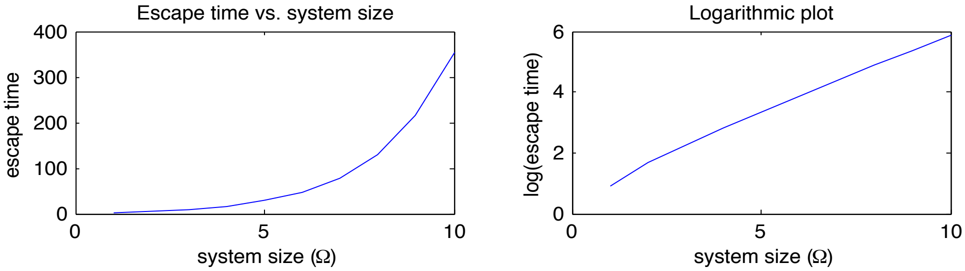


Figure 6.9: Escape time $\tau_{2,1}$ versus system size Ω for system (6.6) (left), computed as mean first-passage time via equation (5.20). The plot of $\log(\tau_{2,1})$ vs. Ω (right) is linear, confirming that the escape time grows exponentially with the system size.

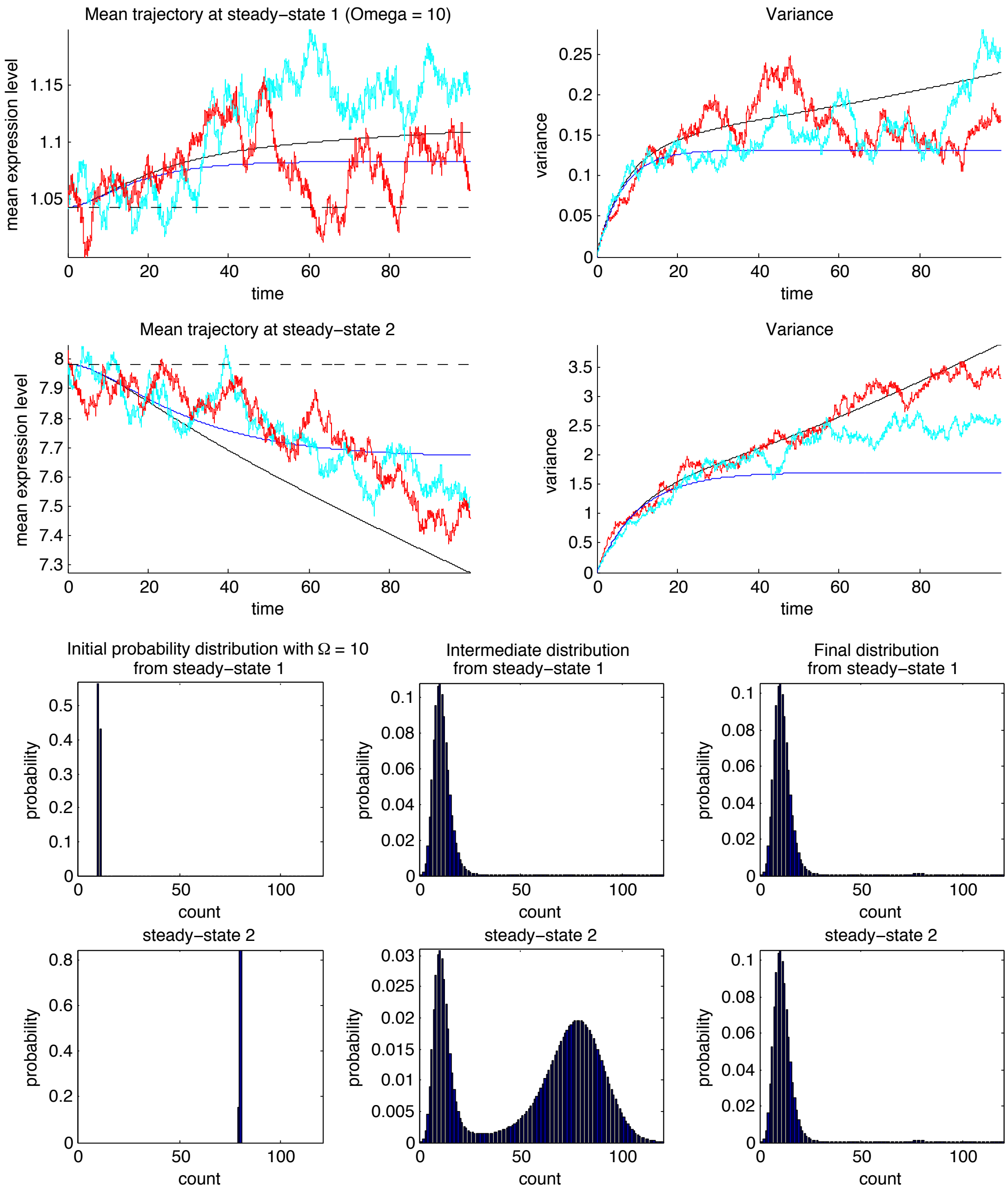


Figure 6.10: One gene system with two stable steady-states (6.6), $\Omega = 10$. Just as in Figures 6.7 and 6.8, regardless of the initial condition, the probability converges to a bimodal distribution with a strong peak at e_1 ($\pi_1 = 97\%$) and weaker peak at e_2 ($\pi_2 = 3\%$), and the mean converges to the weighted average $\pi_1 e_1 + \pi_2 e_2 \approx 1.24$ predicted in our stability analysis for $\Omega = 10$.

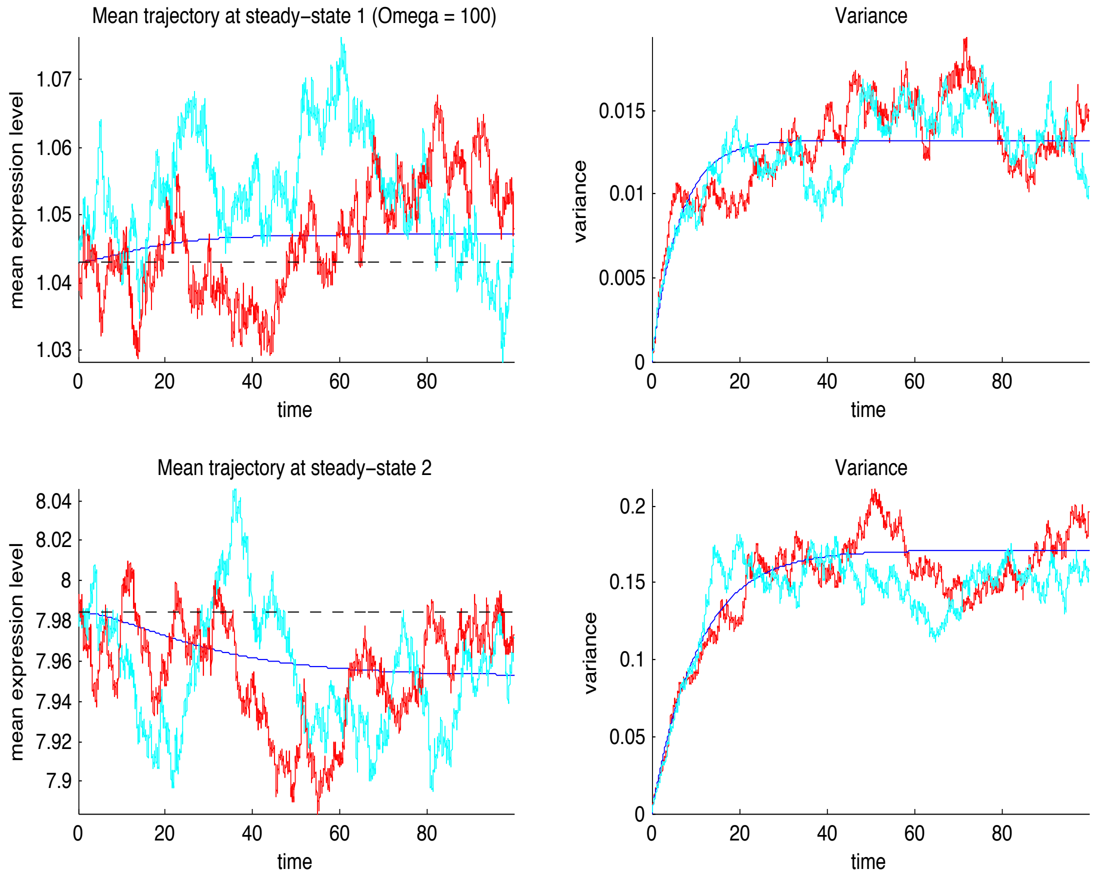
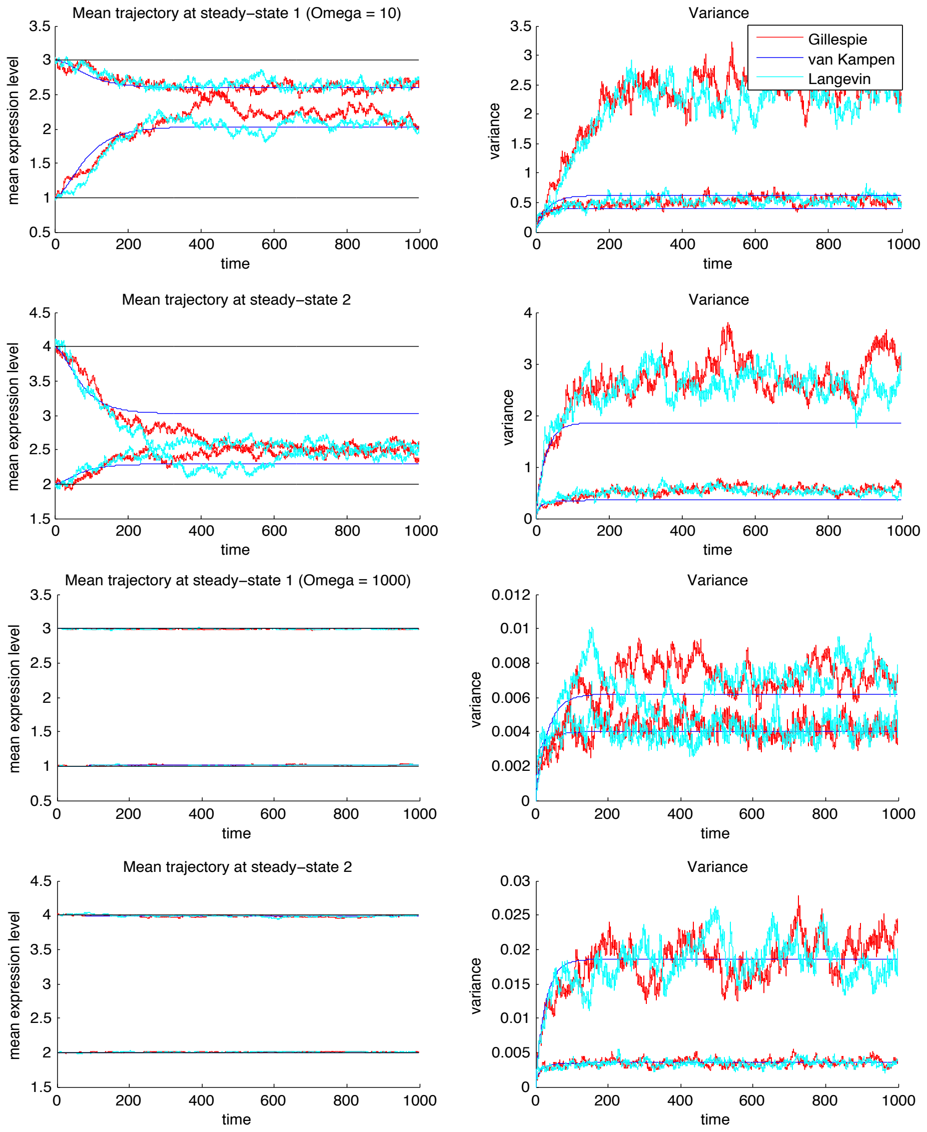


Figure 6.11: One gene system with two stable steady-states (6.6), $\Omega = 100$. Mean (left) and variance (right) trajectories via (improperly applied) van Kampen approximation (blue) and average of 100 trajectories of the Gillespie (red) and Langevin simulation (cyan). The exact Master equation calculation suffered from instability (oscillations) so the trajectory is not shown here. Since the system is so large, the probability of a jump between e_1 and e_2 is extremely small, so the escape time is longer than the length of the simulation. Therefore the stochastic trajectories remain close to the deterministic steady-state where they originated for the duration of the simulation. Since the two deterministic steady-states operate mostly independently of each other in the simulation time-frame, the van Kampen approximation agrees quite well, unlike for smaller system sizes. The variance and the difference between the mean and deterministic trajectory are both on the order of $O(\Omega^{-1})$.

and the van Kampen approximation is quite accurate within this time-frame.

6.6 Conclusions

Our simulation studies support and illustrate the theory discussed in chapter 5 by comparing the van Kampen expansion, Gillespie simulation, and Langevin simulation for systems with one or multiple stable steady-states, hence very different qualitative characteristics. For one gene systems, we can compare the performance of each approach to the exact trajectory of the Master equation. Our study of a one-gene system with one stable steady-state shows that for system-size Ω, both the variance and the difference between the stochastic mean and deterministic trajectory are $O(\Omega^{-1})$, and the van Kampen expansion, Gillespie simulation and Langevin simulation are all in excellent agreement with Master equation, (except for slight inaccuracy in the van Kampen and Langevin approximations for very small systems). Furthermore, the deterministic and stochastic trajectories are almost identical for large systems. As the system size increases, the final probability distribution of the stochastic system becomes increasingly sharply peaked at the deterministic steady-state. The two gene system with one stable steady-state confirms these observations. The bistable systems exhibit much more complex behavior. Rather than staying near the initial deterministic steady-state, the Gillespie and Langevin simulations (and exact Master equation, for the one-gene system) deviate dramatically from both the (improperly applied) van Kampen expansion and the deterministic trajectory, at least for small Ω. The explanation is that each stochastic trajectory has a reasonable probability of escaping from the domain of attraction of one stable steady-state and being attracted to another. In the long run, the system settles to a bimodal steady-state distribution, in which both stable steady-states are represented proportional to their relative stability, and the mean is the weighted average of the two stable steady-states (as predicted by the alternative van Kampen theory for multiple stable steady-states). However, for large bistable systems, the escape time can far exceed the length of the simulation, since escape time scales exponentially with system size. Therefore, the stochastic trajectories remain close to the deterministic steady-state where they originated for



Mean trajectory at steady–state 1 (Omega = 10)

Variance

Mean trajectory at steady–state 2

Variance

Mean trajectory at steady–state 1 (Omega = 1000)

Variance

Mean trajectory at steady–state 2

Variance

Figure 6.12: Two gene system with two stable steady-states (6.7), with $\Omega = 10$ (top) and $\Omega = 1000$ (bottom): Mean and variance via van Kampen approximation (blue), and average over 100 simulations of the Gillespie (red) and Langevin simulation (cyan). For small systems ($\Omega = 10$), the stochastic mean trajectory converges to a weighted average of e_1 and e_2 corresponding to a bimodal stochastic steady-state. For large systems ($\Omega = 1000$), jumps between e_1 and e_2 are very rare, so the escape time is large and the trajectories remain near their initial conditions for the duration of the simulation, hence the van Kampen approximation is quite accurate (though technically not applicable) and the variance and mean-deterministic discrepancy are both $O(\Omega^{-1})$.

the duration of the simulation.

We can draw several important conclusions from the theory of chapter 5 and the results of these studies. The first is that for large systems with a single stable steady-state, the deterministic model is sufficient for almost any practical purpose. In particular, we are entirely justified in applying the deterministic model-based inference method of chapter 4 to biological data if these conditions hold. Of course, systems with only one stable steady-state are not of much interest biologically. Fortunately, for large multistable systems, the escape time is so large that the steady-states operate independently of each other practically indefinitely (assuming the system can be initialized with sharp peaks at stable steady-states). Even for moderately-sized or randomly initialized multistable systems, the final probablity distributions are multimodal with peaks at the locations of the deterministic steady-states. Hence we can still measure the deterministic steady-state expression levels needed for our algorithm if, rather than averaging the expression levels, we instead locate the expression peaks. For the large system sizes typical in gene expression studies, the expression peaks will be extremely close to the deterministic steady-states.

It is worth specifically relating the effects of stochasticity to gene perturbation, since perturbed steady-states are central to our inference algorithm. A gene regulatory system immediately following a perturbation like gene knockdown is not in steady-state, so the expression distribution will be in flux for some period of time before reaching a final stochastic steady-state consistent with the perturbation. This steady-state is, in general, a multimodal distribution, different from the system's natural steady-state distribution due to the perturbation. The peaks of the distribution correspond to deterministic stable steady-states consistent with the fixed expression levels of the perturbed genes. If there is only one such deterministic steady-state, the final distribution will be unimodal; if there are multiple, the system will eventually explore them all. Generally, a perturbed system does not start out very close to a particular deterministic steady-state, so it has a reasonable probability of initial attraction to any possible state and the distribution quickly reaches its multimodal steady-state (on a very short time scale relative to the escape time, as discussed in section 5.3). To collect data for the inference algorithm, the experimenter should apply each perturbation, wait for the system to settle to its stochastic steady-state distribution, and measure the expression peaks, which correspond to deterministic perturbed steady-states.

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

Appendix A

A1 Thermodynamic model

In equation 2.1, the function $f_i(y)$ represents the probability that RNAP binds to the ith gene promoter. We claim that $f_i(y)$ has the form:

$$f_i(y) \equiv p^{(i)}_{\text{bound}}(y) = \frac{\sum_j e^{-\beta\Delta\epsilon^{\text{RNAP}}_{ij}} P e^{-\beta\Delta\epsilon_{ij}} \Pi_{k\in S_{ij}} y_k}{\sum_j (1 + e^{-\beta\Delta\epsilon^{\text{RNAP}}_{ij}} P) e^{-\beta\Delta\epsilon_{ij}} \Pi_{k\in S_{ij}} y_k},$$

where $\Delta\epsilon_{ij}$ is the binding energy of the jth complex to the promoter, $\Delta\epsilon^{\text{RNAP}}_{ij}$ is the binding energy of RNAP to the jth promoter-bound complex, and P, x_j are the concentrations of RNAP and gene product j [BBG$^+$05a, BBG$^+$05b].

Any type of regulator (including no regulator at all) can be represented in this framework. For no regulator, we take $S_{ij} = \emptyset$ with the convention that $\Pi_{k\in\emptyset} y_k = 1$, set $\Delta\epsilon_{ij} = 0$, and take $\Delta\epsilon^{\text{RNAP}}_{ij}$ as the base binding energy of RNAP to the promoter. For a repressor, $\Delta\epsilon_{ij} < 0$ and $\Delta\epsilon^{\text{RNAP}}_{ij} > 0$; for an activator, $\Delta\epsilon_{ij} < 0$ and $\Delta\epsilon^{\text{RNAP}}_{ij} < 0$.

Setting

$$b_{ij} = e^{-\beta\Delta\epsilon^{\text{RNAP}}_{ij}} P e^{-\beta\Delta\epsilon_{ij}}$$

$$c_{ij} = (1 + e^{-\beta\Delta\epsilon^{\text{RNAP}}_{ij}} P) e^{-\beta\Delta\epsilon_{ij}},$$

we obtain the form given in Section 1:

$$f_i(y) = \frac{b_{ij} \Pi_{k\in S_{ij}} y_k}{\sum_j c_{ij} \Pi_{k\in S_{ij}} y_k}.$$

Constant terms in the numerator and denominator correspond to the no-regulator case. Letting c_{i0} denote the constant appearing in the denominator, our convention will be to divide all of the coefficients in the numerator and denominator by c_{i0} so that the constant 1 appears in the denominator.

A1.1 Simplified derivation

The derivation we present here follows Bintu et al and Garcia et al [BBG$^+$05a, BBG$^+$05b, GKO$^+$11]. For simplicity, we will prove the following claim for the simplified case with one regulator y_1 (as well as the possibility of RNAP binding with no regulator):

$$p_{\text{bound}}^{(i)} = \frac{e^{-\beta\Delta\epsilon_{i0}^{\text{RNAP}}}p + e^{-\beta\Delta\epsilon_{i1}^{\text{RNAP}}}pe^{-\beta\Delta\epsilon_{i1}}y_1}{(1+e^{-\beta\Delta\epsilon_{i0}^{\text{RNAP}}}p) + (1+e^{-\beta\Delta\epsilon_{ij}^{\text{RNAP}}}p)e^{-\beta\Delta\epsilon_{i1}}y_1}.$$

We will use the following notation: $\epsilon_{P,i1}^{S}$ is the energy of the state in which RNAP is specifically bound to the regulator-promoter complex, $\epsilon_{P,i0}^{S}$ is the energy of the state in which RNAP is specifically bound to the promoter without the regulator, ϵ_P^{NS} is the energy when RNAP is bound to a nonspecific binding site, ϵ_{i1}^{S} is the energy when y_1 is specifically bound to the promoter, and ϵ_{i1}^{NS} is energy when y_1 is bound to a nonspecific binding site. Then

$$\Delta\epsilon_{i0}^{\text{RNAP}} = \Delta\epsilon_{P,i0} \equiv \epsilon_{P,i0}^{S} - \epsilon_P^{NS}, \quad \Delta\epsilon_{i1}^{\text{RNAP}} = \Delta\epsilon_{P,i1} \equiv \epsilon_{P,i1}^{S} - \epsilon_P^{NS}, \qquad \Delta\epsilon_{i1} \equiv \epsilon_{y_1}^{S} - \epsilon_{y_1}^{NS}.$$

Suppose that we have j RNA polymerase molecules and k molecules of gene product 1 (the regulator). We model the genome as a "reservoir" with n nonspecific binding sites (to which either RNAP or regulator can bind). One of these sites is the promoter of gene i. Three different classes of configurations interest us:

1. empty promoter
2. regulator bound to promoter
3. regulator and RNAP bound to promoter

4. RNAP only bound to promoter

These correspond to the following partial partition functions, which represent the "unnormalized probabilities" of each configuration.

1. $Z(j, k)$
2. $Z(j, k-1)e^{-\beta\epsilon_{i1}^S}$
3. $Z(j-1, k-1)e^{-\beta\epsilon_{i1}^S}e^{-\beta\epsilon_{P,i1}^S}$
4. $Z(j-1, k)e^{-\beta\epsilon_{P,i0}^S}$

where $Z(j,k) = \frac{n!}{j!k!(n-j-k)!}e^{-\beta r\epsilon_{i1}^{NS}}e^{-\beta\epsilon_P^{NS}}$.

$Z(j, k)$ is equal to the total number of arragements of RNAP and regulator on the nonspecific binding sites times the Boltzmann factor, which gives the relative probability $e^{-\beta\epsilon}$ of a particular state in terms of its energy ϵ.

Since RNAP binds the promoter only in the third and fourth classes of configurations, the probability that RNAP binds the promoter is equal to the unnormalized probability of the third and fourth configurations divided by the "total probability" (the sum of the unnormalized probabilities of all classes of configurations). Hence

$$\begin{aligned}
p_{\text{bound}} &= \frac{Z(j-1,k)e^{-\beta\epsilon_{P,i0}^S} + Z(j-1,k-1)e^{-\beta\epsilon_{i1}^S}e^{-\beta\epsilon_{P,i1}^S}}{Z(j,k) + Z(j-1,k)e^{-\beta\epsilon_{P,i0}^S} + Z(j,k-1)e^{-\beta\epsilon_{i1}^S} + Z(j-1,k-1)e^{-\beta\epsilon_{i1}^S}e^{-\beta\epsilon_{P,i1}^S}} \\
&\approx \frac{\frac{n^{j-1}n^k}{(j-1)!k!}e^{-\beta k\epsilon_{i1}^{NS}}e^{-\beta(j-1)\epsilon_P^{NS}}e^{-\beta\epsilon_{P,i0}^S} + \frac{n^{j-1}n^{k-1}}{(j-1)!(k-1)!}e^{-\beta(k-1)\epsilon_{i1}^{NS}}e^{-\beta(j-1)\epsilon_P^{NS}}e^{-\beta\epsilon_{i1}^S}e^{-\beta\epsilon_{P,i1}^S}}{\frac{n^j n^k}{j!k!}e^{-\beta k\epsilon_{i1}^{NS}}e^{-\beta j\epsilon_P^{NS}} + \frac{n^{j-1}n^k}{(j-1)!k!}e^{-\beta k\epsilon_{i1}^{NS}}e^{-\beta(j-1)\epsilon_P^{NS}}e^{-\beta\epsilon_{P,i0}^S} + \dots} \\
&= \frac{\frac{j}{n}e^{\beta\epsilon_P^{NS}}e^{-\beta\epsilon_{P,i0}^S} + \frac{j}{n}\frac{k}{n}e^{\beta\epsilon_{i1}^{NS}}e^{\beta\epsilon_P^{NS}}e^{-\beta\epsilon_{i1}^S}e^{-\beta\epsilon_{P,i1}^S}}{1 + \frac{j}{n}e^{\beta\epsilon_P^{NS}}e^{-\beta\epsilon_{P,i0}^S} + \frac{k}{n}e^{\beta\epsilon_{i1}^{NS}}e^{-\beta\epsilon_{i1}^S} + \frac{j}{n}\frac{k}{n}e^{\beta\epsilon_{i1}^{NS}}e^{\beta\epsilon_P^{NS}}e^{-\beta\epsilon_{i1}^S}e^{-\beta\epsilon_{P,i1}^S}} \\
&= \frac{\frac{j}{n}e^{-\beta\Delta\epsilon_{P,i0}} + \frac{j}{n}\frac{k}{n}e^{-\beta\Delta\epsilon_{i1}}e^{-\beta\Delta\epsilon_{P,i1}}}{1 + \frac{j}{n}e^{-\beta\Delta\epsilon_{P,i0}} + \frac{k}{n}e^{-\beta\Delta\epsilon_{i1}} + \frac{j}{n}\frac{k}{n}e^{-\beta\Delta\epsilon_{i1}}e^{-\beta\Delta\epsilon_{P,i1}}} \\
&= \frac{\frac{j}{n}e^{-\beta\Delta\epsilon_{P,i0}} + \frac{j}{n}\frac{k}{n}e^{-\beta\Delta\epsilon_{i1}}e^{-\beta\Delta\epsilon_{P,i1}}}{1 + \frac{j}{n}e^{-\beta\Delta\epsilon_{P,i0}} + \frac{k}{n}e^{-\beta\Delta\epsilon_{i1}}(1 + \frac{j}{n}e^{-\beta\Delta\epsilon_{P,i1}})} \\
&= \frac{pe^{-\beta\Delta\epsilon_{i0}^{\text{RNAP}}} + py_1e^{-\beta\Delta\epsilon_{i1}}e^{-\beta\Delta\epsilon_{i1}^{\text{RNAP}}}}{1 + pe^{-\beta\Delta\epsilon_{i0}^{\text{RNAP}}} + y_1e^{-\beta\Delta\epsilon_{i1}}(1 + pe^{-\beta\Delta\epsilon_{i1}^{\text{RNAP}}})},
\end{aligned}$$

where in the second line we used the approximation $\frac{n!}{j!k!(n-j-k)!} \approx \frac{n^j n^k}{j!k!}$ whichs hold for $j, k << n$, in the third we divided by $\frac{n^j n^k}{j!k!} e^{-\beta k \epsilon_{i1}^{NS}} e^{-\beta \epsilon_P^{NS}}$, in the fourth we used the identities $\Delta\epsilon_{P,i0} = \epsilon_{P,i0}^S - \epsilon_P^{NS}$, $\Delta\epsilon_{P,i1} = \epsilon_{P,i1}^S - \epsilon_P^{NS}$, $\Delta\epsilon_{i1} \equiv \epsilon_{i1}^S - \epsilon_{i1}^{NS}$, and in the last we substituted in the definitions $\frac{j}{n} = p$, $\frac{k}{n} = y_1$, $\Delta\epsilon_{i0}^{\text{RNAP}} = \Delta\epsilon_{P,i0}$, $\Delta\epsilon_{i1}^{\text{RNAP}} = \Delta\epsilon_{P,i1}$.

A2 Nonidentifiability

To see that the equations

$$\frac{dx_i}{dt} = \frac{b_{i0} + \sum_{j=1}^N b_{ij} \Pi_{k\in S_{ij}} x_k}{1 + \sum_{j=1}^N c_{ij} \Pi_{k\in S_{ij}} x_k} - \gamma_i x_i, \quad b_{i0} < 1.$$

and

$$\frac{dx_i}{dt} = \frac{(wb_{i0} + \gamma_i)x_i + \sum_{j=1}^N wb_{ij} \Pi_{k\in S_{ij}} x_i x_k}{1 + wx_i + \sum_{j=1}^N wc_{ij} \Pi_{k\in S_{ij}} x_i x_k} - \gamma_i x_i, \quad w \geq \frac{\gamma}{1 - b_{i0}}$$

(equations 4.2, 4.3 in the main text) reduce to the same equation at any steady-state where $x_i \neq 0$, choose any $0 \leq b_{i0} < 1$, $0 \leq b_{ij} \leq c_{ij}$, $1 < j \leq N$, $w \geq \frac{\gamma}{1-b_{i0}}$, and calculate:

$$\begin{aligned}
& 0 = \frac{(wb_{i0} + \gamma_i)x_i + \sum_{j=1}^N wb_{ij} \Pi_{k\in S_{ij}} x_i x_k}{1 + wx_i + \sum_{j=1}^N wc_{ij} \Pi_{k\in S_{ij}} x_i x_k} - \gamma_i x_i \\
\iff\ & 0 = (wb_{i0} + \gamma_i)x_i + \sum_{j=1}^N wb_{ij} \Pi_{k\in S_{ij}} x_i x_k - \gamma_i x_i (1 + wx_i + \sum_{j=1}^N wc_{ij} \Pi_{k\in S_{ij}} x_i x_k) \\
& \quad = wx_i (b_{i0} + \sum_{j=1}^N b_{ij} \Pi_{k\in S_{ij}} x_k - \gamma_i x_i (1 + \sum_{j=1}^N c_{ij} \Pi_{k\in S_{ij}} x_k)) \\
\iff\ & 0 = b_{i0} + \sum_{j=1}^N b_{ij} \Pi_{k\in S_{ij}} x_k - \gamma_i x_i (1 + \sum_{j=1}^N c_{ij} \Pi_{k\in S_{ij}} x_k) \quad \text{(provided } x_i \neq 0) \\
\iff\ & 0 = \frac{b_{i0} + \sum_{j=1}^N b_{ij} \Pi_{k\in S_{ij}} x_k}{1 + \sum_{j=1}^N c_{ij} \Pi_{k\in S_{ij}} x_k} - \gamma_i x_i
\end{aligned}$$

A3 Tie-breaking

Assuming that we allow only first and second-order terms, we can determine whether a given equation is ambiguous as follows. If it includes no self-regulation at all, it is of the simple form, and has a class of alternatives of the higher-order form parametrized by $w \geq \frac{\gamma}{1-b_{i0}}$. On the other hand, if it includes self-regulation in every term except for the constant 1 in the denominator, and the coefficient of x_i in the denominator is greater than γ_i, then it is of the higher-order form and has an alternative of the simple form.

Practically, the simplest way to make this decision is solve the convex optimization problem (4.1) twice: once normally, and once without allowing self-regulation (that is, adding the additional constraints that $b_{ij} = 0$ whenever x_i is in the jth complex). We can compare the forms of the recovered equations as well as the quality of the fit (i.e. the unregularized objective). If the equation is ambiguous, the restriction on self-regulation will have little effect on the quality of fit, since it will simply cause the algorithm to choose the simple alternative. Comparing the recovered equations, we will also notice that they are either the same, or have the relationship given by equations 4.2 and 4.3. On the other hand, if the equation is unambiguous, the first recovered equation will not have the form of either equation 4.2 or equation 4.3, and the quality of the fit will be significantly worse when self-regulation is restricted. This test may not always be conclusive (for example, this occurs for the Nanog and Gata6 equations in the noisy simulation), but if we are unsure we can always apply the derivative tie-breaker described below to both versions of the ambiguous equation as well the equation recovered normally, and select the form that makes the best prediction.

In order to choose between the possible forms of an ambiguous equation (and possibly find w), we can measure the derivative $\frac{dx_i}{dt}$ experimentally and check whether it agrees with the value predicted by the simple form of the equation. Specifically, we can choose and perform a perturbation that is likely to have a major impact on the system, measure the concentrations shortly afterwards, and approximate the derivative by:

$$\frac{dx_i}{dt}(t_0) \approx \frac{x_i(t_1) - x_i(t_0)}{t_1 - t_0}.$$

(This type of experiment is not easy to carry out on a large scale, so we must choose which derivatives to measure with care, and do so only when necessary.) Next we predict the derivative following the perturbation using the simple equation. For example, if we knock out term $\Pi_{k \in S_{iJ}} x_k$ starting from steady-state μ, the simple form predicts:

$$\frac{dx_i}{dt} = \frac{b_{i0} + \sum_{j \neq J} b_{ij} \Pi_{k \in S_{ij}} \mu_k}{1 + \sum_{j \neq J} c_{ij} \Pi_{k \in S_{ij}} \mu_k} - \gamma_i \mu_i.$$

If the measured and predicted derivatives agree, we know that the simple form is correct. Otherwise, we conclude that the true equation has the higher-order form, and estimate w as follows:

$$w = \frac{-\frac{dx_i}{dt}}{(\frac{dx_i}{dt} + \gamma_i x_i)(x_i + \sum_{j=1}^{N} c_{ij} \Pi_{k \in S_{ij}} x_i x_k) - b_{i0} x_i - \sum_{j=1}^{N} b_{ij} \Pi_{k \in S_{ij}} x_i x_k}. \tag{A1}$$

From a practical perspective, the measured derivative will not agree exactly with the predicted derivative, so if we are unsure whether we have a match, we can solve for w and determine whether it is possible ($w \geq \frac{\gamma}{1-b_{i0}}$) and reasonable. Furthermore, if we are unsure whether or not the equation is ambiguous, we can also predict the derivative with the equation recovered without restrictions, and compare this prediction with those of the two alternative equations.

A4 Simulated six-gene subnetwork in mouse ESC

We test our method on a synthetic network governed by the system of ODEs (4.4). The $\frac{d[C]}{dt}$, $\frac{d[Gc]}{dt}$, and $\frac{d[G]}{dt}$ equations are ambiguous with the alternative forms given in (4.5) (provided we ignore the very small constant term in the $\frac{d[C]}{dt}$ equation and $[G]$ term in the $\frac{d[G]}{dt}$) . The $\frac{d[C]}{dt}$ equation has the higher-order form and an alternative simple form, while the $\frac{d[Gc]}{dt}$, and $\frac{d[G]}{dt}$ equations have the simple form and alternative higher-order forms. (Although $[S]$ appears in every term of the $\frac{d[S]}{dt}$ equation, it is not ambiguous since $wb_{i0} + \gamma_i = 0$, which is impossible since $w > 0, b_{i0} \geq 0$.) We apply our method twice, once allowing self-regulation and again disallowing it. Then we compare the two recovered forms of each equation and their fit to the data to determine whether nonidentifiability exists in each case. If so, we break the tie by examining derivatives. We do this for both noiseless and noisy data.

Without noise, we use a total of 52 measurements: the expression levels at each of the system steady-states (ESC, DSC, Endo and Trophect) and expression levels at the steady states reached after overexpressing each gene at twice its steady-state level, and knocking it down to one-fifth of its steady-state level, starting from each basic steady-state. We use cross validation (CV) to select the sparsity paramter λ for each equation, with and without self-regulation (Figure A1). We use CVX software to solve the convex optimization problem [GB11, GB08]. When we solve without restricting self-regulation (using the sparsity parameters chosen by CV), we recover the following equations (with coefficients thresholded at 0.1% of the largest recovered coefficient, except Oct4, which is thresholded at 0.01%):

$$\frac{d[O]}{dt} = \frac{[A] + 0.001 + (0.005[O][S] + 0.025[O][S][N])}{1 + [A] + (0.001[O] + 0.005[O][S] + 0.025[O])[S][N] + 10[O][C] + 10[Gc]} - 0.1[O]$$

$$\frac{d[S]}{dt} = \frac{0.001 + 0.005[O][S] + 0.025[O][S][N]}{1 + 0.005[O][S] + 0.025[O][S][N]} - 0.1[S]$$

$$\frac{d[N]}{dt} = \frac{0.1[O][S] + 0.1[O][S][N]}{1 + 0.1[O][S] + 0.1[O][S][N] + 10[O][G]} - 0.1[N]$$

$$\frac{d[C]}{dt} = \frac{0.95}{1 + 2.5[O]} - 0.1[C]$$

$$\frac{d[Gc]}{dt} = \frac{0.1[Gc] + 0.01[C][Gc] + 0.01[G][Gc]}{1 + 0.1[Gc] + 0.01[C][Gc] + 0.01[G][Gc]} - 0.1[Gc]$$

$$\frac{d[G]}{dt} = \frac{0.1 + 0.95[O]}{1 + 0.95[O] + 14.25[N] + 0.04[S][N] + 0.08[N][C] + 0.02[N][Gc] + 0.08[N][G]} - 0.1[G]$$

Table A1: Quality of fit (unregularized objective value) for noiseless data

Equation	unrestricted solution	no self-regulation
Oct4	1.249×10^{-5}	5.7674
Sox2	1.1226×10^{-9}	0.4269
Nanog	6.7426×10^{-8}	0.7664
Cdx2	3.5627×10^{-7}	9.7365×10^{-7}
Gcnf	1.2954×10^{-7}	2.130×10^{-7}
Gata6	7.7505×10^{-7}	7.9072×10^{-5}

When we solve the same problem, disallowing self-regulation (again using the appropriate CV sparsity parameters), we recover the following equations.

$$\begin{aligned}
\frac{d[O]}{dt} &= \frac{[A]+0.14[C]+0.57[G]+0.12[S][G]+0.04[N][C]+0.20[N][Gc]}{1+[A]+20[C]+9.6[Gc]+3.3[G]+0.33[S][C]+0.04[N][C]+0.20[N][Gc]} \\
&\quad - 0.1[O] \\
\frac{d[S]}{dt} &= \frac{0.18[O][N]}{1+0.18[O][N]} - 0.1[S] \\
\frac{d[N]}{dt} &= \frac{0.01+0.12[O][S]}{1+0.12[O][S]} - 0.1[N] \\
\frac{d[C]}{dt} &= \frac{0.95}{1+2.5[O]} - 0.1[C] \\
\frac{d[Gc]}{dt} &= \frac{0.001+0.1[C]+0.1[G]}{1+0.1[C]+0.1[G]} - 0.1[Gc] \\
\frac{d[G]}{dt} &= \frac{0.1+[O]}{1+[O]+0.03[N][Gc]+15[N]} - 0.1[G]
\end{aligned}$$

We measure the quality of the fit by the unregularized objective value $\|G_i b_i - \gamma D_i G_i c_i\|$ from equation 4.1 for each recovered equation. These values are given in Table A1. We can tell that the first three equations are unambiguous, while the last three have alternative forms, since the quality of fit is roughly the same for the last three equations whether or not we restrict self-regulation, while for the first three, the fit is much worse when self-regulation is prohibited. Therefore, we can use the recovered forms of the first three equations, but we need to break ties between alternative forms of the last three equations.

It's easiest to start with the simple forms of the last three equations, and determine the corresponding higher-order forms. To break the tie, we look at derivatives following Oct4, Cdx2, and Nanog knockouts, respectively, since these regulators are important in each of the three ambiguous equations (Figure A2). After each knockout we estimate the derivative with a finite difference, compare it to the derivative prediction made by the simple form of the equation, and accept this form if the match is good. Otherwise we compute w using equation A1, and (provided it is reasonable), accept the higher-order form with this choice of w.

In this case, we measure $\frac{d[C]}{dt} \approx 0.6004825$ immediately after Oct4 knockout from ESC steady-state using a finite difference. The simple form of the equation yields $\frac{d[C]}{dt} \approx 0.78$ immediately following the knockout, which is a poor match. Therefore we select the higher-order form and use the measured derivative to compute $w = 2$, which yields $\frac{d[C]}{dt} = 0.60$. For the $\frac{d[Gc]}{dt}$ equation we measure $\frac{d[Gc]}{dt} \approx -0.13$ immediately after Cdx2 knockout from SC, and the simple form is a good match at -0.13 (computing w for the higher-order form using the derivative yields an unreasonably large $w \approx 86$; using the minimum value $w = 0.1$ in the higher-order form yields $\frac{d[Gc]}{dt} = -0.017$). Similarly, for the $\frac{d[G]}{dt}$ equation we measure $\frac{d[Gc]}{dt} \approx 0.69$ immediately after Nanog knockout from SC, and the simple form is a good match at 0.69. In the end, we select the equations given in (4.6).

Next we test the algorithm using noisy data by adding zero-mean Gaussian noise to each measurement, with standard deviation 1% of the measurement magnitude. We use the 4 basic steady-states, the steady states reached after knockdown and overexpression of each individual gene from each basic steady state, and those reached after knocking one gene up and one gene down from each pair of genes, starting from ESC and DSC. Again, we use cross validation to select the sparsity parameters (Figure A3).

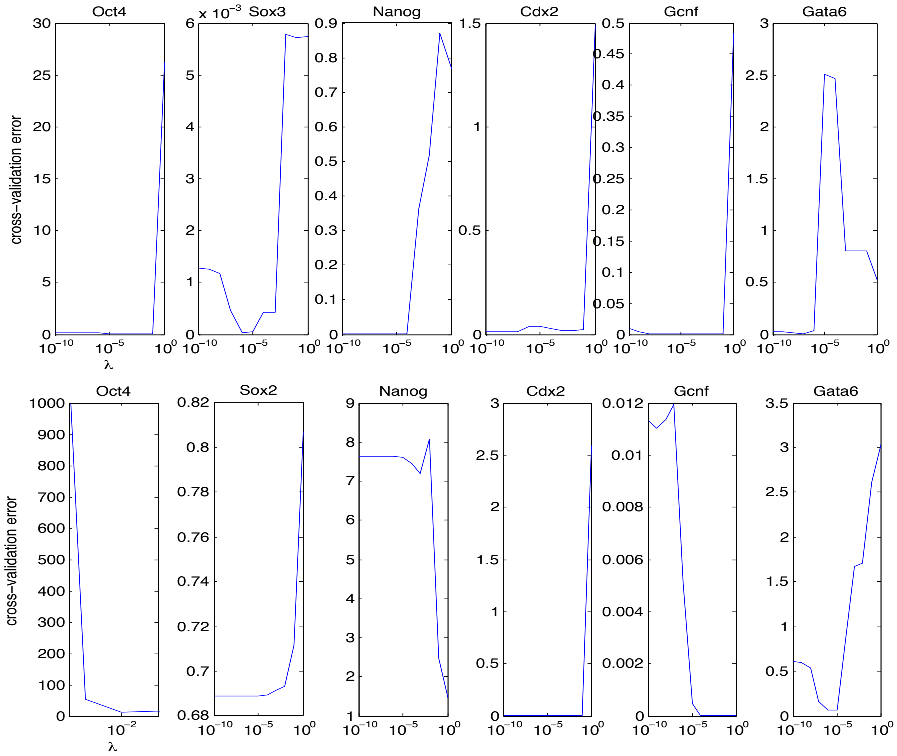


Figure A1: Cross validation (8-fold) on noiseless data. We estimate the test error for each gene equation and various choices of sparsity parameter by randomly dividing the 52 observations into 8 folds (groups), then leaving out each fold out in turn, training the model on the remaining 7 folds, testing on the omitted fold, and finally averaging the 8 resulting test errors. After repeating this process for each gene equation and several choices of sparsity parameter, we select the sparsity parameters corresponding to the lowest error for each equation. (a) Unrestricted case: we selected sparsity parameters $[10^{-1}, 10^{-5}, 10^{-6}, 10^{-4}, 10^{-2}, 10^{-6}]$. (b) No self-regulation: we selected $[10^{-2}, 10^{-5}, 10^{-1}, 10^{-1}, 10^{-4}, 10^{-5}]$.

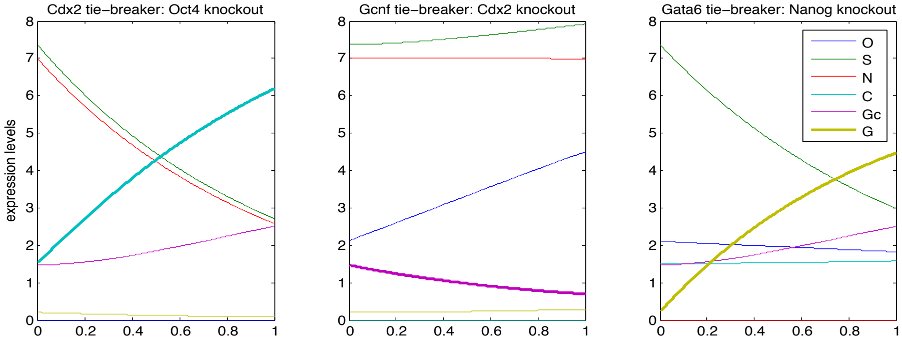


Figure A2: Derivative measurements to break ties between alternative forms. (left) Trajectory of Cdx2 following Oct4 knockout from ESC: $\frac{d[C]}{dt}(t_0) \approx 0.6004825$. (center) Gcnf trajectory following Cdx2 knockout from ESC: $\frac{d[Gc]}{dt}(t_0) \approx -0.1261507$. (right) Gata6 trajectory following Nanog knockout from ESC: $\frac{d[G]}{dt}(t_0) \approx 0.6908821$.

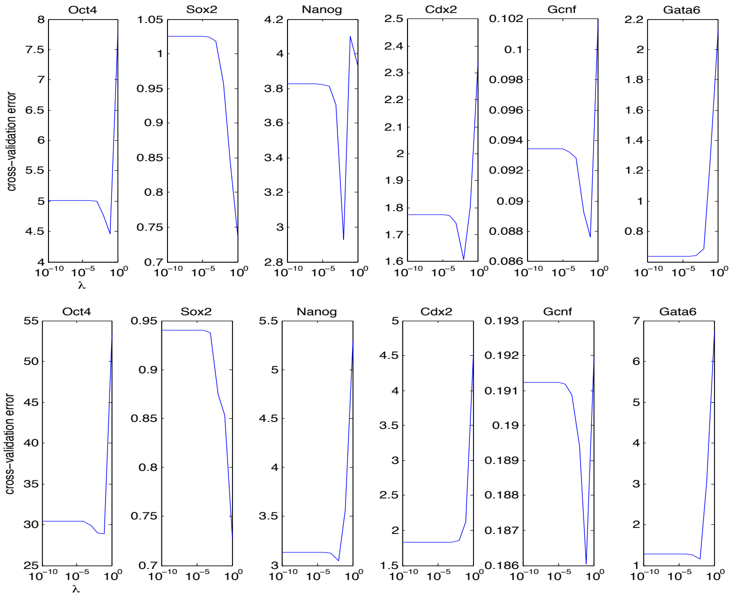


Figure A3: Cross validation (8-fold on 108 observations, with the approach described in Figure A1) on noisy data (1% Gaussian noise): (left) unrestricted: we selected sparsity parameters $[0.1, 1, 0.01, 0.01, 0.1, 0.00001]$ (right) No self-regulation: we selected $[0.1, 1, 0.01, 0.001, 0.1, 0.01]$.

When we solve the problem with noisy data (with no restriction on self-regulation) and threshold at a level of 1% of the largest recovered coefficient, we recover:

$$
\begin{aligned}
\frac{d[O]}{dt} &= \frac{[A]}{1+[A]+9.9[Gc]+9.9[O][C]} - 0.1[O] \\
\frac{d[S]}{dt} &= \frac{0.001[O][S]+0.0005[S][N]+0.025[O][S][N]}{1+0.001[O][S]+0.0005[S][N]+0.025[O][S][N]} - 0.1[S] \\
\frac{d[N]}{dt} &= \frac{0.09[O][S][N]}{1+0.1[G][Gc]+0.09[O][S][N]+9.1[O][G]} - 0.1[N] \\
\frac{d[C]}{dt} &= \frac{0.94}{1+2.4[O]} - 0.1[C] \\
\frac{d[Gc]}{dt} &= \frac{0.1[Gc]+0.01[C][Gc]+0.01[G][Gc]}{1+0.1[Gc]+0.01[C][Gc]+0.01[G][Gc]} - 0.1[Gc] \\
\frac{d[G]}{dt} &= \frac{0.1[G]+0.1[O][G]}{1+0.2[N]+0.1[G]+0.05[O][N]+0.1[O][G]+0.05[S][N]+1.4[N][G]} \\
&\quad - 0.1[G]
\end{aligned}
$$

When we solve without allowing self-regulation, we recover:

$$
\begin{aligned}
\frac{d[O]}{dt} &= \frac{[A]+0.3[G]}{1+[A]+15.4[C]+9.7[Gc]+3.1[G]+0.5[S][C]+0.6[N][C]} - 0.1[O] \\
\frac{d[S]}{dt} &= \frac{0.2[O][N]}{1+0.2[O][N]} - 0.1[S] \\
\frac{d[N]}{dt} &= \frac{0.03+0.17[S]+0.03[S][C]}{1+0.17[S]+0.03[S][C]} - 0.1[N] \\
\frac{d[C]}{dt} &= \frac{0.95}{1+2.5[O]} - 0.1[C] \\
\frac{d[Gc]}{dt} &= \frac{0.1[C]+0.1[G]}{1+0.1[C]+0.1[G]} - 0.1[Gc] \\
\frac{d[G]}{dt} &= \frac{0.1+0.9[O]}{1+0.9[O]+14.2[N]} - 0.1[G]
\end{aligned}
$$

Table A2 shows that for noisy data, the quality of fit does not indicate as clearly which equations are ambiguous. For the Oct4 and Sox2 equations, the quality of fit drops dramatically when we restrict self-regulation, while it changes very little for Cdx2, Gcnf and Gata6, revealing that the Oct4 and Sox2 equations are correct, while Cdx2,

Table A2: Quality of fit (unregularized objective value) for noisy data

Equation	unrestricted	no self-reg
Oct4	1.5170	8.1241
Sox2	0.2800	1.006
Nanog	0.6877	1.9347
Cdx2	0.5634	0.8208
Gcnf	0.0278	0.0599
Gata6	0.1278	0.2224

Gcnf and Gata6 have a simple and a higher-order form. We break the tie between the two forms of the last three equations using derivatives as before. The Nanog equation isstill unclear, so we analyze derivatives to decide between the two alternative forms *and* the solution of the unrestricted optimization problem. First we observe that the higher-order version of the ambiguous form is illegal as it contains third-order terms, so we only need to choose between the unrestricted equation and the simple equation recovered without self-regulation. We simulate the trajectory after Gata6 knockout from ESC and compare the derivative ($\frac{d[N]}{dt} = 0.023$) to the predictions of the first equation ($\frac{d[N]}{dt} = 0.044$) and the simple version ($\frac{d[N]}{dt} = -0.070$), concluding that the first equation is correct. Finally, we obtain the equations given in equation 4.7.

A5 Derivation of the Master Equation

The following derivation, simplified and adapted from Chapters IV and X of van Kampen's *Stochastic Processes in Physics and Chemistry* [VK07], is provided here for the reader's convenience.

A5.1 The Chapman-Kolmogorov equation

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

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

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

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

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

$$\begin{aligned} P(y_1, t_1; y_3, t_3) &= P(y_1, t_1) \int P(y_2, t_2|y_1, t_1)P(y_3, t_3|y_2, t_2)dy_2 \\ \implies P(y_3, t_3|y_1, t_1) &= \int P(y_2, t_2|y_1, t_1)P(y_3, t_3|y_2, t_2)dy_2. \end{aligned} \tag{A2}$$

A5.2 The Master equation

The Master equation is an equivalent form of the Chapman-Kolmogorov equation for Markov processes, but it is more convenient and easier to relate to physical concepts. In order to derive it, we first assume for convenience that the process is time-homogeneous, so we can write the transition probabilities as T_τ, i.e.

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

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

$$T_{\tau'}(y_2|y_1) = (1 - a_0\tau')\delta_{y_2,y_1} + \tau' W(y_2|y_1) + o(\tau'), \tag{A3}$$

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

$$a_0(y_1) = \int W(y_2|y_1)dy_2.$$

Inserting (A3) in place of T'_τ in the Chapman-Kolmogorov equation (A2) yields:

$$\begin{aligned}
T_{\tau+\tau'}(y_3|y_1) &= \int T_{\tau'}(y_3|y_2)T_\tau(y_2|y_1)dy_2 \\
&= \int((1-a_0(y_2)\tau')\delta_{y_3,y_2}+\tau'W(y_3|y_2))T_\tau(y_2|y_1)dy_2 \\
&= (1-a_0(y_3)\tau')T_\tau(y_3|y_1)+\tau'\int W(y_3|y_2)T_\tau(y_2|y_1)dy_2 \\
\frac{\partial}{\partial\tau}T_\tau(y_3|y_1) &= \lim_{\tau'\to 0}\frac{T_{\tau+\tau'}(y_3|y_1)-T_\tau(y_3|y_1)}{\tau'} \\
&= -a_0(y_3)T_\tau(y_3|y_1)+\int W(y_3|y_2)T_\tau(y_2|y_1)dy_2 \\
&= \int\{W(y_3|y_2)T_\tau(y_2|y_1)-W(y_2|y_3)T_\tau(y_3|y_1)\}dy_2
\end{aligned}$$

We can rewrite this equation as (5.1) from the main text as follows:

$$\begin{aligned}
P(y_3,\tau) &= \int T_\tau(y_3|y_1)P(y_1,0)dy_1 \quad \text{as } \tau\to 0 \\
\implies \quad \frac{\partial P(y_3,\tau)}{\partial\tau} &= \int\frac{\partial}{\partial\tau}T_\tau(y_3|y_1)P(y_1,0)dy_1 \\
&= \int\int P(y_1,0)\{W(y_3|y_2)T_\tau(y_2|y_1)-W(y_2|y_3)T_\tau(y_3|y_1)\}dy_1dy_2 \\
&= \int\{W(y_3|y_2)P(y_2,\tau)-W(y_2|y_3)P(y_3|\tau)\}dy_2
\end{aligned}$$

Or, changing the names of the variables:

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

A6 Eliminating intermediate species with the QSSA

The following calculation, adapted from Rao and Arkin [RA03] to fit the needs of our application, shows how to use the quasi-steady-state-assumption (QSSA) to eliminate an intermediate species from the multivariate Master equation.

Consider a chemical reaction with n species, m different reactions with propensities

$a_k(x)$, and stoichiometries v_k, $1 \leq k \leq m$. Let $x \equiv (y, z)$, where y is a primary and z is an intermediate species. Assume that the following QSSA holds:

$$\frac{dP(z|y;t)}{dt} \approx 0.$$

Then we can simplify a Master equation in $x = (y, z)$ to an equation in y, as follows:

$$\begin{aligned}
\frac{dP(x;t)}{dt} &= \sum_{k=0}^{m} [a_k(x - v_k)P(x - v_k; t) - a_k(x)P(x;t)] \\
\frac{dP(y,z;t)}{dt} &= \sum_{k=0}^{m} [a_k(y - v_k^y, z - v_k^z)P(y - v_k^y, z - v_k^z; t) - a_k(y,z)P(y,z;t)] \\
P(y,z;t) &= P(z|y;t)P(y;t) \\
\frac{dP(y,z;t)}{dt} &= P(y;t)\frac{dP(z|y;t)}{dt} + P(z|y;t)\frac{dP(y;t)}{dt} \approx P(z|y;t)\frac{dP(y;t)}{dt} \qquad \text{(QSSA)} \\
P(z|y;t)\frac{dP(y;t)}{dt} &= \sum_{k=0}^{m} [a_k(y - v_k^y, z - v_k^z)P(z - v_k^z|y - v_k^y)P(y - v_k^y) - a_k(y,z)P(z|y)P(y)] \\
\frac{dP(y;t)}{dt} &= \sum_{z} P(z|y;t)\frac{dP(y;t)}{dt} \\
\Longrightarrow \quad \frac{dP(y;t)}{dt} &= \sum_{k=0}^{m} [b_k(y - v_k^y)P(y - v_k^y; t) - b_k(y)P(y;t)], \\
\text{where } b_k(y) &= \sum_{z} a_k(y,z)P(z|y).
\end{aligned}$$

A7 van Kampen's Master equation expansion

This calculation is adapted from van Kampen, chapter X ([VK07]); it is simplified from the original by assuming a birth-and-death process, and provided here for the reader's convenience.

The Master equation for a birth-and-death process given by:

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

Assume that the transition probabilties have the special form:

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

and define

$$\alpha_\nu(x) = \sum_r r^\nu \Phi_0(x;r) = \Phi_0^+(x) + (-1)^\nu \Phi_0^-(x).$$

For birth-and-death processes, we have

$$W_\Omega(X{+}1|X) = \Omega\Phi_0(\frac{X}{\Omega};+1) \equiv \Omega\Phi_0^+(\frac{X}{\Omega}), \quad W_\Omega(X{-}1|X) = \Omega\Phi_0(\frac{X}{\Omega};-1) \equiv \Omega\Phi_0^-(\frac{X}{\Omega})$$

$$\alpha_1(\phi) = \Phi_0^+(\phi(t)) - \Phi_0^-\phi(t), \quad \alpha_2(\phi) = \Phi_0^+(\phi(t)) + \Phi_0^-(\phi(t)).$$

Hence the Master equation becomes

$$\begin{aligned}\frac{\partial P(X,t)}{\partial t} &= \Omega\{\Phi_0^+(\frac{X-1}{\Omega})P(X-1,t) + \Phi_0^-(\frac{X+1}{\Omega})P(X+1,t) \\ &\quad - (\Phi_0^+(\frac{X}{\Omega}) + \Phi_0^-(\frac{X}{\Omega}))P(X,t)\}. \end{aligned} \tag{A4}$$

As discussed in the main text, we make the Ansatz:

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

and define Π by:

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

The partial derivatives are Π are given by:

$$\begin{aligned}\frac{\partial^\nu \Pi}{\partial \xi^\nu} &= \Omega^{\frac{1}{2}} \frac{\partial^\nu P}{\partial X^\nu} \\ \frac{\partial \Pi}{\partial t} &= \frac{\partial P}{\partial t} + \Omega \frac{d\phi}{dt}\frac{\partial P}{\partial X} = \frac{\partial P}{\partial t} + \Omega^{\frac{1}{2}}\frac{d\phi}{dt}\frac{\partial \Pi}{\partial \xi}\end{aligned}$$

Therefore we can rewrite (A4) as:

$$\begin{aligned}\frac{\partial \Pi}{\partial t} - \Omega^{\frac{1}{2}}\frac{d\phi}{dt}\frac{\partial \Pi}{\partial \xi} = \Omega\{&\Phi_0^+(\phi(t) + \Omega^{-\frac{1}{2}}(\xi - \Omega^{-\frac{1}{2}}))\Pi(\xi - \Omega^{-\frac{1}{2}}, t)\\ &+ \Phi_0^-(\phi(t) + \Omega^{-\frac{1}{2}}(\xi + \Omega^{-\frac{1}{2}}))\Pi(\xi + \Omega^{-\frac{1}{2}}, t)\\ &- (\Phi_0^+(\phi(t) + \Omega^{-\frac{1}{2}}\xi) + \Phi_0^-(\phi(t) + \Omega^{-\frac{1}{2}}\xi))\Pi(\xi, t)\}\end{aligned}$$

Taylor expanding $\alpha_1(\phi + \Omega^{-\frac{1}{2}}(\xi - \Omega^{-\frac{1}{2}})\Pi(\xi - \Omega^{-\frac{1}{2}})$ about ξ yields

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

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

$$\begin{aligned}\frac{\partial \Pi}{\partial t} - \Omega^{\frac{1}{2}}\frac{d\phi}{dt}\frac{\partial \Pi}{\partial \xi} = &-\Omega^{\frac{1}{2}}\alpha_1(\phi)\frac{\partial \Pi}{\partial \xi} - \alpha_1'(\phi)\frac{\partial \xi \Pi}{\partial \xi} - \frac{1}{2}\Omega^{-\frac{1}{2}}\alpha_1''(\phi)\frac{\partial \xi^2 \Pi}{\partial \xi}\\ &+ \frac{1}{2}\alpha_2(\phi)\frac{\partial^2 \Pi}{\partial \xi^2} + \frac{1}{2}\Omega^{-\frac{1}{2}}\alpha_2'(\phi)\frac{\partial^2 \xi \Pi}{\partial \xi^2} - \frac{\Omega^{-\frac{1}{2}}}{3!}\alpha_3(\phi)\frac{\partial^3 \Pi}{\partial \xi^3} + O(\Omega^{-1})\end{aligned}$$

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

$$\frac{d\phi}{dt} = \alpha_1(\phi).$$

$$\begin{aligned}\implies \frac{\partial \Pi}{\partial t} = &-\alpha_1'(\phi)\frac{\partial \xi \Pi}{\partial \xi} + \frac{1}{2}\alpha_2(\phi)\frac{\partial^2 \Pi}{\partial \xi^2}\\ &+ \frac{1}{2}\Omega^{-\frac{1}{2}}(\alpha_2'(\phi)\frac{\partial^2 \xi \Pi}{\partial \xi^2} - \alpha_1''(\phi)\frac{\partial \xi^2 \Pi}{\partial \xi} - \frac{1}{3!}\alpha_3(\phi)\frac{\partial^3 \Pi}{\partial \xi^3}) + O(\Omega^{-1})\end{aligned}$$

This is the final form of the expansion. It can be truncated at any level of detail desired and translated back into the original variables to yield various approximations of the Master equation. Note that it is only applicable for systems with a single stable steady-state, as discussed in the main text and in more detail in [VK07].

A8 Mean first-passage time

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

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

Then we have:

$$\begin{aligned}
\tau_k &= t_k(r_k \tau_{k-1} + g_k \tau_{k+1} + 1), \quad k = 0, \ldots, n-1 \\
\implies \tau_{k+1} - \tau_k &= \frac{1}{g_k}[r_k(\tau_k - \tau_{k-1}) - 1], \quad \text{by noting that } t_k(g_k + r_k) = 1 \\
\implies \tau_{k+1} - \tau_k &= \frac{1}{g_k} \sum_{i=0}^{j} \prod_{j=k}^{i-1} \frac{g_j}{r_j} = \frac{1}{g_k p_k^s} \sum_{i=0}^{j} p_i^s \\
\implies \tau_m &= \sum_{k=m}^{n-1} (\tau_{k+1} - \tau_k) = \sum_{k=m}^{n-1} \frac{1}{g_k p_k^s} \sum_{i=0}^{j} p_i^s,
\end{aligned}$$

where p^s is the stationary distribution (5.5). Observe that if n, m are stable points and l with $m \leq l \leq n$ is an unstable point, then the stationary distribution will have peaks at n and m and a valley at l. The most important terms in the sum are therefore those with p_l^s is the denominator, and inner sum is then π_m, which is $O(1)$. Hence the escape rate is on the order of p_l^s; that is,

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

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

A9 Coefficients of the bistable two-gene system

The coefficients of system (6.7) are given by:

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

Bibliography

[AJL$^+$07] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell.* Garland Science, 5th edition, 2007.

[AJS82] G. K. Ackers, A. D. Johnson, and M. A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proc. Natl. Acad. Sci. U.S.A.*, 79(4):1129–1133, Feb 1982.

[Alo07] U. Alon. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, 8(6):450–461, Jun 2007.

[ARM98] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage -infected escherichia coli cells. *Genetics*, 149(4):16331648, 1998.

[AW92] L. Avery and S. Wasserman. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet.*, 8(9):312–316, Sep 1992.

[BBAIdB07] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. How to infer gene networks from expression profiles. *Mol. Syst. Biol.*, 3:78, 2007.

[BBG$^+$05a] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: applications. *Curr. Opin. Genet. Dev.*, 15(2):125–135, Apr 2005.

[BBG+05b] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.*, 15(2):116–124, Apr 2005.

[BBO+09] Dmitry R Bandura, Vladimir I Baranov, Olga I Ornatsky, Alexei Antonov, Robert Kinach, Xudong Lou, Serguei Pavlov, Sergey Vorobiev, John E Dick, and Scott D Tanner. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical Chemistry*, 81(16):6813–6822, 2009.

[BHK98] Arie Bar-Haim and Joseph Klafter. Geometric versus energetic competition in light harvesting by dendrimers. *The Journal of Physical Chemistry B*, 102(10):1662–1664, 1998.

[BJGL+03] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, 21(11):1337–1342, Nov 2003.

[BKCC03] W. J. Blake, M. Krn, C. R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[Cho12] B. Choi. Learning Networks in Biological Systems, Ph.D. thesis, Department of Applied Physics, Stanford University, Stanford, California (thesis supervisor: W.H.Wong). 2012.

[CP08] V. Chickarmane and C. Peterson. A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. *PLoS One*, 3(10):e3478, 2008.

[Cri70] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[dBTG$^+$05] D. di Bernardo, M. J. Thompson, T. S. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. E. Schaus, and J. J. Collins. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat. Biotechnol.*, 23(3):377–383, Mar 2005.

[DCL$^+$08] M. J. Dunlop, R. S. Cox, J. H. Levine, R. M. Murray, and M. B. Elowitz. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature genetics*, 40(12):14931498, 2008.

[DESGS11] M. Dehmer, F. Emmert-Streib, A. Graber, and A. Salvador. *Applied Statistics for Network Biology: Methods in Systems Biology.* Wiley-VCH, 2011. to appear.

[DIB97] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, Oct 1997.

[DSM10] R. De Smet and K. Marchal. Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.*, 8(10):717–729, Oct 2010.

[DvdHM$^+$00] Joseph DeRisi, Bart van den Hazel, Philippe Marc, Elisabetta Balzi, Patrick Brown, Claude Jacq, and André Goffeau. Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. *FEBS letters*, 470(2):156–160, 2000.

[Ein06] Albert Einstein. Eine neue bestimmung der moleküldimensionen. *Annalen der Physik*, 324(2):289–306, 1906.

[EL00] M. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335338, 2000.

[ELSS02] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science Signaling*, 297(5584):1183, 2002.

[FCJ$^+$08] K. Foygel, B. Choi, S. Jun, D.E. Leong, A. Lee, C.C. Wong, E. Zuo, M. Eckart, R.A. Reijo Pera, W.H. Wong, and M.W. Yao. A novel and critical role for Oct4 as a regulator of the maternal-embryonic transition. *PLoS One*, 3(12):e4109, 2008.

[FCSI12] D. Frigola, L. Casanellas, J. Sancho, and M. Ibanes. Asymmetric stochastic switching driven by intrinsic molecular noise. *PloS one*, 7(2):e31407, 2012.

[FHT$^+$07] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, 5(1):e8, Jan 2007.

[Fri04] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, Feb 2004.

[GB08] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.

[GB11] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21, Apr 2011.

[GCC00] T. Gardner, C. Cantor, and J. Collins. *Construction of a genetic toggle switch inescherichia coli*, volume 403. Nature, 2000.

[GdBLC03] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629):102–105, Jul 2003.

[Gil77] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[Gil00] D. T. Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113:297, 2000.

[GKO+11] H. G. Garcia, J. Kondev, N. Orme, J. A. Theriot, and R. Phillips. Thermodynamics of biological processes. *Meth. Enzymol.*, 492:27–59, 2011.

[GMD12] P. Gutierrez, D. Monteoliva, and L. Diambra. Cooperative binding of transcription factors promotes bimodal gene expression response. *PloS one*, 7(9):e44812, 2012.

[HBS+07] M. Hegland, C. Burden, L. Santoso, S. MacNamara, and H. Booth. A solver for the stochastic master equation applied to gene regulatory networks. *Journal of computational and applied mathematics*, 205(2):708724, 2007.

[HGYI05] S. Huang, Eichler G., Bar-Yam Y., and D.E. Ingber. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys Rev Lett*, 94(12):128701, Apr 2005.

[HHG+10] L. Hartwell, L. Hood, M. Goldberg, A. Reynolds, and L. Silver. *Genetics: From Genes to Genomes.* McGraw-Hill Science/Engineering/Math, 4th edition, 2010.

[Hil13] A. V. Hill. The combinations of haemoglobin with oxygen and with carbon monoxide. *i. Biochemical Journal*, 7(5):471, 1913.

[HJW+98] F. C. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5):717–728, Nov 1998.

[HKI07] Z. Hu, P. J. Killion, and V. R. Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, 39(5):683–687, May 2007.

[HMC$^+$02] J. Hasty, D. McMillen, J. Collins, et al. Engineered gene circuits. *Nature*, 420(6912):224–230, 2002.

[HMJ$^+$00] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, Jul 2000.

[HMM09] Gordon L Hager, James G McNally, and Tom Misteli. Transcription dynamics. *Molecular cell*, 35(6):741–753, 2009.

[IKN06] Jonathan M Irish, Nikesh Kotecha, and Garry P Nolan. Mapping normal and cancer cell signalling networks: towards single-cell proteomics. *Nature Reviews Cancer*, 6(2):146–155, 2006.

[JM61] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, Jun 1961.

[KE01] T. Kepler and T. Elston. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical Journal*, 81(6):31163136, 2001.

[KEBC05] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, 2005.

[KH08] R. Khanin and D. J. Higham. Chemical Master Equation and Langevin Regimes for a Gene Transcription Model. *Theor. Comput. Sci.*, 104(1):31–40, 2008.

[KMK73] R. Kubo, K. Matsuo, and K. Kitahara. Fluctuation and Relaxation of Macrovariables. *J. Stat. Phys.*, 9:51–96, 1973.

[KS10] Mark Kittisopikul and Gürol M Süel. Biological role of noise encoded in a genetic network motif. *Proceedings of the National Academy of Sciences*, 107(30):13300–13305, 2010.

[KSK$^+$07] S. Krishnamurthy, E. Smith, D. Krakauer, W. Fontana, et al. The stochastic behavior of a molecular switching circuit with feedback. *Biology direct*, 2(1):1–17, 2007.

[LB03] S. L. Lacy and D. S. Bernstein. Subspace Identification with Guaranteed Stability using Constrained Optimization. *IEEE Trans. Automat. Control*, 48(7):1259–1263, 2003.

[LCH$^+$11] C.J. Lengner, F.D. Camargo, K. Hochedlinger, G.G. Welstead, S. Zaidi, S. Gokhale, H.R. Scholer, A. Tomilin, and R. Jaenisch. Oct4 expression is not required for mouse somatic stem cell self-renewal. *Cell Stem Cell*, 1(4):403–415, 2011.

[LQ10] J. Liang and H. Qian. Computational cellular dynamics based on the chemical master equation: A challenge for understanding complexity. *Journal of Computer Science and Technology*, 25(1):154–168, 2010.

[LRR$^+$02] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, Oct 2002.

[MBBS08] S. MacNamara, A. Bersani, K. Burrage, and R. Sidje. Stochastic chemical kinetics and the total quasi-steady-state assumption: Application to the stochastic simulation algorithm and chemical master equation. *The journal of chemical physics*, 129:095105, 2008.

[MLCW13] A. Meister, Y. H. Li, B. Choi, and W. H. Wong. Learning a nonlinear dynamical system model of gene regulation: A perturbed steady-state approach. *Annals of Applied Statistics*, 2013. arXiv preprint arXiv:1207.3137.

[MM13] L. Michaelis and M. L. Menten. Die kinetik der invertinwirkung. *Biochem. z*, 49(352.):333–369, 1913.

[MNvO12] B. Munsky, G. Neuert, and A. van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 336(6078):183187, 2012.

[MPS+10] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U.S.A.*, 107(14):6286–6291, Apr 2010.

[MTK09] B. Munsky, B. Trinh, and M. Khammash. *Listening to the noise: random fluctuations reveal gene network parameters*, volume 5. Mol Syst Biol, 2009.

[MWHL12] R. Ma, J. Wang, Z. Hou, and H. Liu. Small-number effects: A third stable state in a genetic bistable toggle switch. *Physical Review Letters*, 109(24):248107, 2012.

[MWM+08] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, Jul 2008.

[NT97] B. Novak and J. Tyson. Modeling the control of dna replication in fission yeast. *Proceedings of the National Academy of Sciences*, 94(17):9147, 1997.

[OBB+10] Olga Ornatsky, Dmitry Bandura, Vladimir Baranov, Mark Nitz, Mitchell A Winnik, and Scott Tanner. Highly multiparametric analysis by mass cytometry. *Journal of immunological methods*, 361(1):1–20, 2010.

[OTK+02] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature genetics*, 31(1):6973, 2002.

[OTL+04] E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. Van Oudenaarden. Multistability in the lactose utilization network of escherichia coli. *Nature*, 427(6976):737740, 2004.

[Pal11] B. Palsson. *Systems Biology: Simulation of Dynamic Network States.* Cambridge University Press, Cambridge, U.K., 2011.

[Pau04] Johan Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–418, 2004.

[PG58] Max Planck and Verband Deutscher Physikalischer Gesellschaften. Physikalische abhandlungen und vorträge. 1958.

[PMK06] S. Peles, B. Munsky, and M. Khammash. Reduction and solution of the chemical master equation using time scale separation and finite state projection. *The journal of chemical physics*, 125:204104, 2006.

[PSdlF10] A. Pinna, N. Soranzo, and A. de la Fuente. From knockouts to networks: establishing direct cause-effect relationships through graph analysis. *PLoS ONE*, 5(10):e12912, 2010.

[PvO05] J. M. Pedraza and A. van Oudenaarden. Noise propagation in gene networks. *Science*, 307(5717):1965–1969, Mar 2005.

[RA03] Christopher V Rao and Adam P Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the gillespie algorithm. *Journal of Chemical Physics*, 118(11):4999–5010, 2003.

[Ral08] A Ralston. Simultaneous gene transcription and translation in bacteria. *Nature Education*, 1(1), 2008.

[Ray91] Lord Rayleigh. Liii. dynamical problems in illustration of the theory of gases. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 32(198):424–445, 1891.

[RHB$^+$07] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, 4(8):651–657, Aug 2007.

[RO05] Jonathan M Raser and Erin K O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–2013, 2005.

[Ros11] S. Rosenfeld. Mathematical descriptions of biochemical networks: stability, stochasticity, evolution. *Prog. Biophys. Mol. Biol.*, 106(2):400–409, Aug 2011.

[RRW$^+$00] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–2309, Dec 2000.

[RVL$^+$07] R.T. Rodriguez, J.M. Velkey, C. Lutzko, R. Seerke, D.B. Kohn, K.S. O'Shea, and M.T. Firpo. Manipulation of OCT4 levels in human embryonic stem cells results in induction of differential cell types. *Exp Biol Med*, 232(10):1368–1380, Nov 2007.

[RvO08] A. Raj and A. Nature van Oudenaarden. nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

[RWA02] C. V. Rao, D. M. Wolf, and A. P. Control Arkin. Exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231–237, 2002.

[RYA+05] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–1965, Mar 2005.

[SA85] M. A. Shea and G. K. Ackers. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J. Mol. Biol.*, 181(2):211–230, Jan 1985.

[SES02] Peter S Swain, Michael B Elowitz, and Eric D Siggia. Intrinsic and extrinsic contributions to stochasticity in gene express ion. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002.

[SK12] P. Smadbeck and Y. Kaznessis. Stochastic model reduction using a modified hill-type kinetic rate law. *The Journal of Chemical Physics*, 137:234109, 2012.

[SMF11] T. Schaffter, D. Marbach, and D. Floreano. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–2270, Aug 2011.

[SOWES12] J. Stewart-Ornstein, J. S. Weissman, and H. El-Samad. Cellular noise regulons underlie fluctuations in i saccharomyces cerevisiae/i. *Molecular cell*, 45(4):483493, 2012.

[SSR+03] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, 34(2):166–176, Jun 2003.

[Tao04] Y. Tao. Intrinsic noise, gene regulation and steady-state statistics in a two-gene network. *Journal of theoretical biology*, 231(4):563568, 2004.

[TBO$^+$08] Scott D Tanner, Dmitry R Bandura, Olga Ornatsky, Vladimir I Baranov, Mark Nitz, MA Winnik, et al. Flow cytometer with mass spectrometer detection for massively multiplexed single-cell biomarker assay. *Pure and Applied Chemistry*, 80(12):2627–2641, 2008.

[TCN03] J. J. Tyson, K. C. Chen, and B. Novak. Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr. Opin. Cell Biol.*, 15(2):221–231, Apr 2003.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

[TvO01] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):86148619, 2001.

[TYHC03] J. Tegner, M. K. Yeung, J. Hasty, and J. J. Collins. Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proc. Natl. Acad. Sci. U.S.A.*, 100(10):5944–5949, May 2003.

[vHRGW74] P. H. von Hippel, A. Revzin, C. A. Gross, and A. C. Wang. Nonspecific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects. *Proc. Natl. Acad. Sci. U.S.A.*, 71(12):4808–4812, Dec 1974.

[VK65] NG Van Kampen. Fluctuations in nonlinear systems. *Fluctuation Phenomena in Solids, Academic Press, New York*, 1965.

[VK07] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry.* North Holland, 3rd edition, 2007.

[VM12] Christine Vogel and Edward M Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4):227–232, 2012.

[VS06] Marian Von Smoluchowski. Zur kinetischen theorie der brownschen molekularbewegung und der suspensionen. *Annalen der physik*, 326(14):756–780, 1906.

[Wal39] J. A. Walker. *Dynamical systems and evolution equations.* Plenum Press, New York, 1939.

[WSA$^+$99] Elizabeth A. Winzeler, Daniel D. Shoemaker, Anna Astromoff, Hong Liang, Keith Anderson, Bruno Andre, Rhonda Bangham, Rocio Benito, Jef D Boeke, Howard Bussey, et al. Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906, 1999.

[WWA10] S. Waldherr, J. Wu, and F. Allgowe. Bridging time scales in cellular decision making with a stochastic bistable switch. *BMC systems biology*, 4(1):108, 2010.

[YAYG10] K. Y. Yip, R. P. Alexander, K. K. Yan, and M. Gerstein. Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS ONE*, 5(1):e8121, 2010.

[ZCMW07] Q. Zhou, H. Chipperfield, D. A. Melton, and W. H. Wong. A gene regulatory network in mouse embyronic stem cells. *Proc. Natl. Acad. Sci. U.S.A.*, 408:16438–16443, 2007.

[ZJPP11] M. M. Zavlanos, A. A. Julius, Boyd. S. P., and G. J. Pappas. Inferring stable genetic networks from steady-state data. *Automatica*, 47(6):1113–1122, 2011.