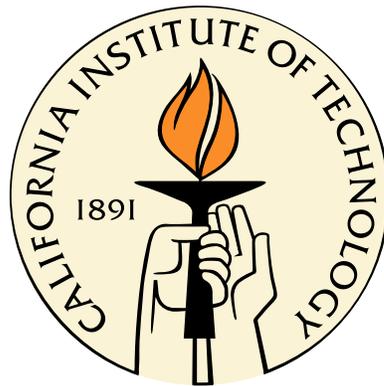


Dynamics and Correlated Noise in Gene Regulation

Thesis by
Mary J. Dunlop

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy



California Institute of Technology
Pasadena, California

2008
(Defended May 21, 2008)

© 2008

Mary J. Dunlop

All Rights Reserved

Acknowledgments

I would like to begin by thanking my advisor, Richard Murray, for his enthusiasm and encouragement. I appreciate Richard's intellectual curiosity, organization, and general level-headedness. He has been a fantastic advisor and a large part of my happiness in graduate school is to his credit.

Michael Elowitz has been like a second advisor to me and I very much appreciate having the opportunity to work in his lab. Michael has an uncanny ability to ask simple questions that get to the heart of an interesting scientific problem; he can also distill confusing and complicated results into simple conclusions. I have learned a great deal from working with him and interacting with other members of the Elowitz lab.

I would also like to thank Erik Winfree and Rob Phillips for serving on my thesis committee.

Mike Wall at Los Alamos National Lab deserves a lot of credit for introducing me to molecular biology and biological modeling. His patience and thorough explanations were very useful as I was getting started.

I have had the good fortune to work with several other students and would specifically like to thank: Elisa Franco, my fellow Murray Group bio-control expert, good friend, and conference companion. R. Sidney Cox, for helping me a great deal with initial experiment designs and for his excellent advice on cloning and imaging. Joe Levine, for his help on my project and being a good sounding board for ideas. Hernan Garcia, for his exceedingly open willingness to help. And Michelle Fontes, who taught me quite a lot of practical molecular biology. I have also benefited from discussions at the bio-control journal club and, in particular, would like to thank Josh Michener and Katie Galloway.

Finally I would like to thank my parents, Lynn and Al, and my brother, David, for all of their encouragement. And Asa for his constant support.

Abstract

Gene regulatory interactions are context dependent, active in some cell types or cellular states but not in others. In this thesis we present a method for determining when a regulatory link is active given temporal measurements of gene expression. Correlations in time-series data are used to determine how genes influence each other and their causal relationships. Natural stochastic noise is shown to aid in the process of network identification by perturbing the expression of genes; the speed and direction at which the noisy signal propagates shows how the network is connected. Cross correlation functions are used to reveal time-delayed correlations.

We develop a stochastic model of gene expression and show that by measuring correlations in cellular noise, it is possible to infer network activity and temporal properties of gene regulation. Using a linearized version of the model, we introduce a method for analytically deriving cross correlation functions for arbitrary networks. These results are validated experimentally using a synthetic gene circuit in *E. coli* bacteria. Single-cell time-lapse microscopy is used to measure noisy expression of multiple genes over time. Extending this work to natural systems, we study feed-forward loops and determine that certain classes of feed-forward loops are more robust to noise and parameter variations than others. Noise in two naturally occurring feed-forward loops involved in galactose utilization is measured experimentally and it is shown that neither is actively regulating its target in the conditions tested.

Contents

Acknowledgments	iii
Abstract	v
Contents	vii
1 Introduction	1
1.1 Network Motifs	2
1.2 Context-Sensitive Regulation	3
1.3 Thesis Overview	8
2 Numerical and Analytic Predictions for Correlations in Gene Expression Noise	9
2.1 Mathematical Models with Noise	9
2.2 Simulation Results	12
2.3 Cross Correlation Functions	12
2.4 Analytic Solutions for Cross Correlation Functions Due to Noise	16
2.4.1 Unregulated Case	17
2.4.2 Regulated Case	19
2.4.3 Summary	20
2.5 Sensitivity Analysis	21
2.6 Transcriptional Cascades	22
2.7 mRNA dynamics	23
2.8 Degenerate Cross Correlation Functions	25
2.8.1 Redundant Network Elements	25
2.8.2 Parametric Degeneracies	26
2.8.3 Implications for Network Identification	27
3 Correlated Noise in a Synthetic Gene Circuit Reveals Regulation	29
3.1 Synthetic Construct	29
3.2 Static Data	31

3.3	Temporal Data	32
3.4	Experimental Cross Correlations	36
3.4.1	Extension of the Cross Correlation Function to Branched Data	36
3.4.2	Experimental Noise Correlations	37
3.5	Alternative Construct	39
3.6	Methods and Characterization	40
3.6.1	Bleaching Times	40
3.6.2	Methods	43
4	Robustness in Feed-Forward Loops: Clustering of Responses	45
4.1	Background	45
4.2	Mathematical Models	46
4.3	Simulations	47
4.4	Clustering Feed-Forward Loop Responses	48
4.4.1	Clustering Algorithm	49
4.4.2	Maximum Error versus Number of Clusters	49
4.4.3	Singular Value Decomposition	50
4.5	Results	51
4.5.1	Representative Cluster Traces	51
4.5.2	Distributions of Responses	53
5	Noise in Feed-forward Loops for Galactose Utilization	57
5.1	Galactose Regulation	57
5.1.1	Galactose Metabolism	58
5.1.2	Galactose Transport	60
5.1.3	Structure of Regulatory Networks	60
5.2	Clustering of Type I Incoherent Feed-Forward Loop Responses	61
5.3	Context Sensitive Regulation	62
5.4	Noise Correlations to Infer Activity	64
5.4.1	Theoretical Predictions	64
5.4.2	Experimental Data	67
5.5	Methods and Characterization	68
5.5.1	Expression Levels from Static Data	68
5.5.2	Methods	69
6	Conclusion	71
6.1	Summary of Contributions	71

6.2	Future Work	72
6.2.1	Biological Persistence of Excitation	72
6.2.2	Network Identifiability	72
6.2.3	Monitoring Dynamic Changes	72
6.2.4	External Inputs for System Identification	73
6.2.5	Context-Sensitive Maps of Gene Regulation	73
A	Mathematica Code for Calculating Cross Correlations of a Cascade	75
B	Synthetic Circuit Movie Frames	77
C	Full Table of Feed-Forward Loop Cluster Percentages	83
	Bibliography	89

Chapter 1

Introduction

Cells use biological circuits to implement diverse cellular and developmental programs. All the information required to construct and control these circuits is encoded in the genome of an organism. Genome complexity spans many orders of magnitude, from the bacterium *C. ruddii* with 160 thousand base pairs of DNA to the amoeba *A. dubia* with 670 billion base pairs [1, 2]. A spectacularly diverse range of organisms fall in between, including the human genome with 3 billion base pairs and *E. coli* bacteria, used for the experiments in this thesis, with 4.6 million. Despite dramatic differences in genome size, the fundamental way information is encoded is conserved across organisms and many of the basic mechanisms for implementing genetic control are universal. In this thesis we will look at examples of genetic control and determine when they are actively being used.

Genes are regions of DNA in the genome that encode for proteins. Humans have about 20,000 genes while *E. coli* have 4000 [1, 3]. Proteins provide a useful function to the cell or may control the expression of other genes. An example of the former is an enzyme that breaks down sugar to fuel the cell. Proteins in the latter category are known as transcription factors. They act by binding to a sequence of DNA upstream of a gene, known as the promoter, and either repressing or activating transcription (see Fig. 1.1).

Since proteins have the ability to control the expression of other genes, they can regulate themselves [4, 5, 6], or a host of other genes. As a result, networks of gene regulation appear when one transcription factor regulates many genes, including other transcription factors. These gene networks (also called gene circuits) are often elaborate and include a wide variety of control strategies, including feed-back and feed-forward loop architectures like those used in engineered control systems. A major goal of systems biology is to connect the regulatory architecture of these networks to the dynamic behavior of individual cells. Biochemistry and genetics can efficiently identify regulatory interactions and there are databases that summarize all documented regulatory connections [7], but it can be unclear what function these networks serve.

From an analytic point of view, it is frequently difficult to use gene network maps to understand and predict cellular behaviors. One problem is that quantitative information about the biochemical

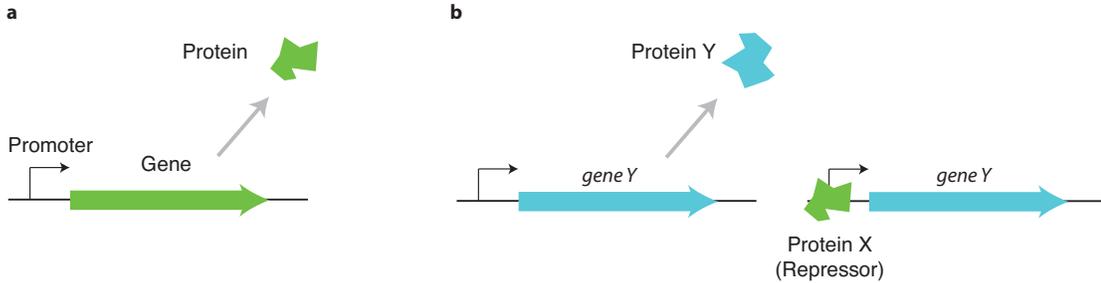


Figure 1.1: Control of gene expression. (a) This diagram depicts a small section of the DNA that is part of the larger genome. A promoter region upstream of the gene is used to control expression of the gene. When gene expression is ON, the gene is transcribed and then translated into a protein. (b) Example of a transcription factor. The gene Y is expressed as a protein in the absence of the transcription factor, protein X, but expression is inhibited when it is present. This is an example of a repressor; activators act in the opposite fashion.

parameters associated with a given regulatory link is often minimal. Also, the complexity of the gene regulatory architecture, even for simple organisms, begs the question of whether cells actually use all of this cellular control at a given time. Many regulatory links, while important in some cellular states, may not be active in a given cellular context. If one can identify the subset of regulatory links that are active in a given state, it could simplify analysis of the circuit as a whole [8].

1.1 Network Motifs

Gene network structure is the focus of several notable studies that search for recurring patterns in databases of documented regulatory connections. If a particular pattern of connections appears more frequently than would be expected in an entirely random network then, in theory, evolution has selected for these so-called *network motifs* because they are useful to the organism [9, 10]. One of the best studied network motifs in model organisms like *E. coli* and the yeast *S. cerevisiae* is the feed-forward loop [11, 12], where two transcription factors regulate the expression of a single downstream gene and one of the transcription factors controls expression of the other (Fig. 1.2). Because there are two pathways that control the final target gene they can act differentially to achieve desirable temporal effects.

Even once network motifs are identified, there is not an immediate map to the functional role they play in regulation. The parameters that describe chemical reaction rates play a large role in determining function. In addition, signaling molecules can bind to network elements rendering them inactive, or enhancing their activity. Several studies look at roles that network motifs can play [5, 12], though this analysis always requires assumptions about network properties. Further, motif analysis often neglects to account for the cellular environment in which the gene circuits are embedded. Work on rewiring natural regulatory networks in *E. coli* suggests that regulatory modules can be strongly

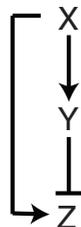


Figure 1.2: Feed-forward loop network motif. X and Y are transcription factors that control Z . Regular arrows indicate activation; T-shaped arrows indicate repression.

affected by the surrounding network [13].

Pattern-searching approaches like those employed to identify network motifs are one way of making analysis of large gene networks tractable. A complementary approach is to ask which regulatory links are functional in a particular condition. Ignoring inactive links can reduce network complexity.

1.2 Context-Sensitive Regulation

Regulatory links may be present, but inactive, for several reasons: In the simplest case, the concentration of a regulatory factor may be well above or below its effective range (Fig. 1.3). For example, it has been shown that cells may maintain transcription factor concentrations outside of their active regulatory regime in order to suppress noise [14]. In other cases, transcription factors may be inactive due to post-translational modification or the absence of necessary co-factors, rendering the transcription factor ineffective on a given target [15, 16, 17]. Since the activity of a regulatory link is highly dependent upon the conditions in which it is operating, we ask: Can the activity of regulatory links in a given cellular state be inferred non-invasively?

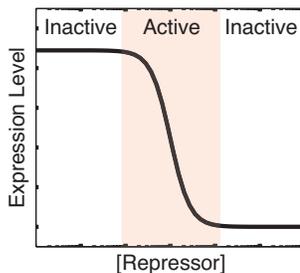


Figure 1.3: Target gene expression versus repressor concentration (schematic). Regions where changes in repressor concentration cause changes in target gene expression are active. When repressor levels saturate (right) or are insufficient to repress (left), then the link is inactive. Regulatory links may be inactive for other reasons as well.

Recent work has shown that noise in gene expression can generate substantial cell-cell variability [18, 19, 20]. Systematic measurements of noise across many genes have helped to broaden under-

standing of where noise comes from and how its effects are mitigated [21, 22]. Although noise plays a role across a broad spectrum of species, some of the best studied examples come from single-celled organisms. For example, when *E. coli* bacteria divide they produce two genetically identical clones. Despite genetic similarities, two cells can show dramatic differences in levels of gene expression. Fig. 1.4 shows a small colony of *E. coli* bacteria that started as a single cell at $t = 0$. Cells elongate and split into two daughters approximately once per hour. A gene encoding for a fluorescent protein was put into the genome of the cells and its expression was monitored over time using a microscope. When the cells divide, the fluorescence expression level drops by approximately a factor of two. By the end of the measurements each cell in the microcolony has a unique level of fluorescent protein.

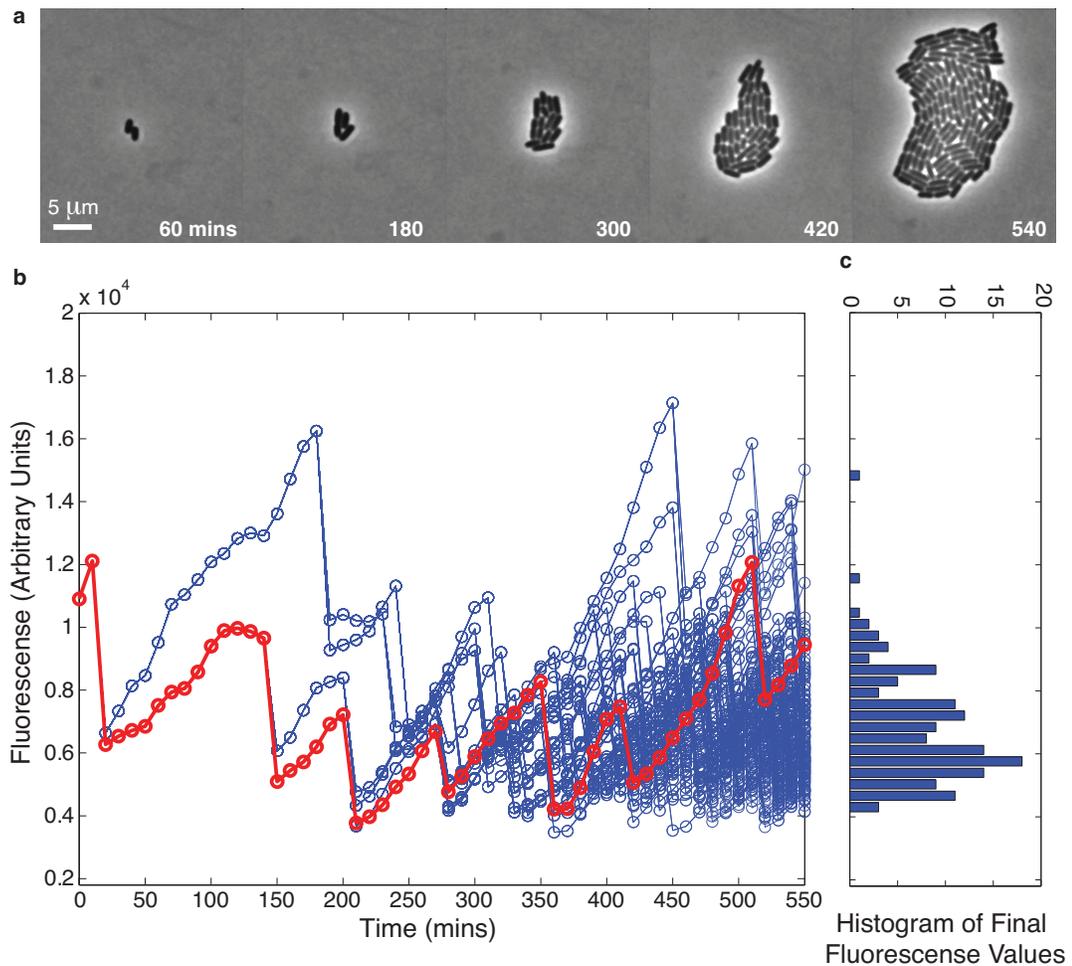


Figure 1.4: Cell-cell variability in *E. coli* gene expression. (a) Snapshots of single cells. The microcolony originated from a single cell. Time and length scales are indicated on the figure. (b) Fluorescent protein expression levels versus time. Blue lines are data for all cells. The red trace follows a single lineage. Sharp drops in expression are cell division events. (c) Distribution of final protein expression levels shows variability in gene expression.

There are several mechanisms that cause diversity among cells, but the underlying reason is small

size. A single *E. coli* bacterium is about 1×10^{-15} L in volume, or $1 \mu\text{m}^3$. Consequently, the number of important proteins, genes, and other molecules of interest in the cell may be small enough that the timing of individual reactions and locations of individual molecules can matter.

We asked whether noise could be used to reveal active regulation. In the context of a transcriptional regulatory circuit, noise in the concentration of a transcription factor can only propagate through one or more active regulatory links. Thus noise may provide information about active regulatory connections without explicit perturbation of cellular components.

Fig. 1.5 illustrates how noise could be used to infer the activity of a regulatory connection. Consider two possible types of interactions between proteins *A* and *B*: an inactive regulatory link and active repression link. If there were no noise, all cells would have exactly the same number of proteins. (If this were true all the lines in Fig. 1.4b would fall onto one line.) Thus a plot of *A* vs. *B* would have data from all cells collapse onto a single point (Fig. 1.5a). Realistically, individual cells show a range of protein concentrations so *A* versus *B* will show a range of points on a plot (Fig. 1.5b, each point represents a single cell). Active regulation between two proteins will result in correlated patterns. Cells that have a small amount of repressor, *A*, will have a larger amount of its target, *B*, which appears as a negative correlation between *A* and *B*.

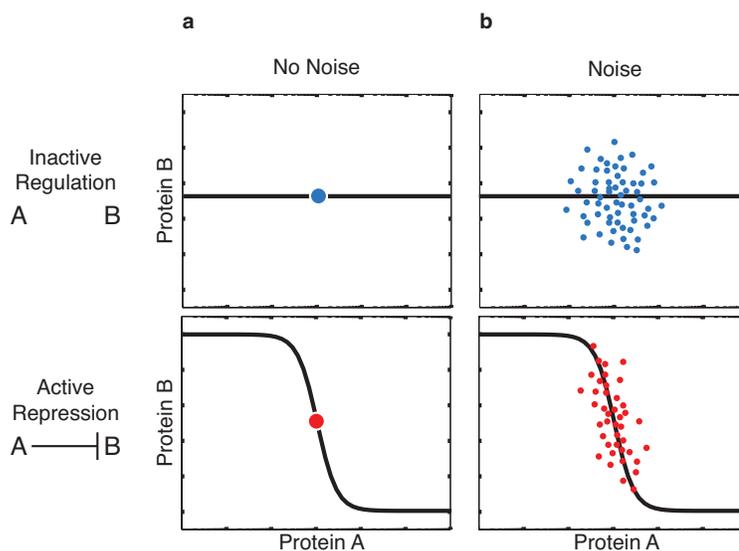


Figure 1.5: Correlations in noise distributions show active regulation. Each dot represents a single cell. Gene regulation function is shown as a solid black line. (a) Without noise, all cells show the same levels of expression. (b) With noise, correlations can suggest active regulation. T-shaped arrow between *A* and *B* indicates repression. Note that this schematic assumes that noise between two independent genes is uncorrelated (intrinsic noise only).

Noise propagation through active regulatory links is not the only factor correlating expression between genes. The expression of many or all genes in the cell may be correlated due to global variations in the overall rate of gene expression. In prior work, noise was broken down into two

broad classes: extrinsic and intrinsic noise [19, 23]. Noise sources that are global to a single cell, but vary from one cell to the next are extrinsic noise sources. For example, fluctuating cell size, numbers of ribosomes, and polymerase components can affect the expression of all genes in a cell; a cell that has a small number of polymerases will produce fewer proteins than a cell with many. Intrinsic noise, in contrast, is specific to an individual gene. Expression of a protein requires many discrete chemical reactions to happen and the timing and order of these reactions is a stochastic process. Consequently, even two identical genes may be expressed at different levels. Thus extrinsic noise can be thought of as global to a single cell, while intrinsic noise is local to a particular gene.

Because extrinsic noise acts globally it positively correlates the expression of all genes. Intrinsic noise, in contrast, is uncorrelated between genes. Fig. 1.6 illustrates how these conflicting effects prevent discrimination between noise and regulation as a source of correlation. Positive correlations from extrinsic noise superimposed on negative correlations due to repression can look very similar to uncorrelated genes. Thus, static correlation-based reasoning may not correctly identify active regulatory interactions.

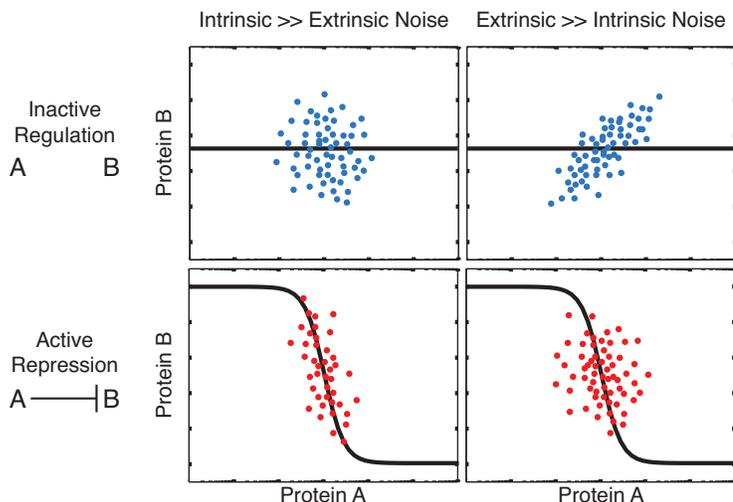


Figure 1.6: Noise can produce different types of static correlations. In each plot, dots represent individual cells from a hypothetical population. Top plots show correlations without an active regulatory link, while the bottom show correlations with active repression. We consider noise regimes in which either intrinsic (uncorrelated) or extrinsic (correlated) noise dominates. Active repression causes negative correlations between the transcription factor and its target, intrinsic noise decorrelates the two, and extrinsic noise causes positive correlations even without active regulation. Thus, correlations derived from static snapshots are ambiguous.

Gene regulation occurs with a delay; it takes time for protein concentrations to build up sufficiently to have a regulatory effect on the downstream genes they control (Fig. 1.7) [19]. The sign of the delay provides information about the direction of the link. Such a delay does not occur for extrinsic noise, which affects all genes simultaneously. Thus, dynamic measurements, in which one can follow the expression of multiple genes over time, can be used to decouple noise from regulatory

correlations.

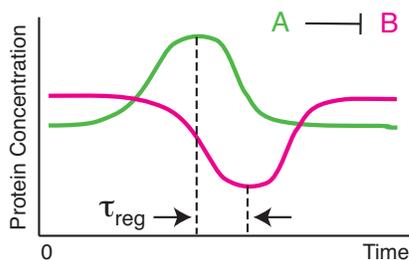


Figure 1.7: Temporal gene expression patterns for a repressor, A , (green line) and its target, B , (magenta line) are anti-correlated at a delay time denoted τ_{reg} (schematic).

Since time is an important factor in distinguishing between correlations due to extrinsic noise and those due to active regulation, we use the cross correlation function, which is a measure of how well two signals are correlated when one signal is shifted by a time τ . A positive pulse in a signal $f(t)$ followed by a negative pulse in a signal $g(t)$ will appear as a dip in the cross correlation function where the minimum occurs at a time $\tau < 0$ when the signals are maximally anti-correlated (Fig. 1.8). Note that whether the dip occurs at $\tau < 0$ or $\tau > 0$ indicates which pulse appeared first. This information is useful for inferring the direction of gene regulation, while the sign of the peak indicates the type of regulation (activation or repression). Similar approaches have been used to infer connectivity of *in vitro* metabolic networks [24, 25]. These experiments were not conducted in living cells so they used a prescribed time-varying input to perturb the system and it was unnecessary to consider many of the details particular to cellular noise sources. Other work has used temporal correlations in combination with large-scale microarray experiments [26].

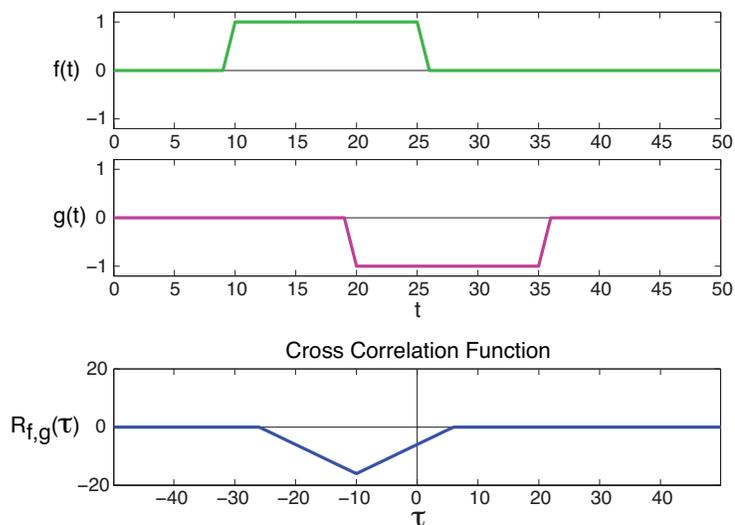


Figure 1.8: Two signals $f(t)$ and $g(t)$ are compared using the cross correlation function $R_{f,g}(\tau)$.

Inferring how gene networks are connected is a large area of research and methods for identifying

the connectivity of regulatory networks have improved as genetic assays have advanced. Large-scale network identification has focused on steady-state measurements and statistical inference algorithms [27, 28, 29, 30]. Alternative approaches for smaller-scale networks have been suggested using temporal data [31, 32, 33]. The work in this thesis focuses on smaller-scale network identification using temporal data, but extends current understanding by allowing for non-invasive measurements of network activity.

1.3 Thesis Overview

In Chapter 2 we develop a mathematical model for gene expression in a simple circuit. Numerical simulations of noisy gene expression are used to calculate cross correlation functions under different conditions, such as environments that are dominated by extrinsic or intrinsic noise. Using simulations, we explore the effect of active and inactive regulatory links. Further, an analytic method for calculating cross correlation functions due to noise is developed and shows excellent agreement with the full nonlinear simulations.

Chapter 3 tests these predictions experimentally with a three-color synthetic gene circuit. Time-lapse movies are used to measure gene expression in single cells under two noise conditions, where one has an appreciably higher level of extrinsic noise than the other. The resulting cross correlation functions are consistent with those predicted from simulation.

Mathematical models of feed-forward loops are explored in Chapter 4 and it is shown that certain feed-forward loop architectures are more robust to parameter variation than others. Chapter 5 looks at two examples of naturally occurring feed-forward loops in the galactose utilization pathway. By measuring correlations in gene expression noise we determine these feed-forward loops are inactive in measurement conditions, but can become active as a result of changes in the genetic background.

Chapter 2

Numerical and Analytic Predictions for Correlations in Gene Expression Noise

In this chapter we develop a mathematical model for stochastic gene regulation that is motivated by previous work on noise in λ cI repressor [19]. Simulated time-series data are used to predict features of cross correlation functions and to understand the effect of active and inactive forms of regulation. We extend the simulation-based analysis by developing analytic solutions to arbitrary cross correlation functions using a linearized approximation of the mathematical model. The analytic framework lends insight into the origins of the cross correlation function shape and simplifies analysis. Two applications are demonstrated: (1) Sensitivity analysis of cross correlation function features reveals which system parameters are most significant. (2) The cross correlation function is calculated for a cascade of arbitrary length to demonstrate generality to larger scale networks. Finally we discuss some limitations of correlation-based analysis methods.

2.1 Mathematical Models with Noise

We analyzed a simple three-gene circuit, shown in Fig. 2.1. The protein A is a transcription factor that represses production of B . Proteins A and C are constitutively expressed, meaning they are produced at a constant level and are not under the control of other transcription factors. A deterministic model for this system can be written using a Hill function to describe repression [34, 35]:

$$\dot{A} = \alpha_A - \beta A \tag{2.1}$$

$$\dot{B} = \frac{\alpha_B}{1 + (A/K)^n} - \beta B \tag{2.2}$$

$$\dot{C} = \alpha_C - \beta C. \tag{2.3}$$

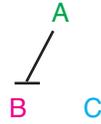


Figure 2.1: Three-gene circuit used in simulation. T-shaped arrow indicates repression of B by A .

In this model each protein is produced at a rate α_i and decays at a rate β ; the parameters K and n determine the properties of repression of B by A . The decay rate for all three proteins is assumed to be the same. This is a reasonable assumption if the proteins are stable (do not break down quickly). There are two ways that the level of protein in a cell can decrease: (1) As the cell grows in size the protein will dilute and (2) if the cell has an active mechanism for degrading proteins, such tagging them for recognition by proteases. Stable proteins are only governed by the first form of decay and thus their degradation is dependent upon the rate of cell division. By approximating dilution by a decaying exponential function, β can be calculated as

$$\beta = \frac{\log(2)}{T_{cc}}, \quad (2.4)$$

where T_{cc} is the length of the cell cycle.

The deterministic model has a single equilibrium point

$$A_{eq} = \frac{\alpha_A}{\beta} \quad (2.5)$$

$$B_{eq} = \frac{\alpha_B}{\beta(1 + (A_{eq}/K)^n)} \quad (2.6)$$

$$C_{eq} = \frac{\alpha_C}{\beta} \quad (2.7)$$

that is stable for all realistic biological parameters ($\alpha_i, \beta, K > 0$).

A more realistic model of gene expression accounts for noise in the expression of genes. Here we model the two classes of noise discussed in the Introduction. Extrinsic noise is assumed to affect all genes in the same way, while intrinsic noise is distinct for each individual gene. Thus, we add an extrinsic noise term, E , and intrinsic noise terms, I_i for $i = \{A, B, C\}$, to Eqns. (2.1)–(2.3):

$$\dot{A} = E + I_A + \alpha_A - \beta A \quad (2.8)$$

$$\dot{B} = E + I_B + \frac{\alpha_B}{1 + (A/K)^n} - \beta B \quad (2.9)$$

$$\dot{C} = E + I_C + \alpha_C - \beta C. \quad (2.10)$$

By setting the mean of these noise processes to zero we preserve the average equilibrium point of the system. We model noise as additive; other models, such as multiplicative noise, give similar qualitative results (not shown). Other properties of the noise sources are modeled explicitly using

biologically realistic parameters, described below.

Cellular noise sources have a finite correlation time that has been measured experimentally [19, 36]. We use Ornstein-Uhlenbeck processes to model noisy gene expression [37, 38]. These processes are described by the statistical values of mean, standard deviation, and correlation time, generating a noisy data trace that is continuous, allowing for numerical integration. In general, an Ornstein-Uhlenbeck process $X(t)$ can be written as

$$\frac{dX(t)}{dt} = -\frac{1}{\tau}X(t) + c^{1/2}\eta(t), \quad (2.11)$$

where τ is the correlation time of $X(t)$, c is a diffusion constant, and $\eta(t)$ is Gaussian white noise. As calculated in [38], as $t \rightarrow \infty$

$$\text{mean}\{X(t)\} = 0 \quad (2.12)$$

$$\text{var}\{X(t)\} = \frac{c\tau}{2}. \quad (2.13)$$

Rewriting Eqn. (2.11) in terms of the standard deviation of the noise, σ , and the correlation time, τ , we have

$$\dot{X} = -\frac{1}{\tau}X + \sqrt{\frac{2}{\tau}}\sigma\eta.$$

Extrinsic and intrinsic noise are modeled using Ornstein-Uhlenbeck processes by

$$\dot{E} = -\beta E + \theta\eta_E \quad (2.14)$$

$$\dot{I}_i = -\kappa I_i + \lambda_i\eta_i. \quad (2.15)$$

We assume the white noise terms η_E , η_A , η_B , and η_C are independent, identically distributed processes. The parameters β and κ define the time scale of the noise, while θ and λ set its standard deviation. These values were measured directly in [19]. Extrinsic noise was found to have a correlation time on the order of the cell cycle, while intrinsic noise had a much shorter characteristic time scale, $T_{int} = 5$ minutes. Thus β is described by Eqn. (2.4) (identical to the decay time of proteins) and $\kappa = \log(2)/T_{int}$. The standard deviation of extrinsic noise was measured directly for the λ cI system in [19] as $\sigma_{ext} = 0.35$. We assume the standard deviation of the intrinsic noise is related to the signal strength by setting $\sigma_{int,i} = \sqrt{\alpha_i}$, as in a Poisson process.

2.2 Simulation Results

We simulated the noisy expression of proteins A , B , and C numerically using the differential equations given in Eqns. (2.8)–(2.10) and (2.14)–(2.15) with the parameters listed in Table 2.1.

Parameter	Value	Notes/Reference
α_A	1.39 molecules/cell/min	chosen so that $\alpha_A/\beta = K$
α_B	4.5 molecules/cell/min	arbitrary
α_C	1.39 molecules/cell/min	chosen to match α_A
β	0.0116 1/min	calculated from Eqn. (2.4) assuming 60 min cell cycle
K	120 nM	[19]
n	1.7	[19]
κ	0.139 1/min	[19]
θ	0.0532 (molecules/cell) ^{1/2} /min	from Eqn. (2.13) and [19] $\sqrt{\frac{2}{T_{cc}}}\sigma_{ext}$
λ_A	0.621 (molecules/cell) ^{1/2} /min	from Eqn. (2.13) $\sqrt{\frac{2\alpha_A}{T_{int}}}$
λ_B	1.12 (molecules/cell) ^{1/2} /min	from Eqn. (2.13) $\sqrt{\frac{2\alpha_B}{T_{int}}}$
λ_C	0.621 (molecules/cell) ^{1/2} /min	from Eqn. (2.13) $\sqrt{\frac{2\alpha_C}{T_{int}}}$
T_{cc}	60 mins	measured from experiments
T_{int}	5 mins	[19]

Table 2.1: Simulation Parameters

Sample simulation traces are shown in Fig. 2.2. The equilibrium point calculated from the deterministic model (Eqns. (2.5)–(2.7)) is an accurate description of the average behavior. The different time scales associated with extrinsic and intrinsic noise are apparent when considering limiting cases where only one source of noise exists, as seen in Fig. 2.3. With only extrinsic noise the three signals are positively correlated, though expression of B is also repressed by A . The cell cycle length governs the time scale of fluctuations. With only intrinsic noise, repression of B by A generates a delayed anti-correlation in the expression of these two genes. The intrinsic noise time scale is much faster than that due to extrinsic noise.

2.3 Cross Correlation Functions

The cross correlation between two signals $f(t)$ and $g(t)$ is $R_{f,g}$, defined as

$$\begin{aligned}
 S_{f,g}(\tau) &= \begin{cases} \frac{1}{N-|\tau|} \sum_{n=0}^{N-\tau-1} \tilde{f}(n+\tau)\tilde{g}(n) & \tau \geq 0 \\ S_{g,f}(-\tau) & \tau < 0 \end{cases} \\
 R_{f,g} &= \frac{S_{f,g}(\tau)}{\sqrt{S_{f,f}(0)S_{g,g}(0)}}, \tag{2.16}
 \end{aligned}$$

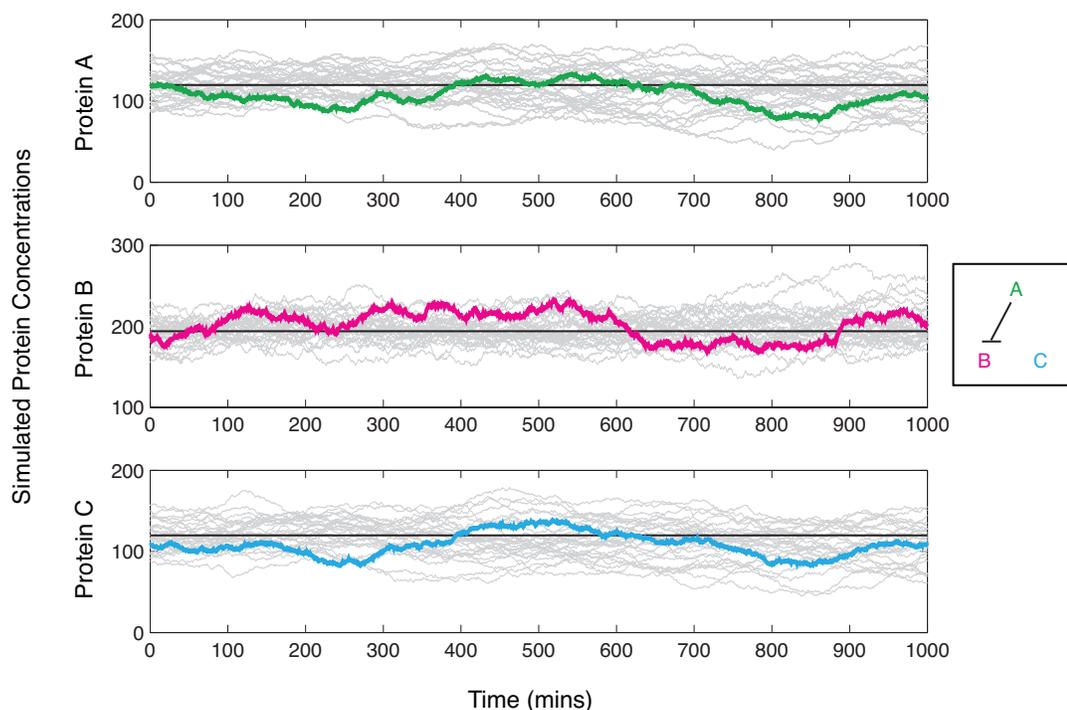


Figure 2.2: Simulated protein concentrations. Gray traces show 30 numerical examples of noisy gene expression. Black line indicates the steady-state equilibrium point. Individual traces from a single simulation are colored to show representative data.

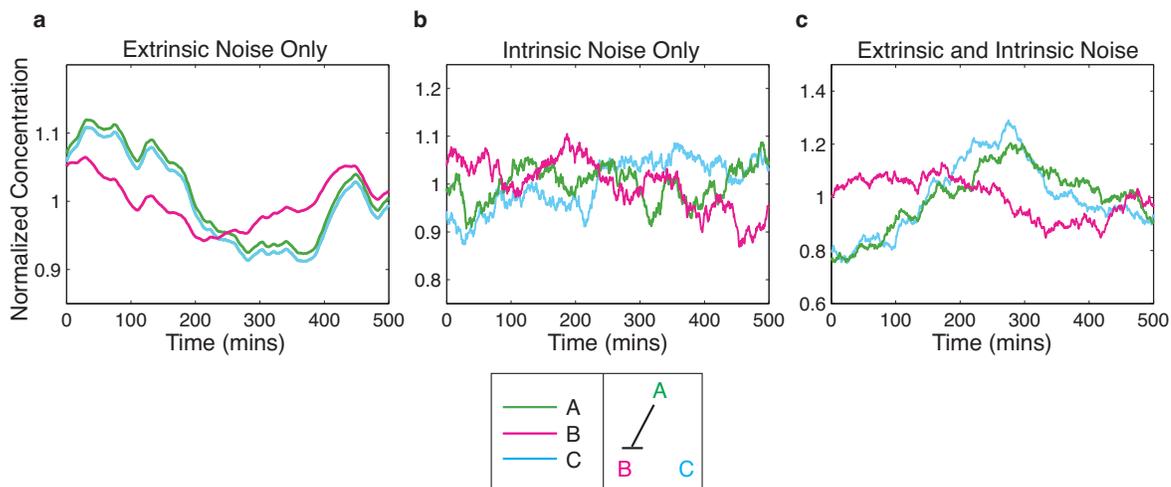


Figure 2.3: Sample simulation data under different types of noise. Protein A (green line) represses production of protein B (magenta line). Protein C (cyan line) is expressed constitutively. Data are normalized by mean concentration. Note that extrinsic noise positively correlates the time traces, while fluctuations in A (green curve) produce opposite fluctuations in B (magenta) at a delay (τ_{reg}). Simulated time traces are shown for three noise regimes, as indicated.

where N is the number of time points, τ is the time shift, and

$$\tilde{f} = f - \frac{1}{N} \sum_{n=0}^{N-1} f(n).$$

The mean-subtracted version of the cross correlation function, is sometimes referred to as the cross covariance.

Temporal correlations were used to measure the propagation of noise through a network. Individual cross correlation curves were generated by calculating results for two time series of noisy gene expression; the mean values of many individual cross correlation functions describe the average behavior (Fig. 2.4). Although the expected value of the cross correlation function due to noise is a fairly simple curve, obtaining it requires generating statistics over many sets of time-series data.

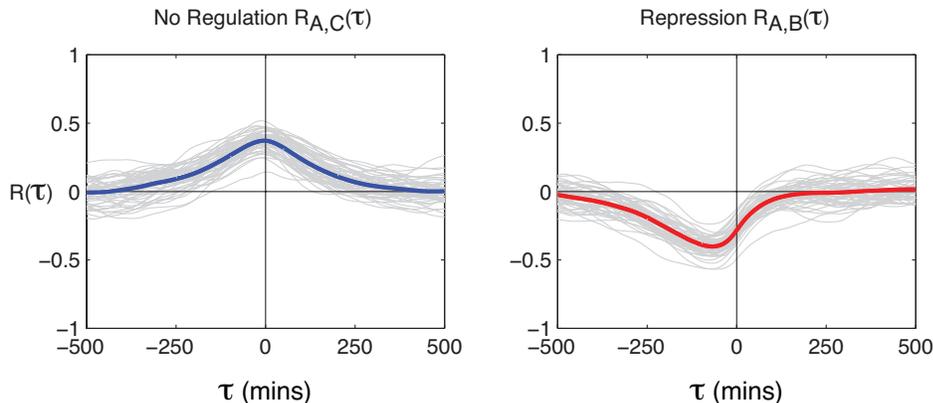


Figure 2.4: Cross correlation functions of simulated gene expression data. Gray curves show 50 cross correlation curves from individual time-series data. Mean values are shown in blue (no regulation) and red (repression).

Mean cross correlation functions for different noise regimes are shown in Fig. 2.5. Several features are apparent: (1) Repression appears as a dip at a delay time denoted by τ_{reg} , the effective regulation time. (2) The direction of regulation is given by the sign of τ_{reg} . Since A represses B, the dip occurs at $\tau_{reg} < 0$. (3) Extrinsic noise causes a positive peak in the cross correlation function close to $\tau = 0$, both with and without regulation. (4) The relative balance of intrinsic and extrinsic noise affects the magnitude of τ_{reg} . Together, these results indicate that cross correlation analysis, in combination with an understanding of physiological levels and types of noise, can be used to analyze the activity and direction of regulatory links.

Simulations were used to explore the effect of network activity on the shape of the cross correlation function. In Fig. 2.6 we varied the ratio of A_{eq} to K , which sets the position of the input on the sigmoidal Hill function curve. The dip is largest when A is in the region of the Hill function with the steepest slope (red dots). As A_{eq} moves to the saturating regions on the Hill function the magnitude

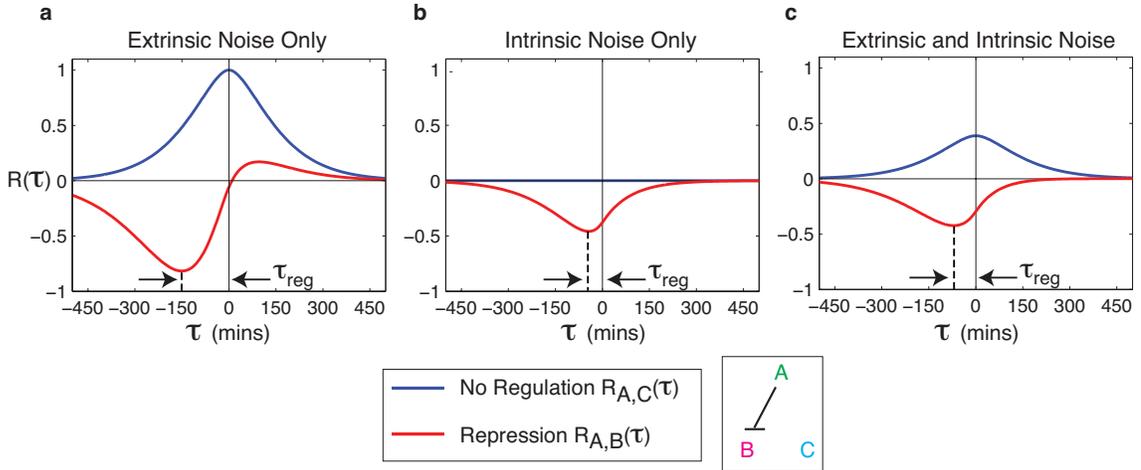


Figure 2.5: Mean cross correlation functions $R_{A,B}(\tau)$ and $R_{A,C}(\tau)$ are shown in red and blue, respectively. Note that active negative regulation causes a dip at τ_{reg} , while extrinsic noise results in positive correlation near $\tau = 0$.

of the dip in the cross correlation returns to zero (blue dots).

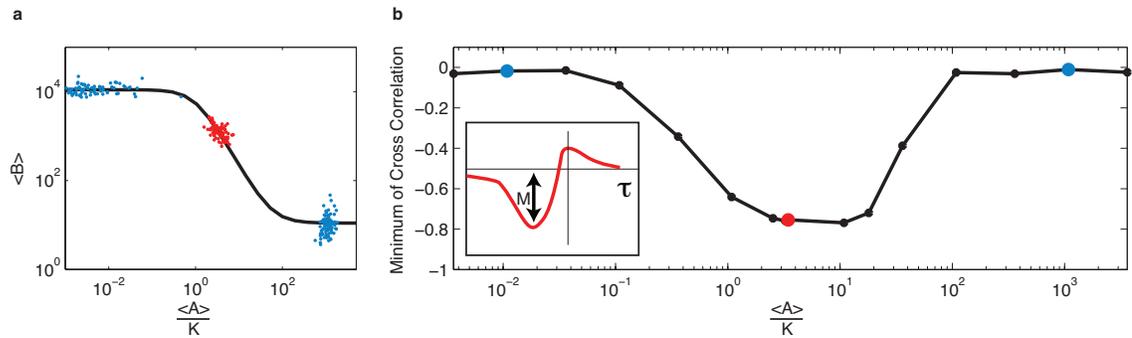


Figure 2.6: Cross correlation function shape depends on the activity of the network. (a) Ratio of the mean value of A (A_{eq}) to K versus the mean value of B . On average the shape of the curve is a Hill function. Several regions of this curve are explored in simulation with colored dots corresponding to data in (b). (b) Magnitude of the dip in the cross correlation function due to repression. As the mean value of A moves through regions on the Hill function with non-zero slope, the cross correlation function exhibits a characteristic dip due to repression. Inset shows schematic of dip magnitude.

The shape of the cross correlation function is highly dependent upon the activity of the regulatory link. We explore the dependence of its features on system parameters further in Section 2.5. To summarize: We found that τ_{reg} is most sensitive to the cell cycle time, with longer cell cycles producing a longer τ_{reg} . The magnitude of the dip due to regulation, in contrast, is determined primarily by the slope of the regulation function—how switch-like the regulation is.

Simulations can be used to predict the shape of cross correlation functions for other network architectures. A transcriptional activator is shown in Fig. 2.7 under different noise environments. Regulation appears as a peak in the cross correlation at $\tau < 0$. With only intrinsic noise, this peak

is a mirror image of the dip in the cross correlation function due to repression. With extrinsic noise, this peak is combined with the positive, symmetric peak that is caused by global noise sources.

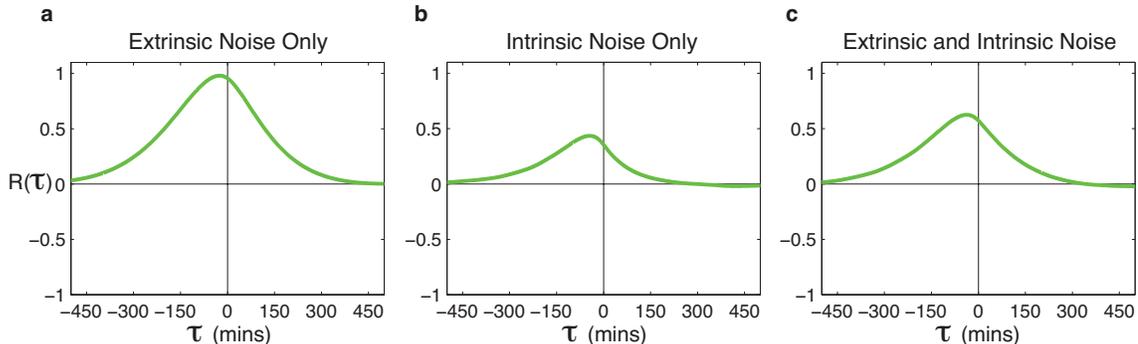


Figure 2.7: Cross correlation functions for an activator with (a) extrinsic noise only, (b) intrinsic noise only, and (c) both noise sources. Parameters are identical to those used in the prior repression simulations except $n = -1.7$.

2.4 Analytic Solutions for Cross Correlation Functions Due to Noise

In this section we develop an analytic method for calculating arbitrary cross correlation functions. This work was conducted in collaboration with Joe Levine.

We consider the stochastic differential equations in Eqns. (2.8)–(2.10) and (2.14)–(2.15). Assuming perturbations due to noise are small, we linearize the system about the equilibrium point given in Eqns. (2.5)–(2.7). Defining $a = A - A_{eq}$, $b = B - B_{eq}$, and $c = C - C_{eq}$ we obtain the following set of linear dynamics

$$\dot{a} = E + I_A - \beta a \quad (2.17)$$

$$\dot{b} = ga + E + I_B - \beta b \quad (2.18)$$

$$\dot{c} = E + I_C - \beta c \quad (2.19)$$

where g is the *local sensitivity*

$$g = -\frac{\alpha_B n \left(\frac{\alpha_A}{\beta K}\right)^{n-1}}{K \left(1 + \left(\frac{\alpha_A}{\beta K}\right)^n\right)^2}.$$

If we assume that the mean value of A is in the center of the Hill function so $A_{eq} = K$, as is the case with our simulated system, then this simplifies to $g = -\frac{\alpha_B n}{4K}$. This analysis is still possible regardless of the steady state value of A , but certain regimes (the middle of the Hill function, saturated edges of the nonlinearity) are better approximated by linear models. Note that the constant g is the only place that information about the nonlinearity enters the equations.

The cross correlation theorem states that cross correlation in the time domain is equal to multiplication in the frequency domain

$$R_{f,g}(\tau) = F^{-1}[\tilde{f}^* \tilde{g}],$$

where * denotes the complex conjugate and $f(t)$ and $g(t)$ are the two signals. However, we average over many cross correlation functions (as seen in Fig. 2.4), so we need to calculate the expected value of the cross correlation function over many realizations of the noise

$$E\{R_{f,g}(\tau)\} = E\{F^{-1}[\tilde{f}^* \tilde{g}]\}. \quad (2.20)$$

Taking the Fourier transform of Eqns. (2.17)–(2.19) we find

$$\tilde{a} = \frac{1}{\beta + i\omega}(\tilde{E} + \tilde{I}_A) \quad (2.21)$$

$$\tilde{b} = \frac{1}{\beta + i\omega}(\tilde{E} + \tilde{I}_B + g\tilde{a}) \quad (2.22)$$

$$\tilde{c} = \frac{1}{\beta + i\omega}(\tilde{E} + \tilde{I}_C) \quad (2.23)$$

$$\tilde{E} = \frac{\theta}{\beta + i\omega} \tilde{\eta}_E \quad (2.24)$$

$$\tilde{I}_i = \frac{\lambda_i}{\kappa + i\omega} \tilde{\eta}_i. \quad (2.25)$$

Below, we calculate cross correlation expressions for two cases: two independent genes (A and C) and a simple regulatory link with repression (A and B). The first, and simpler, case is worked through in detail, while the results of the second case are summarized.

2.4.1 Unregulated Case

We substitute Eqns. (2.21)–(2.25) into Eqn. (2.20), dropping tildes to simplify notation

$$\begin{aligned} E\{R_{a,c}(\tau)\} &= E\{F^{-1}\left[\frac{1}{\beta - i\omega}\left(\frac{\theta}{\beta - i\omega}\eta_E^* + \frac{\lambda_A}{\kappa - i\omega}\eta_a^*\right)\frac{1}{\beta + i\omega}\left(\frac{\theta}{\beta + i\omega}\eta_E + \frac{\lambda_C}{\kappa + i\omega}\eta_c\right)\right]\} \\ &= E\{F^{-1}\left[\frac{1}{\beta^2 + \omega^2}\left(\frac{\theta^2}{\beta^2 + \omega^2}\eta_E^*\eta_E + \frac{\theta\lambda_C}{(\beta - i\omega)(\kappa + i\omega)}\eta_E^*\eta_c\right.\right. \\ &\quad \left.\left. + \frac{\theta\lambda_A}{(\beta + i\omega)(\kappa - i\omega)}\eta_a^*\eta_E + \frac{\lambda_A\lambda_C}{\beta^2 + \omega^2}\eta_a^*\eta_c\right)\right]\}. \end{aligned}$$

Because the Fourier transform is a linear operation we can analyze each of the four terms individually.

We use two features of white noise to simplify analysis. First, white noise has a flat power spectral

density $\eta_i^*(\omega)\eta_i(\omega) = W_i$ and second $E\{\eta_i(t)\eta_j(t)\} = 0$ for $i \neq j$. Thus,

$$\begin{aligned} E\{\eta_i(t)\eta_j(t)\}_{i \neq j} &= E\{F^{-1}[F[\eta_i(t)\eta_j(t)]]\} \\ &= E\{F^{-1}[\eta_i^*(\omega)\eta_j(\omega)]\} \\ &= 0. \end{aligned}$$

Therefore if we have a deterministic function $G(\omega)$

$$\begin{aligned} E\{F^{-1}[G(\omega)\eta_i^*(\omega)\eta_j(\omega)]\}_{i \neq j} &= E\left\{\frac{1}{\sqrt{2\pi}}F^{-1}[G(\omega)] \star F^{-1}[\eta_i^*(\omega)\eta_j(\omega)]\right\} \\ &= \frac{1}{\sqrt{2\pi}}F^{-1}[G(\omega)] \star E\{F^{-1}[\eta_i^*(\omega)\eta_j(\omega)]\} \\ &= 0, \end{aligned}$$

where \star represents convolution. Due to these white noise properties the last three terms in the cross correlation expression become zero and the remaining term simplifies to

$$E\{R_{a,c}(\tau)\} = F^{-1}\left[\frac{1}{\beta^2 + \omega^2} \frac{\theta^2}{\beta^2 + \omega^2} W_E\right].$$

Applying the inverse Fourier transform we find

$$E\{R_{a,c}(\tau)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\beta^2 + \omega^2} \frac{\theta^2}{\beta^2 + \omega^2} W_E e^{-i\omega\tau} d\omega.$$

This integral can be solved using Cauchy's Residue theorem. Specifically, we need to consider two cases: $\tau < 0$ and $\tau \geq 0$. In the first case we can apply Jordan's lemma if we use a contour that encircles the upper half plane (Fig. 2.8a). Using Cauchy's Residue theorem we find

$$\lim_{R \rightarrow \infty} \int_{C_R} f(z) dz + \int_{-R}^R f(z) dz = 2\pi i \sum \text{Res.}$$

By Jordan's lemma the contour at infinity, C_R , becomes zero and we are left with the integral we want to evaluate. In our integral we have a second-order pole at $z = i\beta$ and a second-order pole at $z = -i\beta$. Since we are closing the contour in the upper half plane we need to evaluate the residue at $z = i\beta$:

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\beta^2 + \omega^2} \frac{\theta^2}{\beta^2 + \omega^2} W_E e^{-i\omega\tau} d\omega &= \frac{1}{2\pi} 2\pi i \text{Res}(i\beta) \\ &= \frac{\theta^2 W_E}{4\beta^3} e^{\beta\tau} (1 - \beta\tau). \end{aligned}$$

For $\tau \geq 0$ we can use Jordan's lemma if we choose a contour that encircles the lower half plane

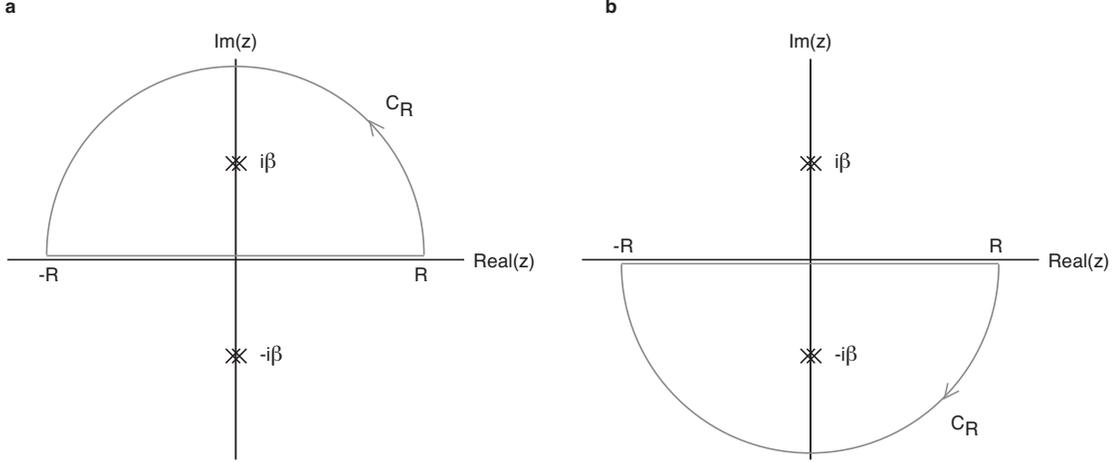


Figure 2.8: Contours used for evaluating Cauchy's Residue theorem. Pole locations shown are for the unregulated cross correlation expression.

(Fig. 2.8b). Since the direction of encirclement is now clockwise, the residue theorem has an additional negative sign:

$$\begin{aligned} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\beta^2 + \omega^2} \frac{\theta^2}{\beta^2 + \omega^2} W_E e^{-i\omega\tau} d\omega &= \frac{1}{2\pi} (-2\pi i \text{Res}(-i\beta)) \\ &= \frac{\theta^2 W_E}{4\beta^3} e^{-\beta\tau} (1 + \beta\tau). \end{aligned}$$

Combining these two results we find

$$E\{R_{a,c}(\tau)\} = \frac{\theta^2 W_E}{4\beta^3} e^{-\beta|\tau|} (1 + \beta|\tau|).$$

Note that this expression for the cross correlation used a and c , the mean subtracted versions of A and C . These expressions are consistent with those calculated directly from the nonlinear simulations because we used the mean subtracted version of the cross correlation function (Eqn. (2.16)).

2.4.2 Regulated Case

Analysis of the cross correlation function due to repression is similar. After simplification using the white noise properties discussed in the previous section we find

$$\begin{aligned} E\{R_{a,b}(\tau)\} &= F^{-1}\left[\frac{\theta^2 W_E}{(\beta + i\omega)^2(\beta - i\omega)^2}\right] \\ &+ F^{-1}\left[\frac{g \theta^2 W_E}{(\beta + i\omega)^3(\beta - i\omega)^2}\right] \\ &+ F^{-1}\left[\frac{g \lambda_A^2 W_A}{(\beta + i\omega)^2(\beta - i\omega)(\kappa + i\omega)(\kappa - i\omega)}\right]. \end{aligned}$$

These three terms have a convenient interpretation. The first term is the artificial correlation due to extrinsic noise, the second term is extrinsic noise that has propagated through the link, the third term is intrinsic noise that has propagated through the link. Cauchy's Residue theorem and Jordan's lemma are applied to find

$$R_{a,b}(\tau) = \begin{cases} \frac{\theta^2 W_E}{4\beta^3} e^{\beta\tau} (1 - \beta\tau) \\ + \frac{g\theta^2 W_E}{16\beta^4} e^{\beta\tau} (3 - 4\beta\tau + 2\beta^2\tau^2) \\ + \lambda_A^2 g W_A \left(e^{\beta\tau} \frac{\kappa^2(1-2\beta\tau) - \beta^2(5-2\beta\tau)}{4\beta^2(\beta^2 - \kappa^2)^2} + e^{\kappa\tau} \frac{1}{2\kappa(\beta - \kappa)^2(\beta + \kappa)} \right) & \tau < 0 \\ \frac{\theta^2 W_E}{4\beta^3} e^{-\beta\tau} (1 + \beta\tau) \\ + \frac{g\theta^2 W_E}{16\beta^4} e^{-\beta\tau} (3 + 2\beta\tau) \\ + \lambda_A^2 g W_A \left(\frac{e^{-\beta\tau}}{4\beta^2(\kappa^2 - \beta^2)} + \frac{e^{-\kappa\tau}}{2\kappa(\beta + \kappa)^2(\beta - \kappa)} \right) & \tau \geq 0. \end{cases}$$

2.4.3 Summary

The cross correlation relations are summarized as

$$E\{R_{A,C}(\tau)\} = N_{A,C} \frac{\theta^2 W_E}{4\beta^3} e^{-\beta|\tau|} (1 + \beta|\tau|)$$

$$E\{R_{A,B}(\tau)\} = \begin{cases} N_{A,B} \left(\frac{\theta^2 W_E}{16\beta^4} e^{\beta\tau} (2g\beta^2\tau^2 - 4\beta(g + \beta)\tau + 3g + 4\beta) \right. \\ \left. + \lambda_A^2 g W_A \left(e^{\beta\tau} \frac{\kappa^2(1-2\beta\tau) - \beta^2(5-2\beta\tau)}{4\beta^2(\kappa^2 - \beta^2)^2} + e^{\kappa\tau} \frac{1}{2\kappa(\beta - \kappa)^2(\beta + \kappa)} \right) \right) & \tau < 0 \\ N_{A,B} \left(\frac{\theta^2 W_E}{16\beta^4} e^{-\beta\tau} (2\beta(g + 2\beta)\tau + 3g + 4\beta) \right. \\ \left. + \lambda_A^2 g W_A \left(e^{-\beta\tau} \frac{1}{4\beta^2(\kappa^2 - \beta^2)} + e^{-\kappa\tau} \frac{1}{2\kappa(\beta + \kappa)^2(\beta - \kappa)} \right) \right) & \tau \geq 0 \end{cases}$$

where the normalization factors are

$$N_{A,B} = \frac{1}{\sqrt{R_{A,A}(0)R_{B,B}(0)}}$$

$$N_{A,C} = \frac{1}{\sqrt{R_{A,A}(0)R_{C,C}(0)}}$$

$$R_{A,A}(0) = \frac{\theta^2 W_E}{4\beta^3} + \frac{\lambda_A^2 W_A}{2\beta\kappa(\kappa + \beta)}$$

$$R_{B,B}(0) = \frac{\theta^2(3g^2 + 6g\beta + 4\beta^2) W_E}{16\beta^5} + \frac{\lambda_A^2 g^2(\kappa + 2\beta) W_A}{4\kappa\beta^3(\kappa + \beta)^2} + \frac{\lambda_B^2 W_B}{2\kappa\beta(\kappa + \beta)}$$

$$R_{C,C}(0) = \frac{\theta^2 W_E}{4\beta^3} + \frac{\lambda_C^2 W_C}{2\beta\kappa(\kappa + \beta)}.$$

To compare these results to the nonlinear simulation we use the constants specified in Table 2.1, where $g = -\frac{\alpha_B n}{4K}$. W_E, W_i for $i = \{A, B, C\}$ are treated as binary variables and set to 0 or 1 to turn off and on extrinsic and intrinsic noise for comparison to Fig. 2.5. We assume $W_A = W_B = W_C$

for simplicity. Fig. 2.9 shows that the analytic solutions for cross correlations match the simulated nonlinear system extremely well.

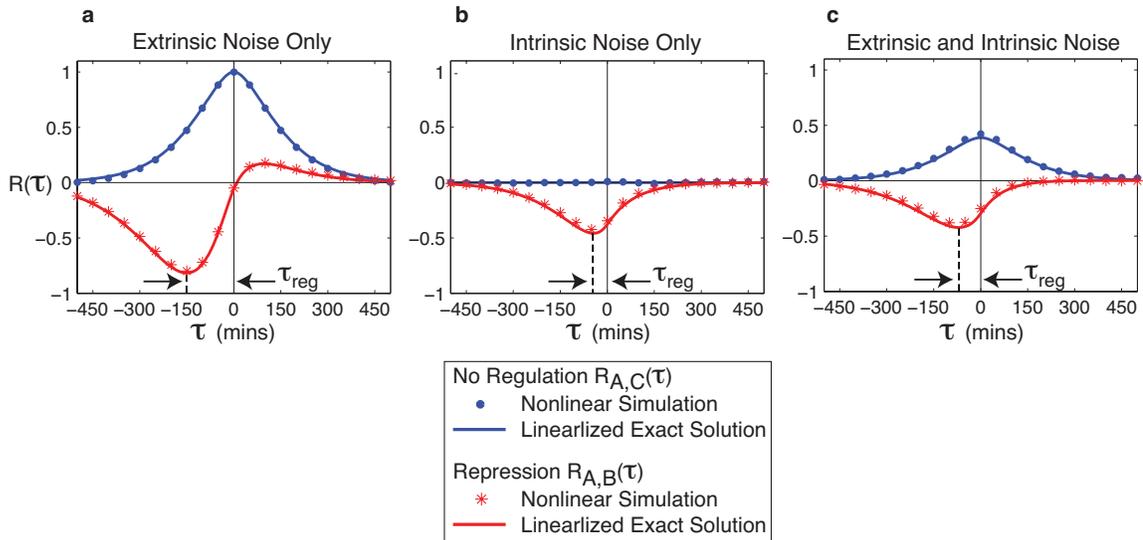


Figure 2.9: Comparison of analytically calculated cross correlation expressions and simulated data. Dots represent simulated data, while solid lines plot analytic solutions for the linearized model.

We have calculated the cross correlation functions for two types of regulation. This method can be applied more generally to larger networks, provided perturbations due to noise are small enough that linearization is a valid approximation.

2.5 Sensitivity Analysis

Two prominent features of the cross correlation curve for repression are the location of the dip, τ_{reg} , and the magnitude of the dip, M (shown schematically in Fig. 2.10). We calculate how sensitive these features are to variations in the system parameters. These results indicate which parameters play a primary role in setting the features of the cross correlation function.

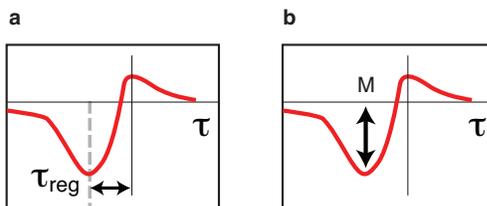


Figure 2.10: Schematic of cross correlation function features (a) τ_{reg} and (b) M .

To find τ_{reg} for repression we take

$$\left. \frac{dR_{a,b}(\tau)}{d\tau} \right|_{\tau=\tau_{reg}} = 0.$$

In general it is not possible to find a closed form solution for τ_{reg} , however numerical root finding methods can be applied. The nominal parameter values for sensitivity analysis are the same as those used in the simulation. For a feature of the cross correlation function, y , we find the normalized sensitivity

$$S_i = \frac{y(p_i + \Delta p_i) - y(p_i - \Delta p_i)}{2 \Delta p_i y(p_i)}.$$

For $\Delta p_i = 0.05 p_i$ the parameteric sensitivities are shown in Table 2.2. Large values indicate that the feature y is very sensitive to that parameter. The sign of the sensitivity indicates whether y will get larger ($S_i > 0$) or smaller ($S_i < 0$) as the parameter is increased.

Parameter	τ_{reg}	M
β	-108.5	2.39
g	18.9	-85.7
θ	11.3	5.5
κ	4.0	-12.4
λ_A	-0.9	2.5
λ_B	0.0	0.0
λ_C	0.0	0.0

Table 2.2: Normalized Sensitivities

τ_{reg} is most sensitive to the parameter β , which sets the time scale of both protein decay and extrinsic noise. As the cell cycle ($\log(2)/\beta$) gets longer, the location of the dip moves further away from zero. M is most sensitive to the local sensitivity g , which is negative for repression. As g becomes less negative, the repressor has less of an effect on its target and the dip gets smaller. In the extreme case when $g = 0$, which indicates an inactive or non-existent regulatory connection, the dip disappears.

2.6 Transcriptional Cascades

It is interesting to ask what the limits are to using these cross correlation functions. For two genetic components that are very far away, will the correlation eventually average to zero or will measurement be the limiting factor (it may be prohibitive to measure noisy signals in single cells for long periods of time)? To explore this question we derived an expression for the cross correlation function of a cascade of arbitrary length. The signals between the first and last element in the cascade are compared. This type of question was asked in [39] where they built a synthetic cascade to measure noise, and in [33] with a numerical study.



Figure 2.11: Cascade of length n . Interactions alternate between repressors (T-arrows) and activators (normal arrows) to preserve the net repression effect.

The cascade shown in Fig. 2.11 is described by a system of linear equations:

$$\begin{aligned} \dot{x}_1 &= E + I_1 - \beta x_1 \\ \dot{x}_2 &= E + I_2 + g_1 x_1 - \beta x_2 \\ \dot{x}_3 &= E + I_3 + g_2 x_2 - \beta x_3 \\ &\dots \\ \dot{x}_n &= E + I_n + g_{n-1} x_{n-1} - \beta x_n. \end{aligned}$$

The Fourier transforms used in the cross correlation calculation are

$$\begin{aligned} \tilde{x}_1 &= \frac{1}{\beta + i\omega} (\tilde{E} + \tilde{I}_1) \\ \tilde{x}_n &= \sum_{k=1}^n \frac{1}{(\beta + i\omega)^{n-k+1}} \left(\prod_{j=k}^n g_j \right) (\tilde{E} + \tilde{I}_k) \end{aligned}$$

where $g_n = 1$. The full cross correlation expressions were calculated using an automated Mathematica script, given in Appendix A. In general, the `Residue` command in Mathematica is an efficient way to calculate cross correlation functions.

Cross correlations from cascades with between two and eight elements are compared in Fig. 2.12. Even for long chains of transcription factors we still see non-negligible correlation values. Thus, it is likely that experimental limitations—such as the length of a movie—will restrict ability to see distant temporal events. In addition, real genetic networks will have other inputs to intermediate elements that may confound analysis.

2.7 mRNA dynamics

To this point we have only considered protein dynamics and have not considered regulation at the mRNA level. Using analytic methods we can show that mRNA dynamics, and other fast time-scale processes, can be neglected to a first approximation.

The systems we have considered so far are of the form

$$\dot{p}_B = -\beta p_B + g_p p_A + n,$$

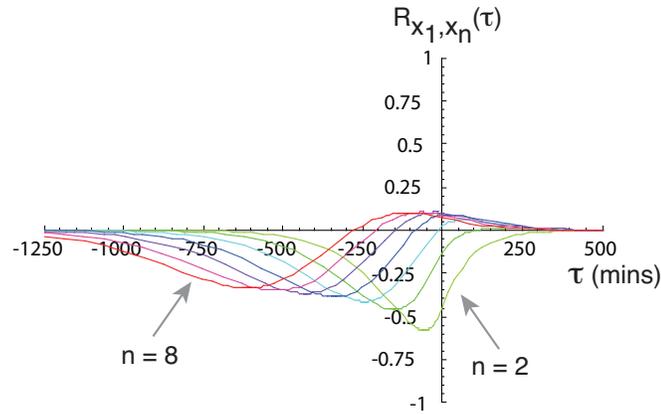


Figure 2.12: Cross correlation functions for different length cascades. Cross correlations are between the signals x_1 and x_n for $n = 2, 3, \dots, 8$. $g = 0.02$ where $g_n = 1$ and $g_i = -g$ for odd i and $g_i = g$ for even i .

where p_A and p_B are the input and output protein concentrations, respectively, g_p is the local sensitivity, and n is a grouped noise term. The Fourier transform of p_B is

$$\tilde{p}_B = \frac{1}{\beta + i\omega} (g_p \tilde{p}_A + \tilde{n}).$$

Extending this model to include mRNA dynamics (as in [23], using a linearized form of the equations), we have

$$\begin{aligned} \dot{m}_B &= -\beta_m m_B + g_m p_A + n_m \\ \dot{p}_B &= -\beta(p_B - m_B) + n_p, \end{aligned}$$

where m_B is the mRNA concentration and n_m and n_p are noise in the mRNA and protein production processes. Taking the Fourier transforms we find

$$\tilde{p}_B = \frac{1}{\beta + i\omega} \left(\frac{\beta g_m}{\beta_m + i\omega} \tilde{p}_A + \tilde{n} \right),$$

where we have grouped all the noise terms into n .

Comparing these two equations, as long as

$$g_p \approx \frac{\beta g_m}{\beta_m + i\omega}$$

we can ignore the mRNA dynamics. If both models have the same steady-state behavior ($i\omega = 0$) then $g_p = \beta g_m / \beta_m$. The protein lifetime is $\beta = 0.0116$ 1/min, based on a 60 minute cell cycle time. mRNA lifetimes, in contrast, are on the order of 2 minutes, or $\beta_m = 0.3466$ 1/min [40]. Thus, the protein dynamics are 30 times slower than the mRNA dynamics and are thus expected to be

dominant in determining the system dynamics. As long as the mRNA dynamics are significantly faster than the protein dynamics it is reasonable to approximate the system using only the protein model.

2.8 Degenerate Cross Correlation Functions

There is not a unique relationship between the shape of a cross correlation function and the network architecture that generated it. Here, we explore two classes of networks that result in degenerate cross correlation functions. In these situations, two cross correlations look similar or identical even though they are the result of different network architectures.

2.8.1 Redundant Network Elements

Consider the network diagram shown in Fig. 2.13, assuming that network connection and expression properties are identical for the links A - B and A - C . Proteins B and C are redundant network elements because they are controlled by the same input, A , and extrinsic noise affects them in the same way. Thus, in the extreme case where only extrinsic noise is present, $B(t)$ and $C(t)$ will be identical. Consequently, the cross correlation function $R_{B,D}(\tau) = R_{C,D}(\tau)$.

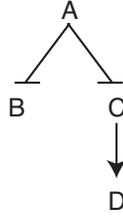


Figure 2.13: Network with redundant components

Intrinsic noise helps to discriminate between $R_{B,D}(\tau)$ and $R_{C,D}(\tau)$ because intrinsic noise in C propagates to D causing additional time-lagged correlation. Intrinsic noise in B , because it is uncorrelated with noise in D , will not have the same effect.

Analytic solutions for both cross correlation functions are summarized by

$$\begin{aligned}
 R_{B,D}(\tau) &= E\{F^{-1}[f_{Be}^* f_{De} + f_{BA}^* f_{DA}]\} \\
 R_{C,D}(\tau) &= E\{F^{-1}[f_{Ce}^* f_{De} + f_{CA}^* f_{DA} + f_{CC}^* f_{DC}]\},
 \end{aligned}$$

where $f_{i,j}$ is the Fourier transform of the differential equations describing the dynamics of protein j in response to noise from i , and e is extrinsic noise. Assuming the network connections are identical

for A - B and A - C we find the difference between the two cross correlation functions

$$R_{C,D}(\tau) - R_{B,D}(\tau) = E\{F^{-1}[f_{CC}^* f_{DC}]\}.$$

This term represents how noise in C propagates to D , but is only non-zero if there is intrinsic noise (or other inputs) affecting C .

Thus, mathematically $R_{C,D}(\tau) \neq R_{B,D}(\tau)$ unless there is no intrinsic noise in C . In practice it may be difficult to distinguish between the two cross correlation functions even if they are not identical.

This simple example illustrates how redundant network elements can confound analysis. More complicated networks will have similar problems any time there are two or more network elements that are controlled by the same inputs. Intrinsic noise or other signals that affect individual genes can help to distinguish correlations due to regulation from correlations due to redundant network elements.

2.8.2 Parametric Degeneracies

A second class of degeneracies comes from uncertainty in parameters. For example, Fig. 2.14 shows two cascades. If we measure $R_{A,C}(\tau)$ is it possible to determine that there is a middle element in the network or will the cross correlation function look like $R_{X,Y}(\tau)$? In this example extrinsic noise

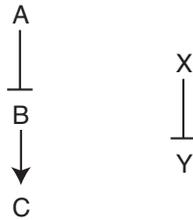


Figure 2.14: Example of a network showing parametric degeneracy

is helpful in discriminating between the two cascades. In the two-step cascade, extrinsic noise affects A , B , and C , and thus enters into the cross correlation function in three ways, while in the one-step cascade it only enters twice.

If we consider the limiting case where only intrinsic noise is present there are still ways adjust parameters to make the two cross correlation functions look nearly identical. For example, if B degrades quickly, but the net strength of the network is the same as in X - Y it is possible to choose parameters that result in very similar cross correlation functions (Fig. 2.15).

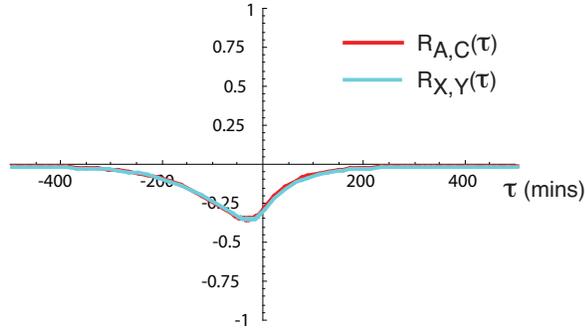


Figure 2.15: Cross correlation functions for two cascades with degenerate parameters. Simulation parameters are $\beta_x = \beta_y = \beta_a = \beta_c = 1/60 \text{ mins}^{-1}$, $\beta_b = 1/5 \text{ mins}^{-1}$, $g_{xy} = g_{ab}g_{bc} = -0.01$, $W_i = 1$.

In this example the intrinsic noise-only dynamics are described by

$$\begin{aligned}\dot{a} &= -\beta_a a + \eta_a \\ \dot{b} &= -\beta_b b + \eta_b + g_{ab} a \\ \dot{c} &= -\beta_c c + \eta_c + g_{bc} b \\ \\ \dot{x} &= -\beta_x x + \eta_x \\ \dot{y} &= -\beta_y y + \eta_y + g_{xy} x\end{aligned}$$

where intrinsic noise has been approximated by white noise and parameters are given in the figure caption.

Since biochemical parameters are often unknown or uncertain there are many possible situations where the shape of two cross correlation functions may look very similar.

2.8.3 Implications for Network Identification

Even though the cross correlation function does not uniquely determine the network architecture and parameters that generate it, it is still a useful tool. In particular, the cross correlation function can be used as a sensitive measure of network activity or it can suggest possible links in networks that are only partially mapped. Indeed, much of our knowledge about biochemical regulation comes from correlation-based reasoning and including temporal measurements extends these tools.

Chapter 3

Correlated Noise in a Synthetic Gene Circuit Reveals Regulation

Simulations and analysis in the previous chapter are based on parameters for the λ cI system. Here, we test those predictions experimentally by building a synthetic circuit using the λ cI repressor and other characterized circuit elements in *E. coli* bacteria. The synthetic circuit uses three fluorescent proteins so that noisy gene expression can be monitored with time-lapse microscopy. To apply the analysis methods in the previous chapter, the cross correlation calculation is extended to allow for branched data where all progeny originate from a single ancestor. Experimental cross correlation curves show features that were predicted in simulation and results are tested further by comparing two different noise backgrounds to study the relative effect of extrinsic and intrinsic noise.

3.1 Synthetic Construct

We built a synthetic gene circuit with one transcription factor and three fluorescent proteins (Fig. 3.1). A protein fusion of the transcription factor λ cI and yellow fluorescent protein (YFP) represses production of red fluorescent protein (RFP), which is under the control of a variant of the λP_R promoter, O_{R2}^* [41]. Cyan fluorescent protein (CFP) is controlled by a strong independent constitutive promoter, based on the σ_{70} consensus sequence. The construct was transformed into *E. coli* strain MG1655Z1, a derivative of MG1655 that overexpresses LacI [42]. The promoter for cI-YFP is controlled by LacI, which can be inactivated by the inducer IPTG. Thus, inducer concentration can be adjusted to place the mean cI-YFP concentration in an active regulatory range.

Two versions of this construct were built: (1) A chromosomally integrated version (single copy) and (2) a low-copy plasmid version (~ 10 copies). Fig. 3.1 shows the plasmid map; integration details are discussed in Section 3.6.2.

The design of this circuit is based upon work by R. Sidney Cox on a similar circuit, which is studied in Section 3.5. The fluorescent proteins are oriented to minimize read-through from one gene

to the next. In addition, multiple terminators are placed in between genes to stop transcriptional read-through. Each promoter in the circuit is bracketed by unique restriction sites to allow for easy interchange of promoters. A kanamycin antibiotic resistance marker is included to select for bacteria that contain the synthetic construct. In the chromosomally integrated version, the region from the kanamycin resistance marker through the end of *cI*-YFP was integrated into the *galK* site of MG1655Z1. The plasmid version uses a low-copy origin of replication, SC101 [42].

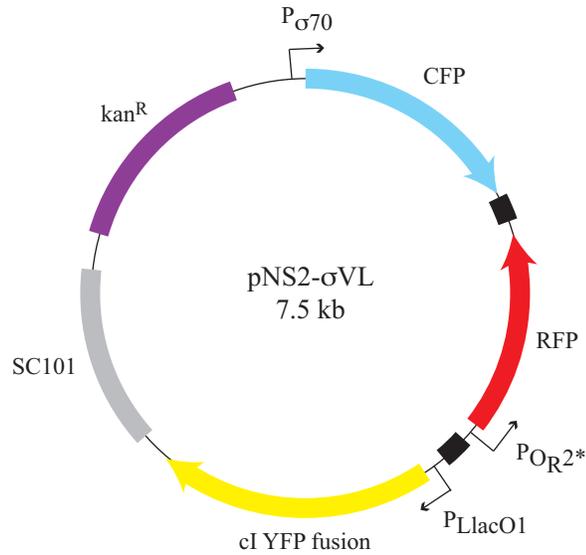


Figure 3.1: Plasmid version of the synthetic gene circuit. Colored arrows are genes, small black arrows are promoters. Black boxes indicate several terminators placed next to each other. The plasmid is a 7,500 base pair loop of DNA that replicates independently inside the cell.

A schematic representation of how noise affects the circuit is shown in Fig. 3.2. Each individual gene had intrinsic noise associated with it, while extrinsic noise affects all genes.

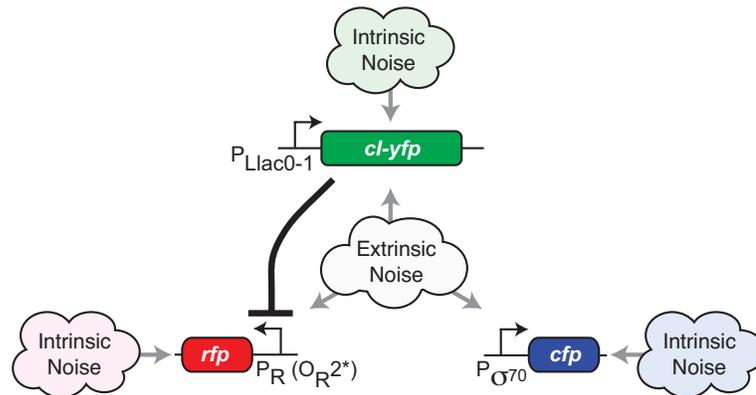


Figure 3.2: Noise sources affecting the three-color circuit.

3.2 Static Data

To quantify the behavior of the gene circuit in response to the inducer IPTG we took snapshots of individual cells and measured expression of the three fluorescent proteins (Fig. 3.3). As expected, CFP is relatively independent of IPTG, YFP increases as LacI is inhibited, and RFP decreases as cI-YFP increases. In the movies discussed below we used between 10–15 μM IPTG concentrations to ensure that cI-YFP was actively regulating RFP.

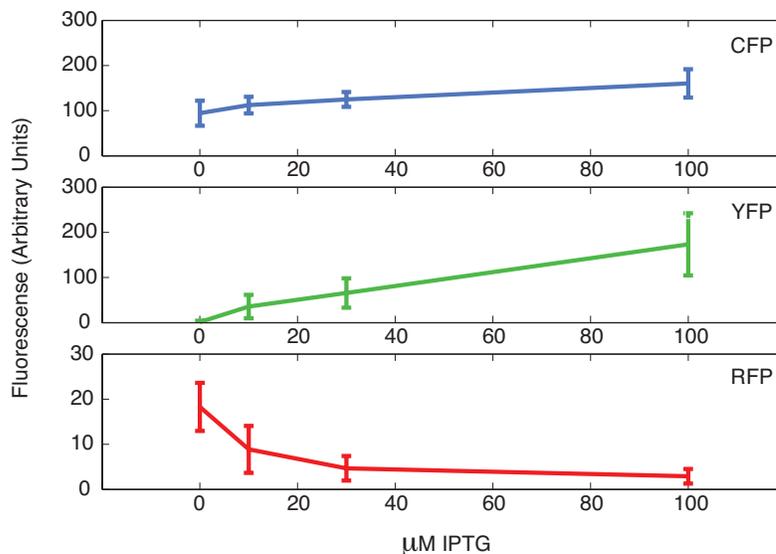


Figure 3.3: Dose response curves for three-color circuit with IPTG. Data comes from single cells (100–200 per data point), measured at 0, 10, 30, 100 μM IPTG. Error bars are the standard deviation of the data. Background has been subtracted from all fluorescence data.

We plotted snapshot data to show raw correlations between genes before considering temporal effects. Fig. 3.4 shows data from the plasmid and the chromosomally integrated versions of the construct; each dot is data from a single cell. Comparing the unregulated genes (CFP and YFP) with the regulated genes (RFP and YFP), a negative correlation can be seen due to repression of RFP by cI-YFP. The pair of unregulated genes are positively correlated, even in the absence of any regulatory elements linking them together. This correlation is the product of extrinsic noise in the system. Moving to the plasmid case, correlation between all pairs of genes increases. Furthermore, RFP and YFP on the plasmid show similar levels of correlation to CFP and YFP on the chromosome. It is clear that from these data alone it is not possible to infer the presence of regulation or details about its strength and the time scale on which it occurs.

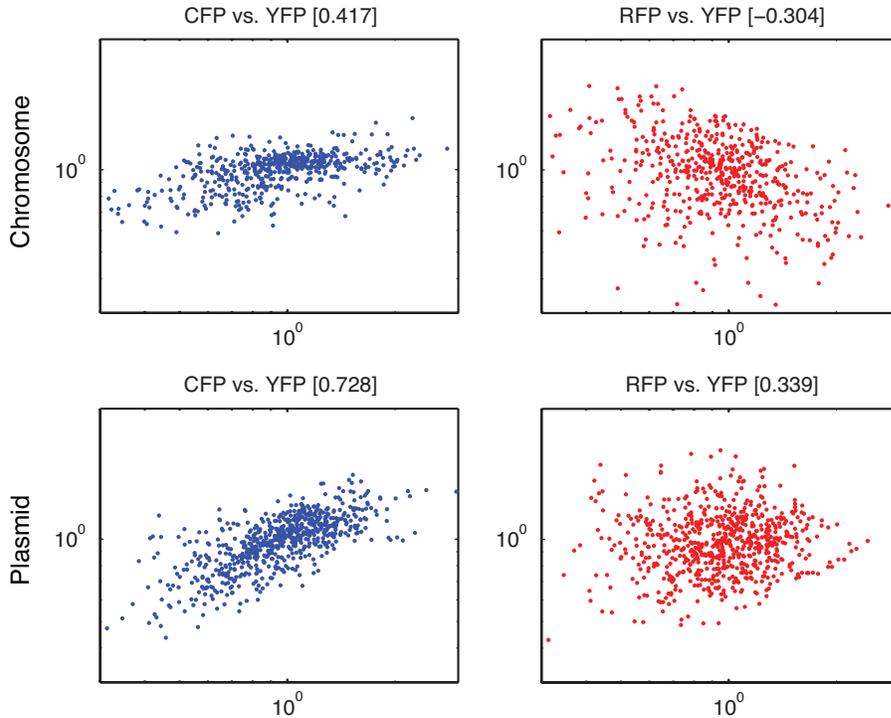


Figure 3.4: Expression of RFP, YFP, and CFP in individual cells plotted against each other. Title bars contain the correlation coefficient for these data. Plots are shown in log scale. Fluorescence data are background subtracted, units are arbitrary.

3.3 Temporal Data

Cells were grown and imaged in three colors using automated time-lapse fluorescence microscopy. A filmstrip from one movie of the chromosomally integrated circuit is shown in Fig. 3.5a. Although cells are imaged in all three colors at once, the two filmstrips show YFP & RFP and YFP & CFP separately for clarity. Note that pixels that contain equal levels of red and green appear as yellow, while equal levels of blue and green appear cyan. Individual colors are shown in grayscale in the right panel of Fig. 3.5a for a single time point. These images show strong anti-correlation between RFP (red) and YFP (green), while CFP (blue) is expressed at a more homogeneous level across all cells (see right panel of Fig. 3.5a). The expression of CFP does not appear to be linked to the expression of YFP or CFP, consistent with the design of circuit. The appearance of spatially grouped sub-populations of cells that display similar fluorescence states occurs because τ_{reg} exceeds the cell cycle time, consistent with simulation predictions (Fig. 2.5) and observations in [36].

A filmstrip of the alternate plasmid-based construct is shown in Fig. 3.5b. We expect an increase in extrinsic noise due to plasmid copy number fluctuations. Thus, the correlations between all three colors should be higher than in the chromosomal case. Although anti-correlation between RFP (red) and YFP (green) is visible, there is a marked increase in the number of yellow pixels, indicating

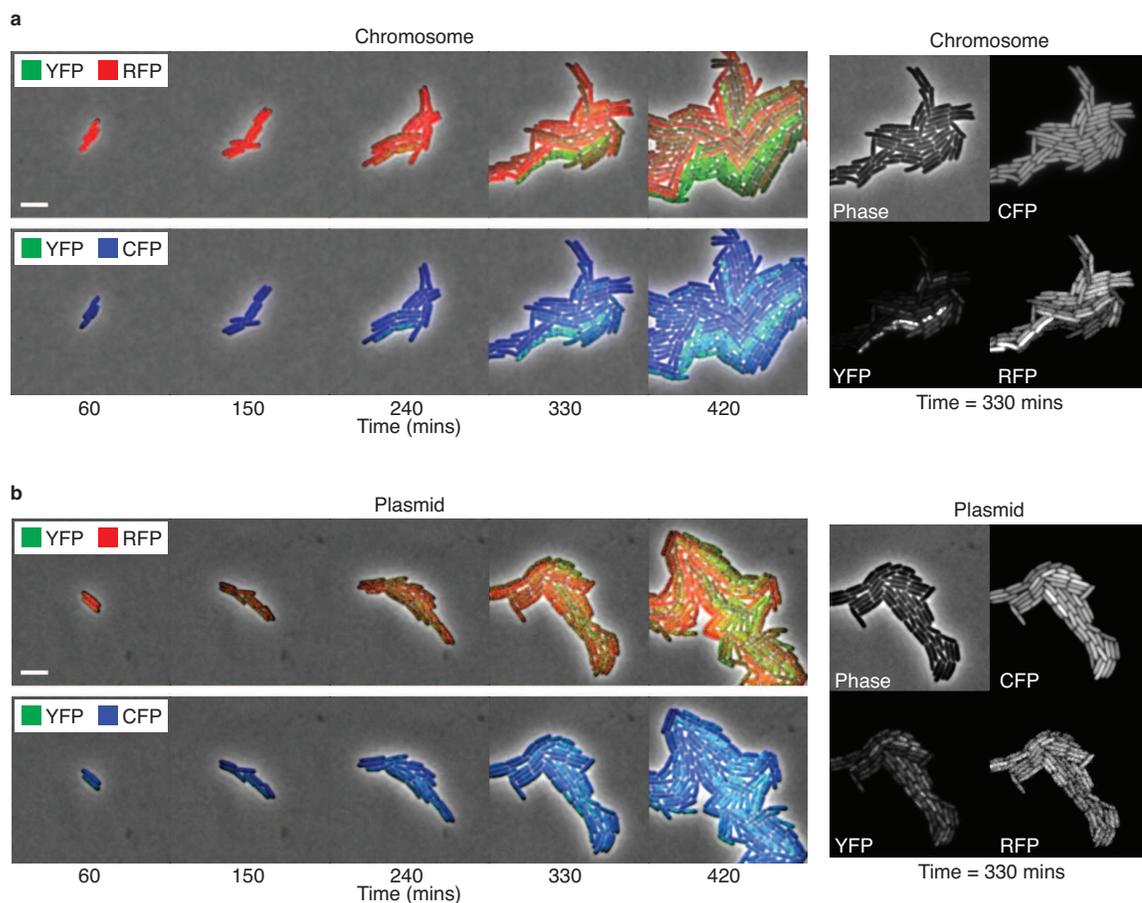


Figure 3.5: Time-lapse movies of gene expression fluctuations in a synthetic genetic circuit. (a) Left panels show filmstrip of cells with chromosomally integrated circuit. Note strong anti-correlation between RFP (red) and YFP (green), and the lower correlation between CFP (blue) and YFP. Scale bar, $5 \mu\text{m}$. Right panel shows individual colors and phase images for $t = 330$ mins. Note anti-correlation between RFP and YFP and the uniform expression of CFP. (b) Filmstrip of the same circuit on a low copy plasmid shows increased variability in all colors. In particular, in the right panel note increased variability of CFP relative to the chromosomal case. Colors and scale bar are the same as in (a). Extended movies are shown in Appendix B.

increased correlation between these two genes. Comparing the black and white panels for CFP in Fig. 3.5a and b, the plasmid shows increased variability in this constitutive promoter over the chromosomal case. The right panel also shows the effects of extrinsic noise, where cells that are bright in one color tend to be bright in all three colors.

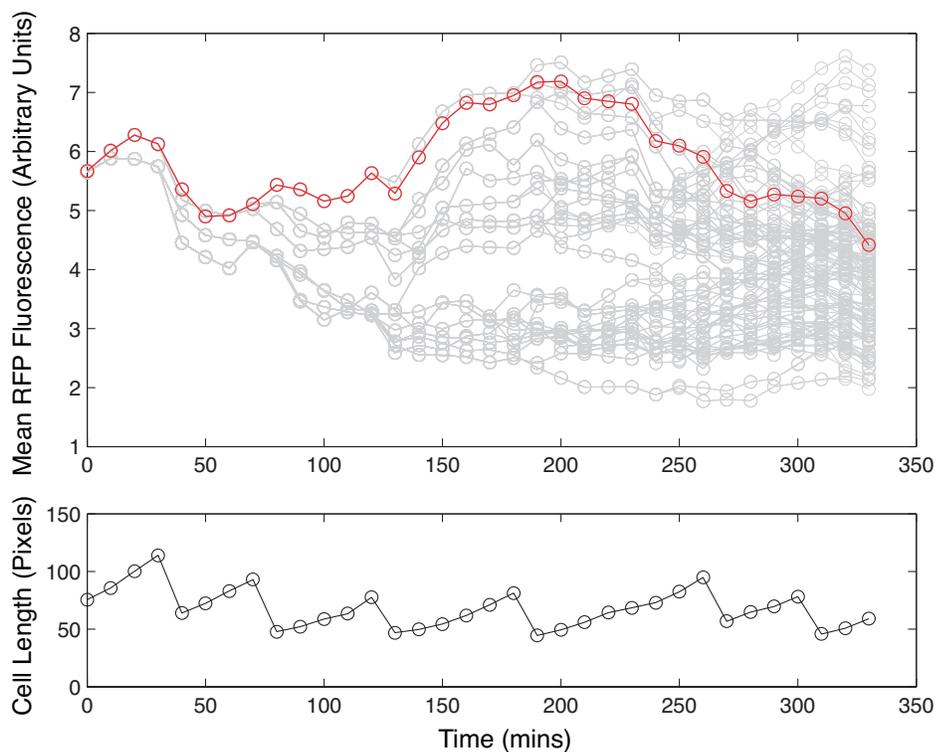


Figure 3.6: Mean (averaged over the cell) expression of RFP for all cells in a growing microcolony. A single representative trace is highlighted in red. Data on cell length for this representative cell lineage are shown in the bottom panel. Decreases in the cell length are cell division events. Note that these decreases are not visible in the RFP data because it has been averaged by cell volume.

To analyze these data quantitatively, we used semi-automated image analysis software to extract fluorescence intensities for individual cells across the lineage tree of a growing microcolony (Fig. 3.6). These data show a wide diversity of expression levels at the end of the movie. A single trace is highlighted in red and the corresponding cell length is shown in the bottom panel—variations in gene expression are not strongly linked to the cell cycle. Typical time traces from a single cell lineage for all three fluorescent genes are shown in Fig. 3.7. More variability is seen in the expression of plasmid-based genes than in their chromosome-based counterparts, consistent with the enhanced level of extrinsic noise. The same data shifted in time reveal temporal anti-correlation between *cI*-YFP and RFP signals (Fig. 3.8).

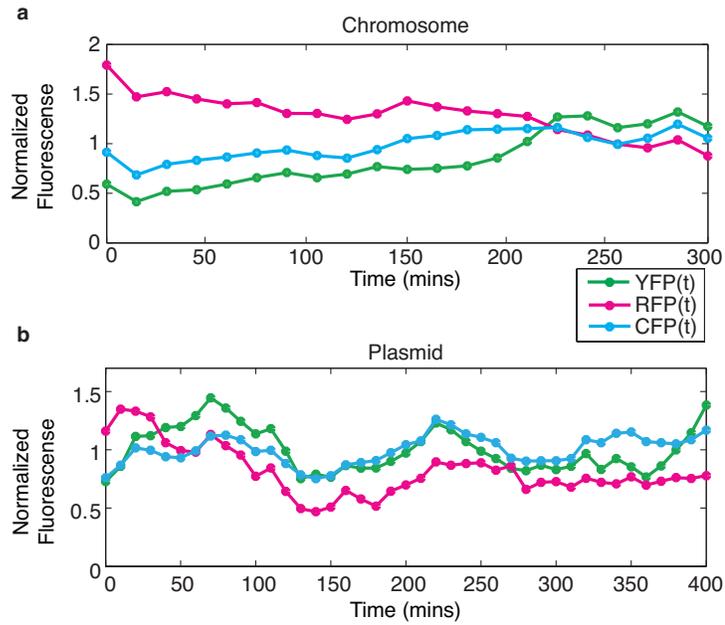


Figure 3.7: Typical lineage traces show noise in gene expression. Data are normalized by mean intensity.

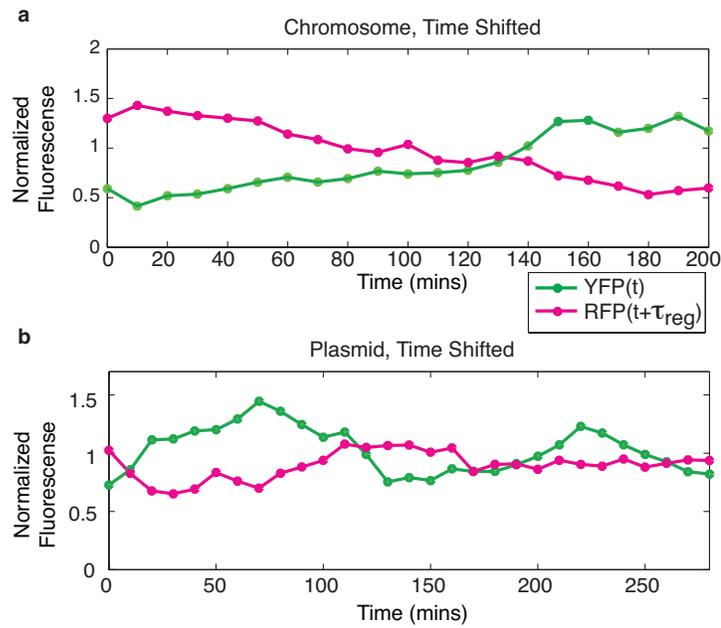


Figure 3.8: Data are the same as in Fig. 3.7, however RFP data are shifted to reveal the delayed anti-correlation. (chromosome: $\tau_{reg} = 100$ mins, plasmid: $\tau_{reg} = 120$ mins)

3.4 Experimental Cross Correlations

3.4.1 Extension of the Cross Correlation Function to Branched Data

The cross correlation function was calculated using these temporal data. However, it was necessary to introduce a modified formula for the cross correlation function to properly account for the branching nature of the data. Fig. 3.9 illustrates the need for an extension to the traditional cross correlation formula. In this example we start with one cell, which divides into two cells, and finally four cells. Although we have four time-series measurements, some of the data between them are shared. Thus, data from the trunk of the tree are counted four times, while data at the leaves are only counted once. This is potentially problematic if the characteristics of the correlation at early times are different from those at the end; the correlation expression will be heavily biased towards the early data. To correct for the problem of over-counting data we introduced a modified expression for the

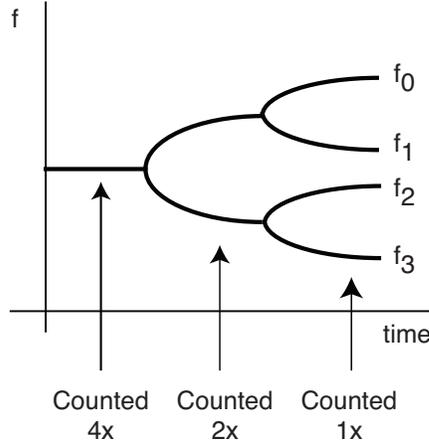


Figure 3.9: Cross correlations on data with a tree structure need to avoid over-counting.

cross correlation function, $R_{f,g}(\tau)$, given by

$$S_{f,g}(\tau) = \begin{cases} \frac{1}{N-|\tau|} \frac{1}{N_{cells}} \left[\sum_{i=0}^{N_{cells}-1} \left(\sum_{n=0}^{N-\tau-1} \tilde{f}_i(n+\tau) \tilde{g}_i(n) \right) - \sum_{i=0}^{N_{cells}-2} \left(\sum_{n=0}^{k_i-\tau-1} \tilde{f}_i(n+\tau) \tilde{g}_i(n) \right) \right] & \tau \geq 0 \\ S_{g,f}(-\tau) & \tau < 0 \end{cases} \quad (3.1)$$

$$R_{f,g} = \frac{S_{f,g}(\tau)}{\sqrt{S_{f,f}(0)S_{g,g}(0)}}.$$

This expression has two terms: The first term is the standard expression for the cross correlation function, averaged over the number of data traces (N_{cells}). The second term accounts for any over-counting of data by subtracting data that have been counted more than once.

N_{cells} is the total number of cells at the final time point (four in Fig. 3.9) and k_i is the branching

point between traces f_i and f_{i+1} , described in further detail below. Other terms are the same as defined in the original cross correlation expression given in Eqn. (2.16) except the mean subtraction is extended to the branched situation using

$$\tilde{f}_i = f_i - \sum_{i=0}^{N_{cells}-1} f_i.$$

Fig. 3.10 further illustrates the data analysis process. The formula in Eqn. (3.1) assumes that the numbering of the data traces is ordered in a manner so that, for all data, each pair of adjacent data traces f_i and f_{i+1} share as much data as possible. Fig. 3.10a illustrates numbering before correction. A preprocessing step must be applied to renumber these data traces, as in Fig. 3.10b. The branching points between data are listed on the x-axis.

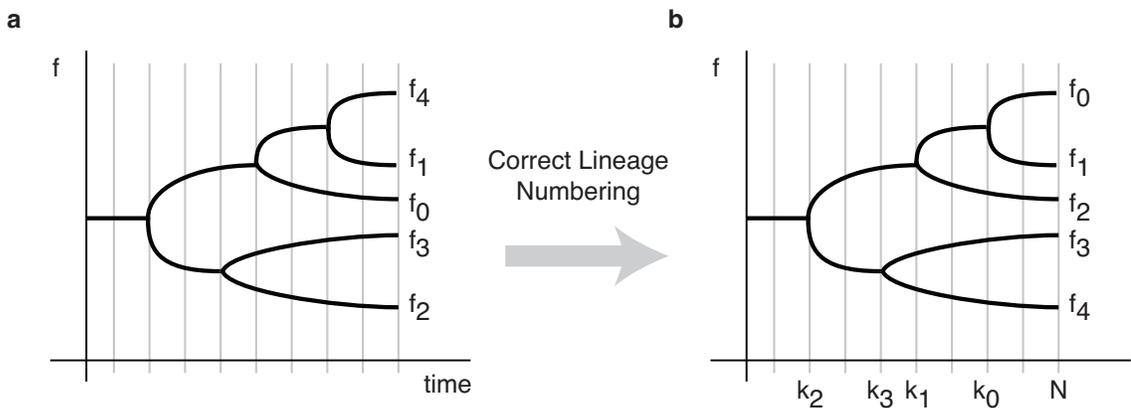


Figure 3.10: Tree data must be preprocessed for optimal numbering. Here $N_{cells} = 5$, $N = 9$, $k_0 = 7$, $k_1 = 5$, $k_2 = 2$, $k_3 = 4$.

3.4.2 Experimental Noise Correlations

Fig. 3.11 shows the resulting cross correlations for cases of active repression (cI-YFP and RFP) and no regulation (cI-YFP and CFP). These functions displayed all features predicted by the model, including a strong dip at a negative lag time due to repression, and positive correlation at zero lag due to global noise in the unregulated case.

Comparing the chromosomal and plasmid-based constructs shows how the relative amplitude of intrinsic and extrinsic noise affects cross correlation functions. Recall that copy number fluctuations in the plasmid increase the effective extrinsic noise level for the genes in the circuit, and reduce the relative importance of intrinsic noise, whose uncorrelated fluctuations average out. Comparing Figs. 3.11a and b shows that the amplitude at $\tau = 0$ is increased on the plasmid relative to the chromosomal construct in both the regulated and unregulated case, reflecting simultaneous correlations. These results confirm model predictions, and demonstrate that regulation can be discriminated even

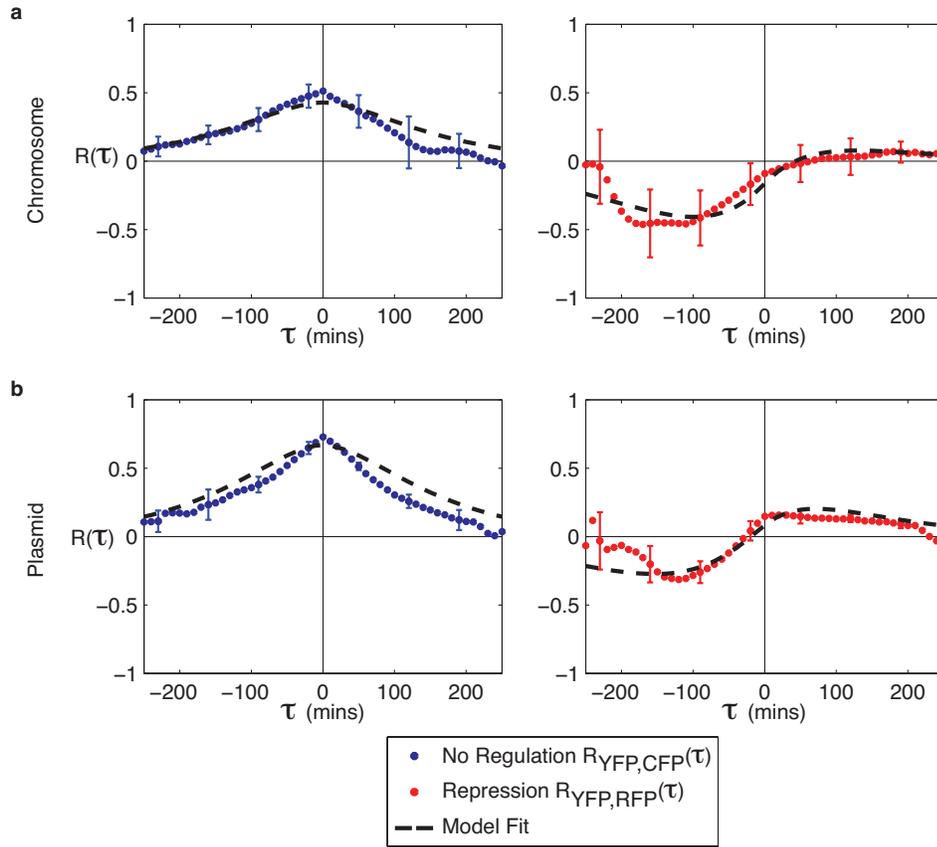


Figure 3.11: Experimental cross correlation functions for chromosomal and plasmid-based circuits. Chromosomal data are averaged over $n = 5$ independent movies (with 100–200 cells per microcolony upon movie completion), $n = 6$ movies for plasmid data. Error bars show standard error of the mean. (a) Cross correlation functions $R_{YFP,CFP}(\tau)$ (blue circles) and $R_{YFP,RFP}(\tau)$ (red circles) and model fits (dashed black line) for the chromosomally integrated construct. For the model fit to the chromosomal construct, $g = -0.01$, $W_E/W_I = 4.5$, and $\alpha_B = 1.7$. (b) Cross correlation functions for the plasmid-based construct. For the model fits, $g = -0.01$, $W_E/W_I = 1.7$, $\alpha_B = 0.5$.

when extrinsic noise amplitudes are large.

The model fits shown in Fig. 3.11 were generated using MATLAB optimization software to minimize the difference between experimental and analytical cross correlation expression from Eqns. (2.26)–(2.26), taking into account standard errors for experimental data points. Three parameters were fit: g , the derivative of the gene regulation function evaluated at the steady-state repressor concentration (see Chapter 2); W_E/W_I , the ratio of extrinsic to intrinsic noise; and α_B , the rate of protein production at steady-state. All other parameters are listed in Table 2.1.

The autocorrelation curves, $R_{x,x}(\tau)$, can also be calculated using these experimental measurements. By using the analytic methods described in Chapter 2 we can calculate expressions for the autocorrelation for the two extreme cases where only extrinsic or intrinsic noise is present. Since our system is influenced by both noise sources, we expect the actual data to fall within these theoretical bounds. With only extrinsic noise acting on a signal $x(t)$ we model gene expression by

$$\begin{aligned}\dot{x} &= -\beta x + E \\ \dot{E} &= -\beta E + \eta_e.\end{aligned}$$

Using the analytic methods described in Chapter 2 we find the normalized autocorrelation expression

$$R_{x,x}^{ext}(\tau) = \exp^{-\beta|\tau|} (1 + \beta|\tau|).$$

With only intrinsic noise the model is

$$\begin{aligned}\dot{x} &= -\beta x + I_x \\ \dot{I}_x &= -\kappa I_x + \eta_x,\end{aligned}$$

finding the normalized autocorrelation to be

$$R_{x,x}^{int}(\tau) = \frac{\kappa \exp^{-\beta|\tau|} - \beta \exp^{-\kappa|\tau|}}{\kappa - \beta}.$$

Note that if $\kappa \gg \beta$ this expression simplifies to $\exp^{-\beta|\tau|}$, as in [36].

The autocorrelation curves for the plasmid data (Fig. 3.12) are bounded by the theoretical limits for autocorrelations due to solely extrinsic or intrinsic noise.

3.5 Alternative Construct

These cross correlation calculations were applied to data from a similar three-color circuit that was built by R. Sidney Cox. The goal of building this circuit, shown in Fig. 3.13, was to create

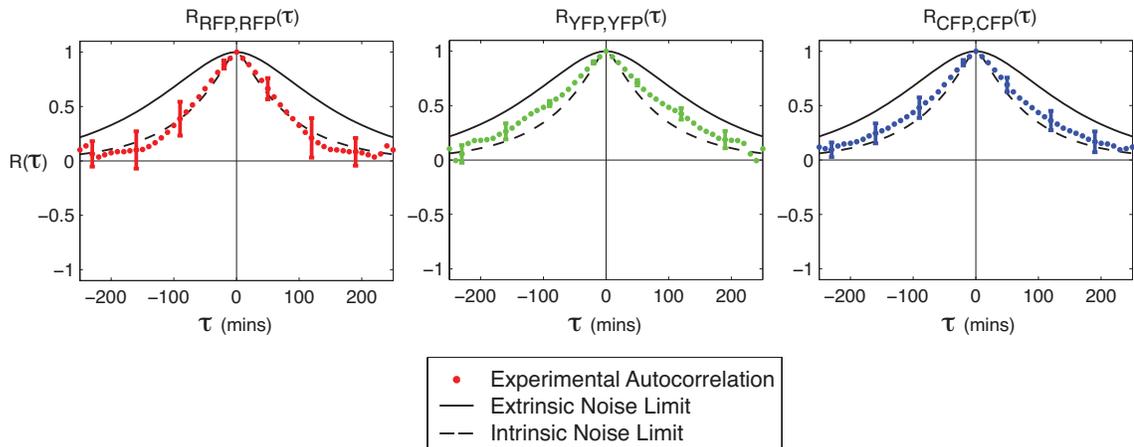


Figure 3.12: Autocorrelation curves for the plasmid data and theoretical noise limits. The parameters $\beta = \log(2)/T_{cc}$ and $\kappa = \log(2)/T_{int}$ are listed in Table 2.1. Error bars are standard error.

a construct for easy measurement of gene expression from three promoters, $P1$, $P2$, and $P3$. To test properties of the circuit, several well-characterized promoters, TetR, LacI, and a combinatorial promoter with LacI and AraC, were placed in the construct. Cox was able to measure cross-talk between the promoters, read-through between genes, and measure noise under different levels of induction for each promoter (R. S. Cox, personal communication).

Although certain features are discernible from standard correlations between genes, temporal correlations contain more information. Fig. 3.14 shows (1) higher correlation between YFP and RFP than the other pairs of genes due to the co-regulation by LacI, (2) correlations between YFP and CFP and CFP and RFP are similar in magnitude and are the result of extrinsic noise in the system, (3) dynamics governing CFP expression and regulation are slower, causing extended periods of positive correlation, (4) all three curves are centered at zero, indicating that no uncharacterized regulation is present.

3.6 Methods and Characterization

3.6.1 Bleaching Times

We measured bleaching times for each of the three proteins in the new synthetic construct. Fig. 3.15 shows the response of fluorescent proteins to prolonged exposure to measurement conditions. These data were gathered by preparing agarose pads with dense cells on them and exposing them to fluorescent light for 10 minutes. After each minute, a snapshot was taken of the cells in all three colors. Exposure times of these measurement snapshots were short—typically less than a second—and are not expected to significantly affect the measurements.

Cells in each column of Fig. 3.15 were bleached in a single color. Rows are measurements of the

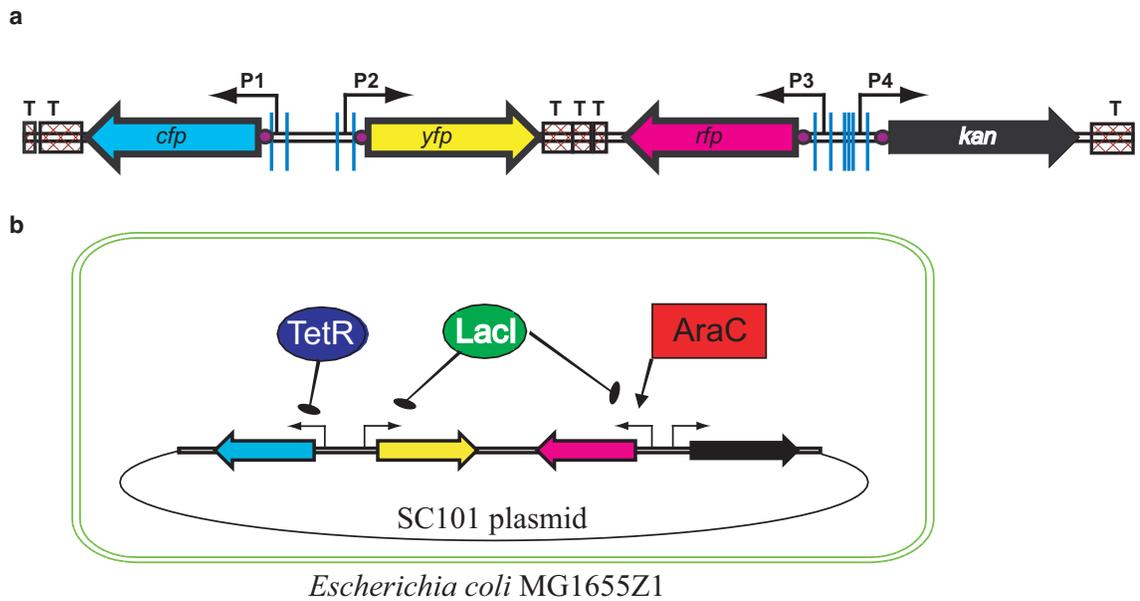


Figure 3.13: Alternate three-color circuit. (a) Construct for measuring expression from three promoters. Terminators are labeled “T,” purple dots are ribosome binding sites, blue lines indicate unique restriction sites for cloning. (b) Promoters used in data analysis. The promoter on RFP is a combinatorial promoter that is repressed by LacI and activated by AraC. Figure courtesy of R. Sidney Cox.

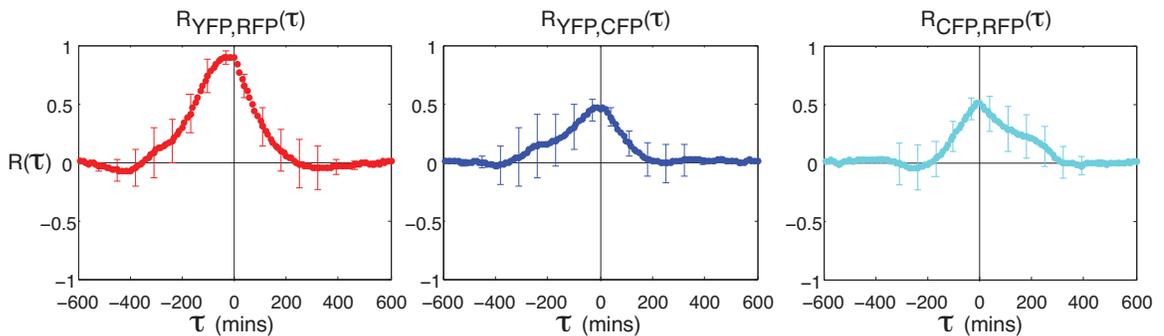


Figure 3.14: Cross correlations between all pairs of genes in the synthetic circuit shown in Fig. 3.13. Error bars are standard error.

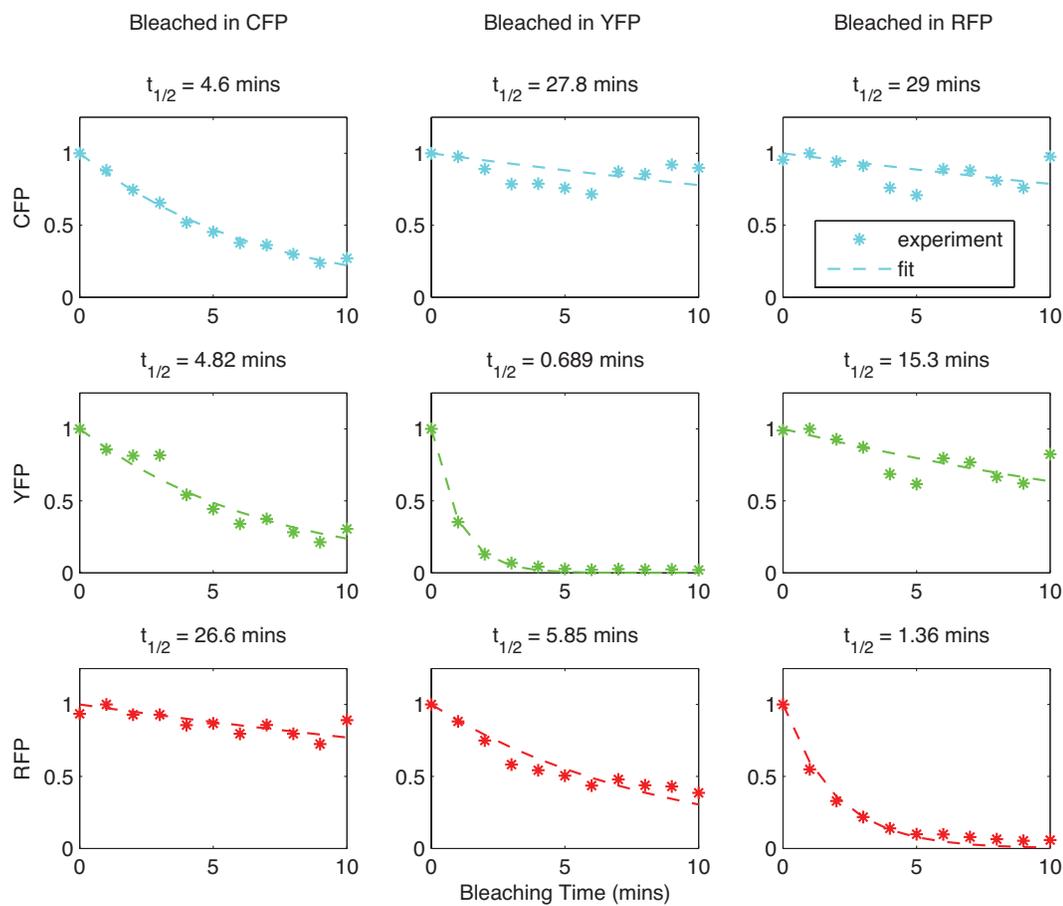


Figure 3.15: Bleaching times for CFP, YFP, and RFP. Times to half signal are shown in the title bars.

three fluorescent protein levels in response to the bleaching. The plots along the diagonal show how the measured signal degrades with time. YFP is the biggest concern for bleaching, with the signal halving every 40 seconds. CFP is remarkably resistant to bleaching. Off-diagonal plots indicate how fluorescence in one channel degrades due to exposure in a different channel. Again, YFP is sensitive to bleaching even for exposure in other channels. For each set of data, a decaying exponential was fit to the curve and the half-life of the protein is listed in the title of each subplot.

Exposure times in movie conditions are typically on the order of 1 second or less per color, taken once every 10 minutes. At these levels bleaching is not expected to be a significant concern.

3.6.2 Methods

The plasmid pNS2- σ VL (Fig. 3.1) was constructed by synthesizing a region starting with the kanamycin promoter and ending just before cI-YFP (synthesis by Blue Heron). The sequences for CFP and the red fluorescent protein mCherry were modified as in GFPuv116 of [43] in regions immediately downstream of the start codon, and were codon-optimized for expression in *E. coli*. The synthesized construct was cloned into the plasmid pZS21-cIYFP [19], replacing the Tet promoter with the synthesized fragment.

The strong promoter on CFP was based on the σ_{70} consensus sequence:



where the capital letters highlight the location of the -35 and -10 boxes.

To integrate the construct into the chromosome, the region from the kanamycin resistance marker through the end of cI-YFP was amplified using PCR with 50 base pair homology regions for galK



included on the end of each primer. The construct was integrated into the *galK* region of MG1655Z1, using the recombineering methods described in [44]. Insertion was verified with colony PCR.

Single colonies were inoculated in selective LB media and grown overnight. This culture was diluted back 1:100 in 1/4 strength LB with 30 $\mu\text{g/ml}$ kanamycin and between 10–15 μM IPTG (varies for different movies). The cells were then grown to OD 0.1–0.2 and diluted back 1:100 in M9 minimal media containing 0.2% glycerol, 0.01% Casamino Acids, 0.15 $\mu\text{g/ml}$ biotin, and 1.5 μM thiamine (we denote this media MGC). Cells were placed on 1.5% MGC low melting temperature agarose pads containing 10–15 μM IPTG and grown at 37°C for 3 hours to equilibrate to the inducer conditions on the pad. The pad was then placed in 200 μl of MGC + IPTG and shaken to release the

cells. These equilibrated cells were placed on a fresh pad for time-lapse imaging. The temperature of the microscope chamber was kept at 32°C for the duration of the movie. Images were taken every 10 minutes in phase and each of the three fluorescent color channels. Fluorescence analysis of cell lineages with done with custom MATLAB software.

Chapter 4

Robustness in Feed-Forward Loops: Clustering of Responses

Even for models of very simple networks, like those in Chapters 2 and 3, describing the system requires many parameters. These parameters are often unknown or uncertain. Consequently, predicting the response of a gene circuit may require inferring gene circuit function from data on circuit structure alone. By using the feed-forward loop as a model system, this chapter introduces a technique for classifying gene circuit function given a set structure. Temporal responses of a comprehensive set of feed-forward loop models are calculated for a range of parameter values. The responses are clustered, and the relation between clusters and circuit types is analyzed. Some designs are robust, producing one unique type of response regardless of parameter selection. Other designs may exhibit a variety of responses, depending upon parameter values.

4.1 Background

As discussed in the Introduction, certain patterns of genetic regulatory interactions occur more frequently than would be expected in randomized networks with similar connection statistics [11]. The feed-forward loop is one such design; an example is regulation of *araBAD* by both the local transcription factor AraC and the global transcription factor CRP in *E. coli* (see review [45] and references therein). Two other naturally occurring feed-forward loops are introduced in Chapter 5.

Given that there are recurring structural designs found in genetic regulatory networks, it is logical to ask: (1) What is the function of a design and (2) why might one design be preferred over others?

First, even once a particular circuit configuration is selected, the function of the circuit is not necessarily transparent. For the feed-forward loop, Mangan and Alon [12] explored several possible circuit functions by using a mathematical model of the feed-forward loop where a signal Sx interacts with X , and a different signal Sy interacts with Y . For a constant level of Sy , they noticed pulsing, ON/OFF, and OFF/ON behaviors in gene expression levels in response to a step input in Sx .

Their efforts produced a preliminary classification of responses for feed-forward loops, but in order to thoroughly characterize feed-forward loop function, it is desirable to explore a larger range of parameters, and to consider circuit types in which the same signal can interact with both X and Y .

Answering the second question—why some designs are preferred—requires an understanding of performance criteria relevant to natural selection in gene circuits [46]. Broad classification of possible circuit functions can eventually help clarify why certain circuit designs are better than others.

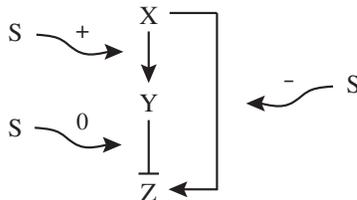


Figure 4.1: Feed-forward loop network motif. X and Y represent transcription factors. Z is the target gene. Activator connections are drawn as normal arrows and repressor connections are drawn as arrows with T-shaped ends. Signal effects are shown with characters $\{+, -, 0\}$.

The feed-forward loop network motif, shown in Fig. 4.1, has two transcription factors, X and Y , which control expression levels of a target gene Z . X additionally regulates transcription of Y . Here, the response of the feed-forward loop refers to the expression level of Z as a function of time.

Signaling molecules also play a significant role in gene expression. Signals may be small-molecule metabolites or other molecules that bind to the transcription factor, enabling or blocking its activity.

As in [12], we consider feed-forward loop models in which each of three genetic regulatory interactions can take on one of two possible values (activator, repressor). Unlike their study, which only considers changes in a signal that enables the global activity of X , we consider models in which a signal may have one of three effects ($+$, $-$, 0 , described below) on each genetic regulatory interaction. Instead of considering just 8 ($= 2^3$), we consider 216 ($= 2^3 3^3$) different ways of wiring a feed-forward loop. Fig. 4.1 is just one example.

4.2 Mathematical Models

The general feed-forward loop is modeled using a pair of nonlinear differential equations:

$$\dot{Y} = \alpha_y \frac{1}{1 + \left(\frac{S_{yx}X}{K_{yx}}\right)^{n_{yx}}} - \beta_y Y \quad (4.1)$$

$$\dot{Z} = \alpha_z \frac{1}{1 + \left(\frac{S_{zy}Y}{K_{zy}}\right)^{n_{zy}}} \frac{1}{1 + \left(\frac{S_{zx}X}{K_{zx}}\right)^{n_{zx}}} - \beta_z Z. \quad (4.2)$$

X is treated as a constitutively expressed protein, modeled here as a constant. α_i is the regulatable transcription rate and β_i is the decay rate through degradation and dilution. S_{ij} , discussed in further

detail below, is a binary value that describes the signal effect. K_{ij} is a threshold value, and the Hill coefficient, n_{ij} , is negative if the connection is an activator and positive if it is a repressor.

Signal interactions are modeled by inserting a binary term, $S_{ij} \in \{0, 1\}$, in the Hill function argument. S_{ij} takes on different values depending upon the level of signal in the environment and the type of signal interaction. Table 4.1 is used to determine S_{ij} .

	Signal < Threshold	Signal > Threshold
+	0	1
-	1	0
0	1	1

Table 4.1: S_{ij} Values

The nonlinear dynamics described in Eqns. (4.1)-(4.2) have a single equilibrium point

$$Y_{eq} = \frac{\alpha_y}{\beta_y} \frac{1}{1 + \left(\frac{S_{yx}X}{K_{yx}}\right)^{n_{yx}}}$$

$$Z_{eq} = \frac{\alpha_z}{\beta_z} \frac{1}{1 + \left(\frac{S_{zy}Y_{eq}}{K_{zy}}\right)^{n_{zy}}} \frac{1}{1 + \left(\frac{S_{zx}X}{K_{zx}}\right)^{n_{zx}}}.$$

Linearizing the system and moving the equilibrium point to the origin we find

$$\begin{pmatrix} \dot{y} \\ \dot{z} \end{pmatrix} = \begin{pmatrix} -\beta_y & 0 \\ \frac{dZ}{dY}|_{eq} & -\beta_z \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y = Y - Y_{eq}$, $z = Z - Z_{eq}$. Since the Jacobian matrix is lower triangular, the eigenvalues are the diagonal matrix elements ($-\beta_y$ and $-\beta_z$). For realistic biological systems $\beta_y, \beta_z > 0$, thus the system has a stable equilibrium point with two real eigenvalues.

4.3 Simulations

The initial conditions for all simulations are the steady state values of Y and Z when the signal is below the threshold level. We are interested in the dynamical behavior that results from changing signal levels.

Fig. 4.2 shows the response of one representative feed-forward loop to changing signal levels. The level of transcription factor Y increases to its steady-state value following a decaying exponential curve. Nonlinear effects cause overshoot in Z before it reaches steady state.

There are seven parameters in Eqns. (4.1)-(4.2) that can be varied. The values of α_i and β_i are selected randomly from a specified range ($[0.1, 10]$ for data shown in the following section). To

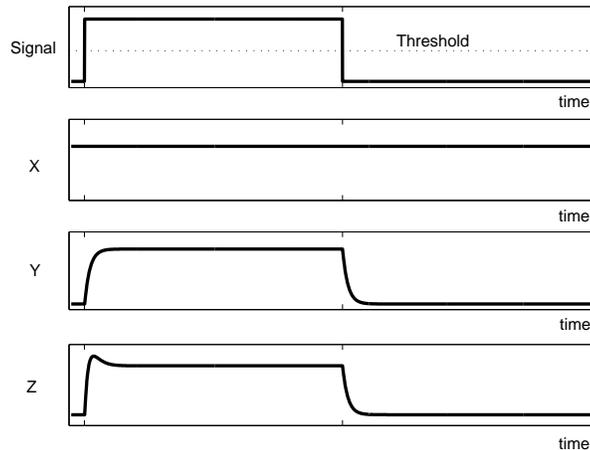


Figure 4.2: Sample response. Simulation results for levels of X , Y , and Z expression as a function of time. The input signal and threshold are shown in the top plot. The configuration of the simulated system is $(X\text{-}Y, Y\text{-}Z, X\text{-}Z) = (\text{activator}, \text{repressor}, \text{activator})$ and $(\text{Signal } X\text{-}Y, \text{Signal } Y\text{-}Z, \text{Signal } X\text{-}Z) = (+, +, +)$. Parameter values shown here are: $\alpha_i = \beta_i = K_{ij} = 1$, $X = 10$, $|n_{ij}| = 2$.

ensure log-uniform parameter selection, for each parameter, a random number, r , is mapped using

$$10^{(2r-1)\log_{10}M_{\alpha\beta}},$$

where $M_{\alpha\beta}$ is the maximum value α_i or β_i can take on (10 for this example). This mapping ensures that we are equally likely to assign values less and greater than 1 to the parameters.

Additionally, the threshold parameters, K_{ij} , are varied. The ratio of transcription factor concentration to the threshold value is the relevant quantity (e.g., $\frac{X}{K_{yx}}$). These three ratios are allowed to take on values less than 1, equal to 1, and greater than 1. All 27 possible combinations are considered.

Recognizing symmetry in signaling effects reduces the size of this problem and decreases computation time. A circuit with signaling interaction type $+$ will respond to an ON pulse in the signal in the same way type $-$ will respond to an OFF pulse.

4.4 Clustering Feed-Forward Loop Responses

Large numbers of feed-forward loops can be modeled with these techniques. For each of the 216 wiring patterns there are multiple threshold and rate parameters that are either unknown or uncertain in biological systems. Broad range limits on parameter values can be assumed to make the problem tractable, but the number of systems remains large.

Although a great number of feed-forward loops can be modeled, many of the feed-forward loop

responses appear to be similar. We use an automated clustering algorithm to classify responses into different categories based upon their similarity.

4.4.1 Clustering Algorithm

A greedy approximation algorithm was used to cluster responses [47]. The algorithm uses a metric $d(x, y)$ that characterizes the distance between x and y . Given an input of a set X of n points x_1, \dots, x_n and a metric d on X , we want to find a set C of K points $c_1, \dots, c_K \in X$ that minimizes $\max_{1 \leq i \leq n} d(x_i, C)$. In other words, we want to cluster the points into K different groups where the size of the largest cluster is as small as possible.

This K-center clustering problem is NP-hard in general, but the approximation algorithm can quickly compute clusters where the maximum error is within a factor of two of the actual solution [47]. Thus, the maximum radius of all clusters is, at worst, two times larger than it needs to be.

The clustering algorithm is performed as follows: First, K points must be selected as cluster centers. The first center, c_1 , is chosen at random. After that ($i = 2, \dots, K$) let c_i be the point x of X that maximizes $d(x, \{c_1, \dots, c_{i-1}\})$. This is equivalent to assigning all the remaining non-center points to clusters, determining which is furthest from its center point, assigning that point as a new center, and throwing the rest of the points back into the pool of non-centers. After all K centers have been assigned, the remaining points x_{K+1}, \dots, x_n are assigned to clusters.

This algorithm is used to cluster feed-forward loop responses. Defining a distance measure, d , is the primary complication in extending the clustering algorithm to the present task. Each response is a vector $z \in \mathbb{R}^N$ where the vector contains the values of Z running from $t = 0$ to $t = N - 1$.

A correlation coefficient is used to measure the distance between two response vectors, z_1 and z_2 :

$$d(z_1, z_2) = \frac{1}{2} - \frac{\langle z_1 - \bar{z}_1, z_2 - \bar{z}_2 \rangle}{2 \|z_1 - \bar{z}_1\|_2 \|z_2 - \bar{z}_2\|_2}, \quad (4.3)$$

where \bar{z} is the mean of z , brackets denote the dot product, and $\|\cdot\|_2$ is the 2-norm. This distance function is designed so that $d(z_1, z_2) = 0$ if $z_1 = z_2$ and $d(z_1, z_2) = 1$ if the two signals are very different. Note that the distance function evaluates to zero for responses that differ only by a multiplicative scaling factor and an offset.

4.4.2 Maximum Error versus Number of Clusters

The maximum error is defined as the largest cluster “radius,” $\max_{1 \leq i \leq n} d(x_i, C)$. As the number of clusters is increased, the maximum error goes down (Fig. 4.3). This value can be plotted as a function of K , the number of clusters. At $K = 1$, we will have a large maximum error value unless

the feed-forward loop responses are all nearly identical. For $K = \text{total } \# \text{ of responses}$, we will have no error since every response is associated with its own individual cluster. If the intermediate curve drops quickly then a small number of clusters can describe almost all of the data.

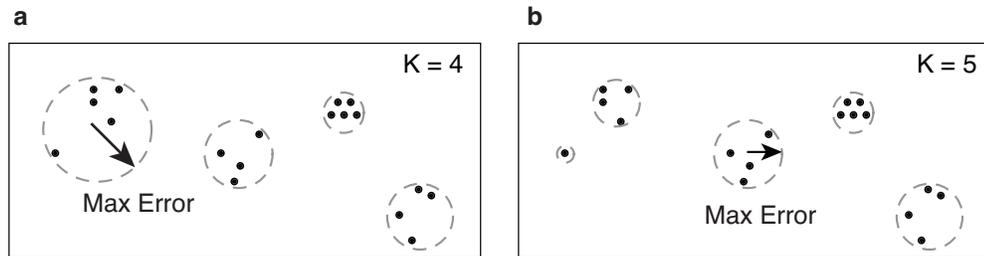


Figure 4.3: Schematic diagram of maximum error for (a) $K = 4$ and (b) $K = 5$. As the number of clusters increases, the maximum cluster error decreases.

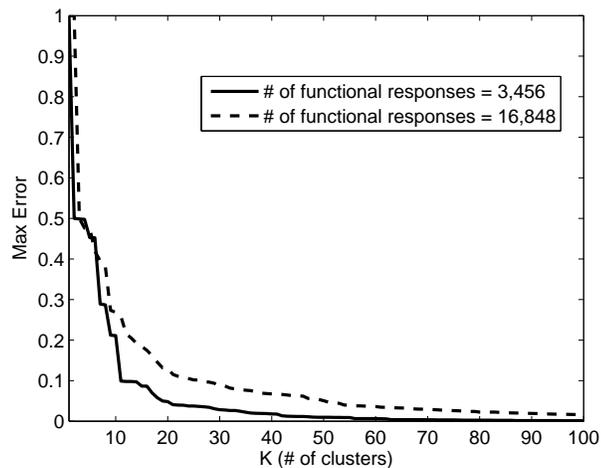


Figure 4.4: Maximum error versus the number of clusters. The results of clustering on two distinct data sets are shown.

Fig. 4.4 shows that, even for large numbers of feed-forward loop responses, the maximum error drops off rapidly with increasing cluster numbers.

4.4.3 Singular Value Decomposition

Since each cluster may contain a large number of responses – thousands in some cases – it would be convenient to have a simple way to represent data. Singular value decomposition is used to generate a representative trace that describes the most significant principal component of all of the responses in a cluster.

Singular value decomposition has been used in other biological applications to compress data into a simplified, more understandable form [48]. In this work the singular value decomposition of

a matrix $A \in \mathbb{R}^{M \times N}$ is taken:

$$A = USV^T.$$

M is the number of feed-forward loop responses we are comparing and N is the number points in time. S , U , and V come from the standard definition of singular value decomposition.

The first right singular vector (the first column of V) is the singular vector associated with the largest singular value. This vector describes the principal component of all of the response data listed in the A matrix and provides a single representative response to associate with a cluster.

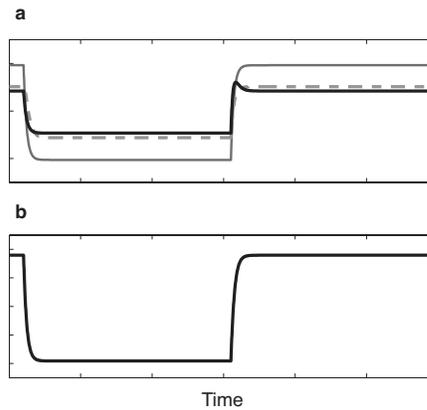


Figure 4.5: Singular value decomposition example. a) Three representative $Z(t)$ responses from one cluster. b) The first right singular vector of a matrix containing all responses from that cluster.

An example of how singular value decomposition can be used to represent many responses is shown in Fig. 4.5. Fig. 4.5a shows 3 responses plotted on top of each other. In reality, this is a small subset of all responses that fall into this cluster type. The primary singular vector associated with the complete set of responses is shown in Fig. 4.5b.

4.5 Results

The clustering approach associates the responses of the 216 feed-forward loop models with a small number of distinct patterns. These patterns can be used to classify the behavior of an individual circuit over a range of parameter values. The number of clusters it takes to describe a particular circuit configuration can be used as a measure of how robust a circuit is to parameter variation.

4.5.1 Representative Cluster Traces

A relatively simple example is presented to illustrate the utility of clustering. The data shown in Fig. 4.7 are the result of clustering on a set of 3,456 responses. All 216 possible circuit configurations

are represented. Within each configuration only parameters α_i and β_i from Eqns. (4.1)–(4.2) are varied.

$K = 11$ clusters is chosen as a cutoff point because the maximum error is acceptably small (see Fig. 4.4). Beyond this point additional cluster types represent similar responses but with differing temporal characteristics. For example, the rise times, settling times, and overshoot behavior may be different for the additional cluster types. The utility of clustering lies in its ability to segregate responses into broad class types, allowing for a qualitative understanding of possible circuit functionality. In particular, this method will be useful for considering circuits that have more complicated responses (e.g., responses to input signals that are more complicated than a step function).

Fig. 4.7 shows representative singular vectors from each of the 11 clusters. These are the responses, $Z(t)$, to the input signal shown in Fig. 4.6.

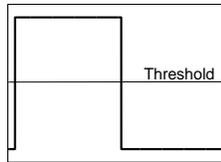


Figure 4.6: Signal level as a function of time.

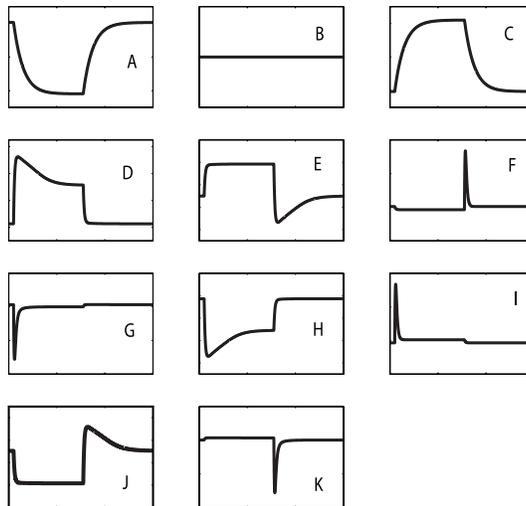


Figure 4.7: Representative responses (Z versus *time*) from each of 11 clusters. Letter labels are used for reference in Table 4.2.

When the threshold values, K_{ij} , are varied in addition to α_i and β_i the result is a large set of functional responses that do not segregate as logically into individual clusters. Even when this is the case, the clustering technique can still be applied to yield a qualitative picture of possible responses. The case with 16,848 responses shown in Fig. 4.4 corresponds to a widely explored range

of parameters, but the maximum error still drops off rapidly. If an acceptable error value is chosen, clustering can be performed to within this margin of error.

In an exploration of the more complete parameter space, the cluster types seen in Fig. 4.7 are preserved, but several additional clusters are added. For example, selecting $K = 15$ clusters produces the four additional cluster types shown in Fig. 4.8.

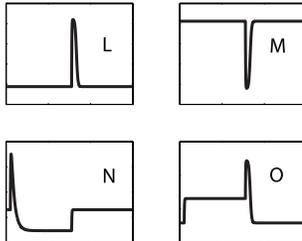


Figure 4.8: Representative responses for cluster types added when additional parameters are varied.

Even when exploring the complete parameter space, some system configurations fall into the same cluster type regardless of the parameter values selected for α_i , β_i , and K_{ij} . The responses of these genetic regulatory configurations are particularly robust to parameter changes.

4.5.2 Distributions of Responses

Table 4.2 lists cluster types for various system configurations. Each row corresponds to one particular configuration: a set of genetic regulatory and signal interaction types. The columns labeled A–O correspond to the cluster types labeled in Figs. 4.7 and 4.8. The numbers in the row tell which cluster this system’s responses fall into. For some system configurations, varying parameter values causes the response to fall into different clusters. The rows are normalized by the total number of cases that were run. The entries shown in Table 4.2 are a subset from a larger table, given in Appendix C.

Table 4.2’s entries do not indicate how “different” responses are within a cluster. Selecting the number of clusters (Fig. 4.4) sets the upper bound on the error within each cluster. For the 15 cluster case, all responses within a cluster are within a distance of 0.18 of each other, as measured by Eqn. (4.3).

The entropy of each response distribution in Table 4.2 is calculated by using the standard definition of Shannon entropy:

$$-\sum_{i=1}^{15} p(i) \log_2 p(i),$$

where $p(i)$ is the percentage of responses that fall into cluster i .

Rows which have a 1 associated with one cluster type and 0s for all the rest (entropy = 0) are

			Signal	Signal	Signal																Entropy
X-YY	ZX-Z	X-Y	Y-Z	X-Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O		
act	rep	rep	+	+	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act	rep	rep	+	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act	rep	rep	+	+	-	0.41	0	0.51	0.02	0	0	0	0.01	0	0	0	0	0.06	0		1.41
act	rep	rep	+	0	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act	rep	rep	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act	rep	rep	+	0	-	0.38	0	0.51	0.03	0	0	0	0.01	0	0	0	0	0.07	0		1.49
act	rep	rep	+	-	+	0.90	0	0	0	0.09	0	0	0	0	0	0	0.01	0	0		0.52
act	rep	rep	+	-	0	0	0.11	0	0	0.06	0	0	0	0	0.50	0	0.33	0	0		1.61
act	rep	rep	+	-	-	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
act	rep	rep	0	+	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act	rep	rep	0	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act	rep	rep	0	+	-	0.48	0	0.52	0	0	0	0	0	0	0	0	0	0	0	0	1.00
act	rep	rep	0	0	+	0.99	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05

Table 4.2: Percentage of cluster types exhibited by circuit configurations. Subsection of a larger table, given in Appendix C. *act* = activator, *rep* = repressor.

particularly robust because parameter variations do not change the function of the system. Exact entries in the table are dependent upon the details of parameter selection and the distance measure used to cluster data.

Clustering provides a logical grouping of response types without prior knowledge of the behavior that a network may exhibit. The response pattern of a cluster can then be interpreted in a biologically meaningful way. For example, cluster *A* (Fig. 4.7) is associated with repressible circuits, where gene expression decreases upon an increase in signal level. Similarly, cluster *C* is associated with inducible circuits, where gene expression increases upon signal increase. For circuits associated with cluster *B*, gene expression is unresponsive to changes in signal. Pulsed gene expression responses, both with and without steady state changes, are seen in the remaining cluster types.

This chapter presents a method for identifying functional capabilities of a genetic network given its structure. In our analysis of feed-forward loop models, responses were organized into a relatively small number of clusters. Some feed-forward loop types show non-robust behavior, suggesting that these circuits do not have unique information processing roles. This clustering technique allows for such quick, qualitative intuition into the function of a system. Insight from clustering will be particularly helpful if the state space and parameter space are even larger than those presented in the feed-forward loop example here.

Although we consider models of feed-forward loops in isolation, in nature gene circuits are embedded within the context of the entire molecular network of the cell. Nevertheless, considering isolated gene circuit models can reveal insights into the cellular response to signals. Such models have already proved to be useful in design of synthetic gene circuits, for example, in the design of a toggle switch [49], an oscillator [40], and a circuit whose design may be selected to exhibit either toggle switch or oscillatory behavior [50]. The present technique can help to narrow down which system types and parameter ranges exhibit a desirable behavior, given a broad class of possible designs.

In the future it will be interesting to explore the implications of robustness of responses in real biological systems. In particular, is robustness necessarily a desirable trait for a genetic circuit? If the circuit is locked into one role it may not be capable of evolving in alternative environments. In addition, natural selection can act to enhance the populations of organisms that are sensitive, rather than robust, to mutations in gene circuits. This process has been used previously to explain patterns in the use of activator and repressor control in natural genetic regulatory interactions [51]. It would be interesting to consider tradeoffs involving robustness in the context of the evolution of feed-forward loop configurations and other aspects of gene circuit design.

Chapter 5

Noise in Feed-forward Loops for Galactose Utilization

In this chapter we study two naturally occurring feed-forward loops that are involved in galactose metabolism and transport. Despite having network structures that are capable of producing dynamic, temporally diverse responses we find, by measuring dynamic noise correlations, that in their natural context these feed-forward loops are inactive. By perturbing genetic conditions the activity can be restored.

5.1 Galactose Regulation

Although *E. coli* prefer glucose as a sugar source, if glucose is not present and other sugars are available, cells will turn on the machinery to metabolize these alternate sugars. The galactose network in *E. coli* contains regulatory circuitry that implements the logic `if NOT glucose AND galactose` [52]. When this logic function is true there are two classes of genes that are turned on: galactose metabolism genes and galactose transport genes, which are known collectively as the gal regulon [53]. The metabolism and transport pathways are regulated by many of the same molecular components and the network diagrams that describe which genes affect each other are nearly conserved.

Genes for galactose metabolism and transport are turned on in response to two signals: cAMP and galactose [52]. When glucose is not present, cAMP is produced in cells and binds to the global regulator, CRP (cAMP Repressor Protein). The cAMP-CRP complex functions as an activator, turning on genes in the galactose regulon. While cAMP acts as a positive signal, galactose acts as a negative signal. Galactose binds to two repressor proteins, GalR (galactose repressor) and GalS (galactose isorepressor). GalS and GalR repress transcription by modulating the α subunit of RNA polymerase when it is bound to the promoter. The addition of galactose interrupts this process, but the galactose-GalR/S complex does not necessarily dissociate from the promoter [54, 55]. Thus,

the presence of both cAMP and galactose are necessary to turn on the metabolic and transport machinery needed so that galactose can be used as a sugar source.

The three proteins CRP, GalR, and GalS control the majority of the genes responsible for the metabolism and transport of galactose. The promoter structures of genes they regulate show many similar features with minor variations from promoter to promoter (Fig. 5.1). GalS and GalR are very similar proteins: 53% of their sequence is identical and 85% is similar. They belong to a larger family of transcriptional repressors known as the GalR-LacI family. As a result of their similarity, GalR and GalS bind to many of the same DNA binding sites and have similar features. For example, both proteins are dimers [56], are autorepressed [52, 57], and are capable of repressing each other [52]. In contrast to *galR*, *galS* is activated by the cAMP-CRP complex; *galR* has a putative CRP binding site, but has repeatedly been shown to be unresponsive to CRP [52, 53]. Despite controlling many of the same targets, the binding affinities of GalS and GalR for different genes in the gal regulon are often quite specific [53, 56].

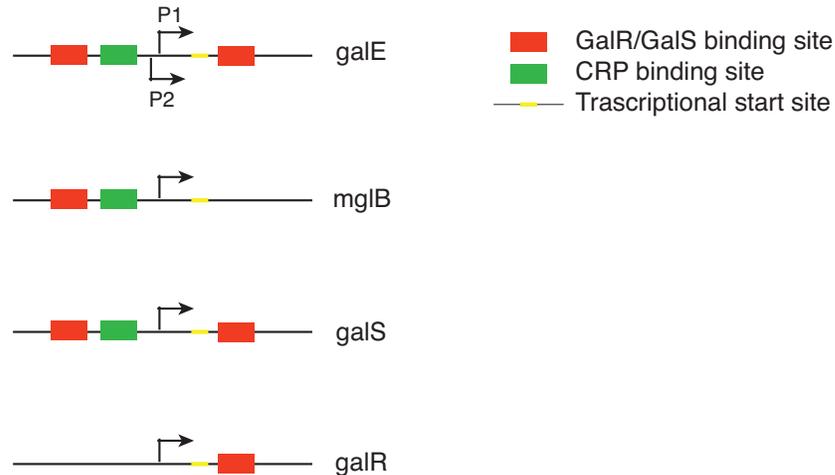


Figure 5.1: Promoter architecture for gal regulon genes. Binding sites for CRP (red boxes), GalR and GalS (green boxes), promoters (black arrows), and the transcriptional start codon (yellow bar). Many genes in the gal regulon have similar promoter and transcription factor binding sites with minor variations. This figure is based upon a diagram from [52].

Below, we go through some of the particular players in the gal regulon and describe in detail how they respond to CRP, GalR, GalS, and the signals cAMP and galactose.

5.1.1 Galactose Metabolism

There are six genes involved in the preliminary steps of galactose metabolism: *galE*, *galT*, *galK*, *galM*, *pgm*, and *galU* [53]. The first four of these genes are arranged in an operon, *galETKM*. We focus on control of this operon, referring to it as *galE* for concise notation.

galE has two promoters, P1 and P2, that control its expression (Fig. 5.2). *In vitro* studies have tested the roles of these promoters individually. In the absence of glucose, CRP activates transcription from P1 and represses it from P2. GalR and GalS play the opposite role, repressing transcription from P1 and activating it from P2 [52, 58]. In the presence of galactose, repression is relieved and transcription occurs primarily from the P1 promoter. *In vivo*, GalR plays a primary role in controlling expression of *galE*. There are two GalR binding sites O_E (external) and O_I (internal) that bracket the P1 and P2 promoters, shown in Fig. 5.3. In the absence of galactose, GalR binds to these two operators and causes the DNA to loop, obscuring the P1 and P2 promoters and inhibiting transcription. Adding galactose interrupts looping and allows for transcription from P1 and P2 as seen in the *in vitro* studies [59]. DNA looping by GalR requires formation of a structure known as the repressosome, which consists of two GalR dimers—one bound to O_E and one to O_I —and HU, a bacterial histone-like protein [60, 61]. Interestingly, operator mutation studies have shown that O_E and O_I can be replaced by LacI binding sites and full repression is maintained, suggesting that DNA looping is the major factor in repressing transcription of *galE* in the absence of galactose [58].

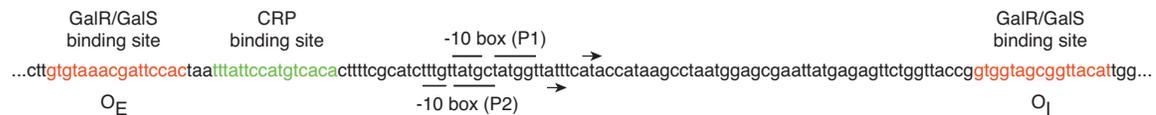


Figure 5.2: Sequence of the promoter region for *galETKM* operon. -10 boxes (and extended -10 boxes) for the two promoters are overlined (P1) and underlined (P2). Small arrows indicate transcriptional start sites for the two promoters. Red text is GalR/GalS binding sites, green is CRP.

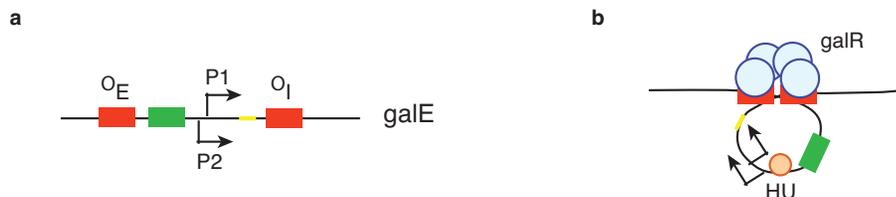


Figure 5.3: GalR repression by looping. (a) Unlooped orientation of the GalR binding operators, O_E and O_I . (b) GalR dimers tetramerize, bringing together O_E and O_I and obscuring the promoter region and transcriptional start site.

Although GalR appears to be the primary repressor of *galE*, GalS is also capable of regulating its expression. When *galR* is deleted, *galE* can be further induced by the addition of galactose, a phenomenon known as ultrainduction [62]. Deleting *galS* removes this effect. Several mechanisms have been proposed to account for this redundant regulation. Weickert and Adhya [62] postulate that GalS may serve as a backup control, alternatively, GalR and GalS may play different temporal roles, or GalR and GalS may respond differentially to levels of galactose. Work by Mangan, et

al. [63] suggests that GalS acts as part of a feed-forward loop to speed the response of galactose metabolism genes after glucose depletion.

5.1.2 Galactose Transport

Four genes are primarily responsible for transporting extracellular galactose into the cell: *mglB*, *mglA*, *mglC*, and *galP* [53]. A three-gene operon, *mglBAC*, makes up the high affinity galactose transport system, which is active when extracellular galactose is low. *mglB* is involved in binding galactose, while MglA and MglC are membrane-associated proteins. The high affinity system is primarily regulated by GalS and only weakly controlled by GalR [52]. In contrast, the low affinity galactose system is active when extracellular galactose is high and is primarily (likely solely) regulated by GalR [56]. GalP, galactose permease, is the major player in the low affinity system and is a membrane transport protein. Additional galactose transport systems exist, but are much less efficient than *mglBAC* and *galP* [53]. We focus on the high affinity transport system, *mglBAC*, abbreviated as *mglB*.

A single promoter controls expression of *mglB*. It contains a single GalR/GalS binding site and a single CRP site. The structure of the promoter is very similar to the P1 promoter on *galE* and has similar behavior: CRP activates expression in the absence of glucose, and GalR/GalS represses expression in the absence of galactose. When galactose is added, repression is relieved and *mglB* is expressed at higher levels. Unlike the *galE* promoter, because there is only a single GalR/GalS binding site DNA looping is not used to inhibit transcription. *galR* deletion experiments had little affect on *mglB* expression, while they had a strong affect on *galE*, suggesting that GalS plays a primary regulatory role [62]. This is further supported by the co-localization of the *galS* and *mglB* genes on the chromosome [53].



Figure 5.4: Sequence of promoter region for *mglBAC* operon. Labeling and symbols are consistent with Fig. 5.2.

5.1.3 Structure of Regulatory Networks

Control of *galE* and *mglB* is implemented by regulatory circuits with very similar structures. Both operons are the target of feed-forward loops involving CRP and GalS that respond to the signals cAMP and galactose. Fig. 5.5 contains all of the regulatory connections for controlling *galE* and *mglB* that have been proposed in the literature. Databases like RegulonDB [7] contain information

in this form, which is a useful starting point for understanding regulation and cataloging all possible interactions. However, as pointed out in [52], not all of the regulatory connections that are listed in Fig. 5.5 are necessary to produce the cellular responses that are observed *in vivo*. In particular, we show that the context in which the circuit operates is very important for determining its function.

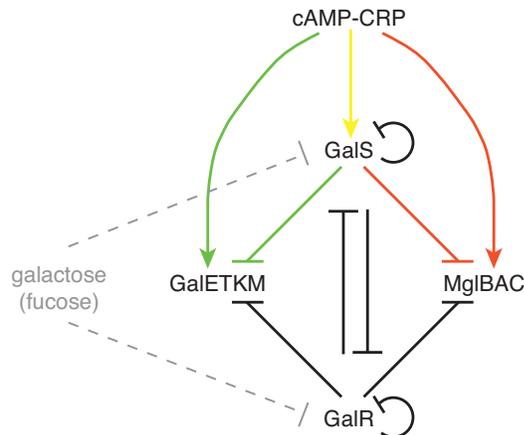


Figure 5.5: Regulatory network controlling expression of *galE* and *mglB*. Two feed-forward loops are highlighted in red and green, with a common element colored in yellow. Solid lines represent transcriptional regulation and dashed lines are non-transcriptional interactions.

5.2 Clustering of Type I Incoherent Feed-Forward Loop Responses

In Chapter 4 we saw that feed-forward loops can exhibit different types of responses depending upon the regulatory circuitry, response to signals, and system parameters. Both the feed-forward loops governing galactose metabolism and transport are Type I Incoherent feed-forward loops [12] that respond to two signals. We consider these two signals independently, though the analysis can be extended to multiple signals, as in [64]. Galactose does not affect the X-Y or X-Z connections, but inhibits repression in Y-Z, thus, in the notation from Chapter 4, the influence of galactose is type $\{0, -, 0\}$. cAMP has a positive effect on the X-Y connection and X-Z connection, but has no effect on Y-Z, so the signal's influence is $\{+, 0, +\}$.

Results from screening circuits of this type over a broad range of parameters are summarized in Fig. 5.6. The response of the feed-forward loop to a pulse in galactose is very stereotyped: for all combinations of parameters explored the circuit exhibits simple activation. A pulse in cAMP can exhibit responses ranging from simple activation to accelerated response with overshoot in reaction to an ON step in the signal. Activation without an accelerated response is the most common response, but a significant fraction (20%) of the conditions tested showed pulsing behavior.

(Fig. 5.8). As in [63], we used fucose, a non-metabolism analog to galactose, as an inducer. Like galactose, fucose binds to GalS and GalR and inhibits ability to repress. Thus, when fucose is present autoregulation and repression are relieved, further simplifying the circuit diagrams shown in Fig. 5.7.

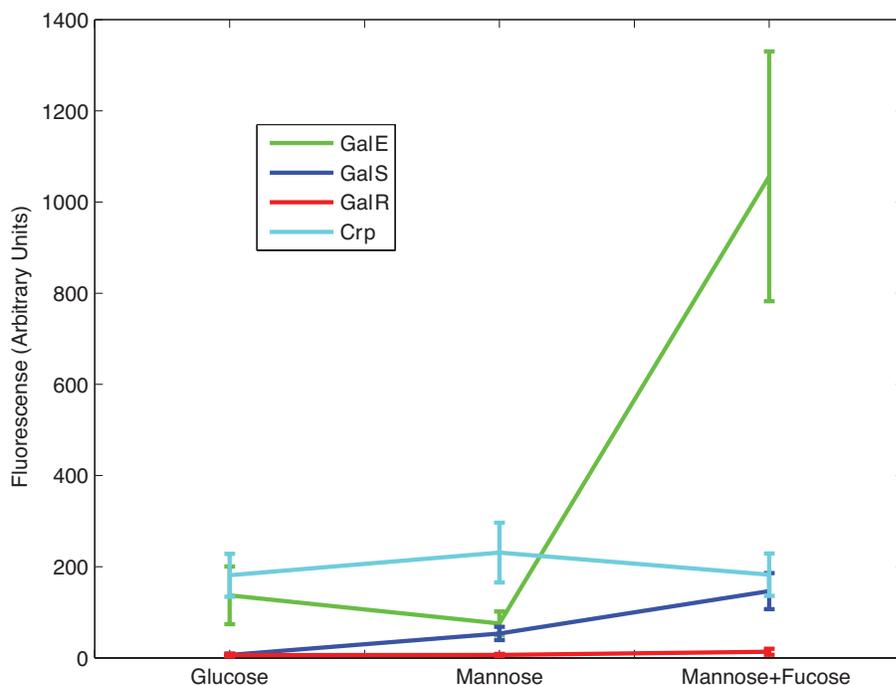


Figure 5.8: Response of P_{galS} , P_{galR} , P_{crp} , and P_{galE} to different sugars. Error bars show standard deviations. Mannose and glucose are 0.8%, fucose is 20 mM, all cells were grown in M0 (see Methods). Expression levels were measured individually using plasmid-based promoter-GFP fusions from [65]. Fluorescence data have been background subtracted. P_{mglB} is expected to have measurable expression levels based on measurements from [66].

To test the activity of the *galE* and *mglB* feed-forward loops we constructed a set of promoter-fluorescent protein fusions for pairwise measurement of gene expression from *galS*, *galE*, and *mglB*. Promoters for these genes are the same as in [65], placed upstream of the *yfp* and *cfp* genes used in the synthetic circuit described in Chapter 3. The promoters and fluorescent proteins were oriented in opposite directions (Fig. 5.9) to minimize read-through, and cloned next to a kanamycin resistance marker for selection. The synthetic constructs were integrated into the *intC* region of the MG1655 chromosome and colonies were screened for correct insertion length and then verified with sequencing.

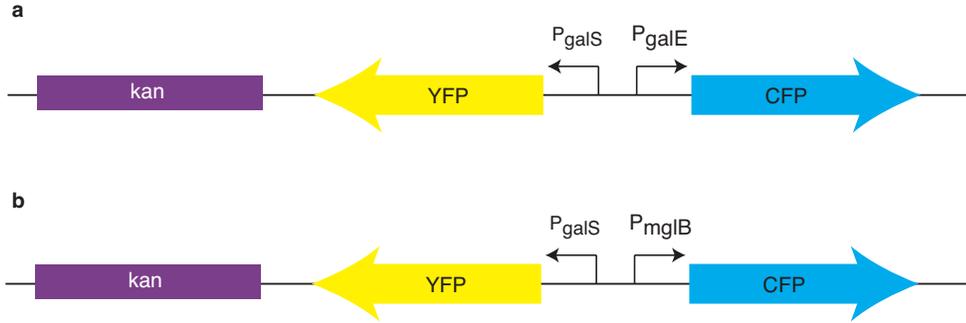


Figure 5.9: Promoter fusions for measuring (a) P_{galS} and P_{galE} and (b) P_{galS} and P_{mglB} .

5.4 Noise Correlations to Infer Activity

5.4.1 Theoretical Predictions

Promoter fusions, as compared to protein-gene fusions like that used in the synthetic circuit in Chapter 2, reduce perturbation of endogenous circuit function, and allows for signal amplification using strong ribosome-binding sites for reporter gene expression. Potential drawbacks are that intrinsic noise is no longer measured directly and if the reporter dynamics differ significantly from those of the gene of interest this will appear in the cross correlation function.

Fig. 5.10 illustrates how the expression levels of two promoters can be measured using fluorescent reporter proteins. Protein A represses B , while F is a reporter for A , and G a reporter for B .

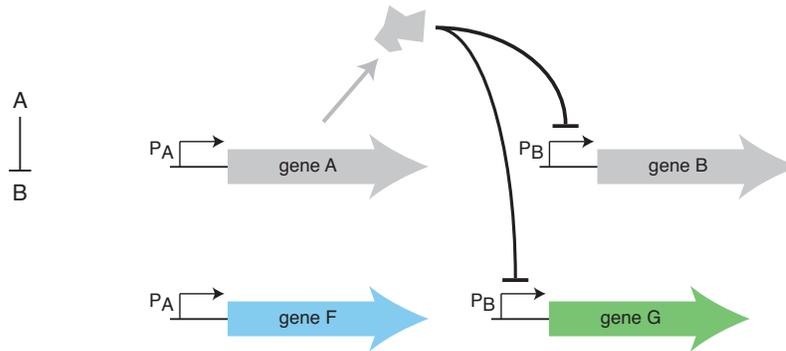


Figure 5.10: Schematic of promoter fusion for A repressing B . $gene F$ and $gene G$ are fluorescent proteins that can be measured to report the expression levels of promoters A and B , respectively.

This system can be modeled using the linearized approximation by

$$\begin{aligned}\dot{a} &= -\beta a + E + I_a \\ \dot{b} &= -\beta b + g_{ab}a + E + I_b \\ \dot{f} &= -\beta_f f + E + I_f \\ \dot{g} &= -\beta_g g + g_{ab}a + E + I_g.\end{aligned}$$

Here, f and g model the reporter proteins, which are expressed in the same way as the original proteins, a and b , but have different sources of intrinsic noise. The degradation rates β_f and β_g are one example of a way that the reporter dynamics could differ from the system dynamics. Fig. 5.11 shows examples of cross correlation functions generated by reporter proteins that decay more quickly than those in the original system. Cross correlation functions for the full system, assuming direct

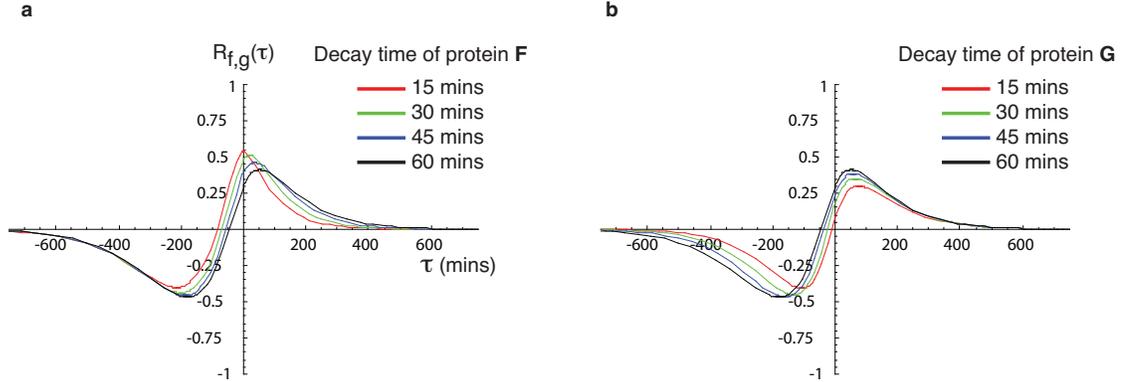


Figure 5.11: Differing reporter and system dynamics. (a) Protein F degrades more quickly than A , B , and G . (b) Protein F degrades more quickly than A , B , and G . For all calculations $g_{ab} = -0.0125$, $\theta = 0.064$, and $\beta = \beta_f = \beta_g = \text{Log}[2]/60$ unless specified in the figure caption where the decay time, T_{decay} , is used to calculate $\beta_i = \text{Log}[2]/T_{decay}$.

measurement of proteins A and B are nearly identical to those generated when $\beta = \beta_f = \beta_g$.

We used the analytic methods discussed in Chapter 2 to calculate expressions for the cross correlation function of a feed-forward loop. Fig. 5.12 shows the expected shape of the cross correlation function with and without an inducer that inhibits repression by Y . The feed-forward loop was modeled by using the linearized system of equations

$$\dot{x} = -\beta x + E + \eta_x \quad (5.1)$$

$$\dot{y} = -\beta y + g_{xy}x + E + \eta_y \quad (5.2)$$

$$\dot{z} = -\beta z + g_{xz}x + g_{yz}y + E + \eta_z, \quad (5.3)$$

with reporter proteins for Y and Z that have dynamics

$$\dot{f} = -\beta f + g_{xy}x + E + \eta_f \quad (5.4)$$

$$\dot{g} = -\beta g + g_{xz}x + g_{yz}y + E + \eta_g. \quad (5.5)$$

Parameter values are listed in the caption of Fig. 5.12. These equations represent a simplification over those in Chapter 2 because we model intrinsic noise as white noise rather than using an Ornstein-Uhlenbeck process with short correlation time. The equations in this form are simpler and give

similar results, but it is straightforward to calculate the same expressions using the more accurate intrinsic noise terms. In addition, for a complete match to the galactose feed-forward loops, the model should include autorepression by Y . This can be modeled by adding the term $g_{yy}y$ to the second equations for \dot{y} and \dot{f} . The cross correlation between Y and Z in the active feed-forward loop

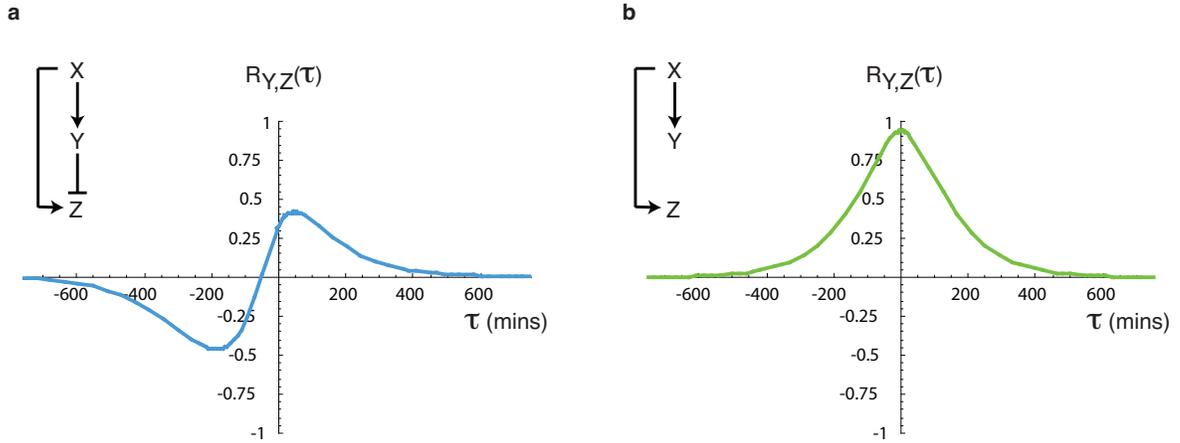


Figure 5.12: Theoretic predictions for cross correlation functions from galactose feed-forward loops (a) without and (b) with fucose. Parameter values used in these calculations are $g_{xy} = g$, $g_{xz} = g$, $g_{yz} = -g$, where $g = 0.0125$, $W_i = 1$ for $i = \{x, y, z\}$ and $W_e = 0.064$. For (b), $g_{yz} = 0$.

is similar to the response of a simple repressor, but additional positive correlation that is symmetric about zero lag is also present. The effect of X is extrinsic to both of the measured variables and consequently acts very similarly to extrinsic noise.

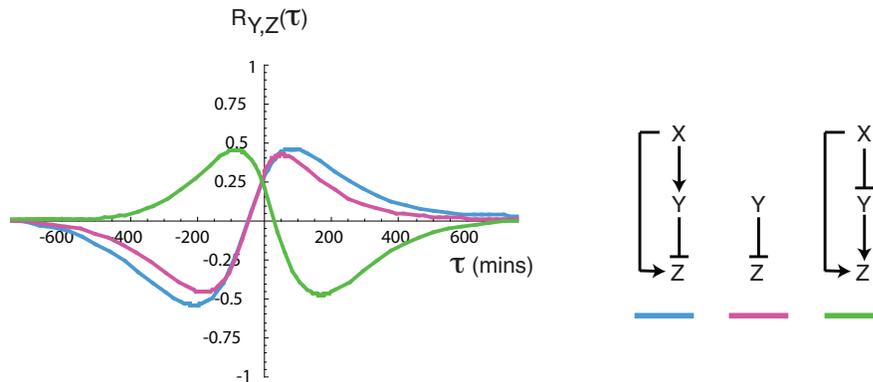


Figure 5.13: Theoretic predictions for cross correlations from alternate forms of regulation. All cross correlations are between Y and Z , regulatory architectures are indicated in the figure. The third gene circuit is an Incoherent Type 4 feed-forward loop [12]. Parameters are the same as those listed in Fig. 5.12, except $g = 0.0125$ and for the third circuit $g_{xy} = -g$ and $g_{yz} = g$.

5.4.2 Experimental Data

We measured the cross correlation between reporters for *galS* (YFP) and *galE* (CFP) in the presence and absence of fucose (Fig. 5.14). The cross correlation function without fucose has a peak at zero and is symmetric. This indicates that *galS* and *galE* are affected by some of the same noise sources and regulatory proteins, but, in these conditions, *galS* does not have a distinct regulatory effect on *galE*. Adding fucose inhibits repression by GalS and GalR, but the cross correlation curves with and without fucose are indistinguishable. These results suggest that, in the conditions we tested, GalS does not play an active regulatory role in controlling expression of *galE*.

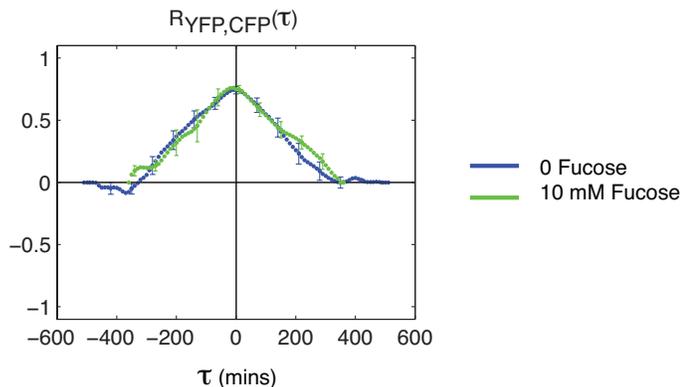


Figure 5.14: Cross correlations between P_{galS} -YFP and P_{galE} -CFP at 0 and 10 mM fucose. Error bars are standard error averaged across multiple movies. $n = 15$ movies for 0 mM fucose, $n = 8$ for 10 mM fucose.

To test whether *galE* could be controlled by *galS* in some contexts, we created a *galR* deletion strain with P_{galS} and P_{galE} reporters. The goal of this experiment was to explore the activity of the CRP/GalS/GalE feed-forward loop in the absence of repression due to looping by GalR. Without fucose the cross correlation between P_{galS} -YFP and P_{galE} -CFP is no longer symmetric and shows clear signs of repression of *galE* by GalS (Fig. 5.15). There is still a strong peak at zero lag as a result of extrinsic noise and noise in CRP, both of which are extrinsic to the measured signals. When fucose is added, repression by GalS is inhibited and the cross correlation curves are symmetric, like those seen in the presence of GalR. These data suggest that although GalS can play a regulatory role in the control of *galE* by acting a repressor of its production, in natural contexts the dominant regulatory role is played by GalR.

Note that the peak value of the cross correlation curves decrease when GalR is deleted (compare Fig. 5.14 and Fig. 5.15). This indicates that the presence of GalR was adding to the correlation between *galS* and *galE*.

Based on information in the literature, the CRP-GalS-MglB feed-forward loop should behave in a simpler fashion since it lacks looping by GalR, which was a confounding factor in the GalE feed-forward loop. Surprisingly, noise-generated cross correlations between P_{galS} -YFP and P_{mglB} -CFP

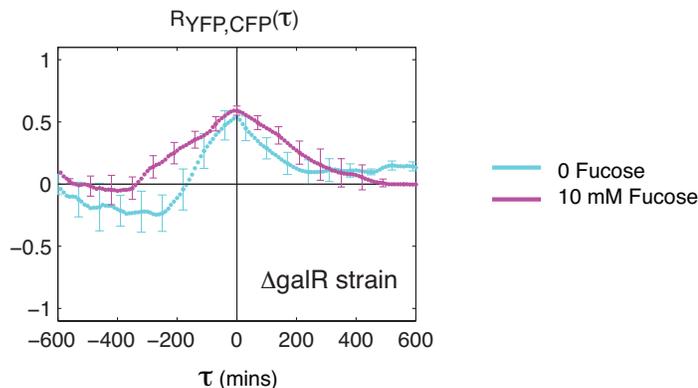


Figure 5.15: Cross correlations between P_{galS} -YFP and P_{galE} -CFP at 0 and 10 mM fucose in a $galR$ deletion strain of MG1655. Error bars are standard error averaged across multiple movies. $n = 9$ movies for 0 mM fucose, $n = 9$ for 10 mM fucose.

indicate that GalS and MglB are not strongly linked, even in the absence of fucose (Fig. 5.16). Thus, it appears that in the cellular contexts that were measured for these experiments, neither galactose feed-forward loop is actively regulating its target gene. Measurements in [66] suggest that GalS may be active in other regulatory regimes.

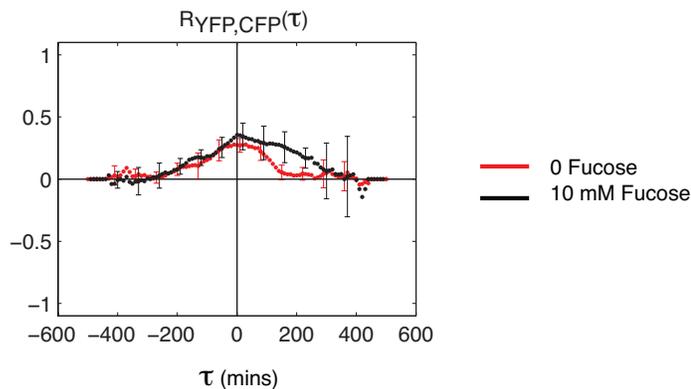


Figure 5.16: Cross correlations between P_{galS} -YFP and P_{mglB} -CFP at 0 and 10 mM fucose. Error bars are standard error averaged across multiple movies. $n = 6$ movies for 0 mM fucose, $n = 5$ for 10 mM fucose.

5.5 Methods and Characterization

5.5.1 Expression Levels from Static Data

Measurements from snapshot data are shown in Fig. 5.17. Addition of 10 mM fucose raises expression levels of all three proteins by relieving repression by GalR and GalS. When GalR is deleted, expression of GalS and GalE increase further with fucose addition, but remain low without fucose—likely

the result of residual repression by GalS.

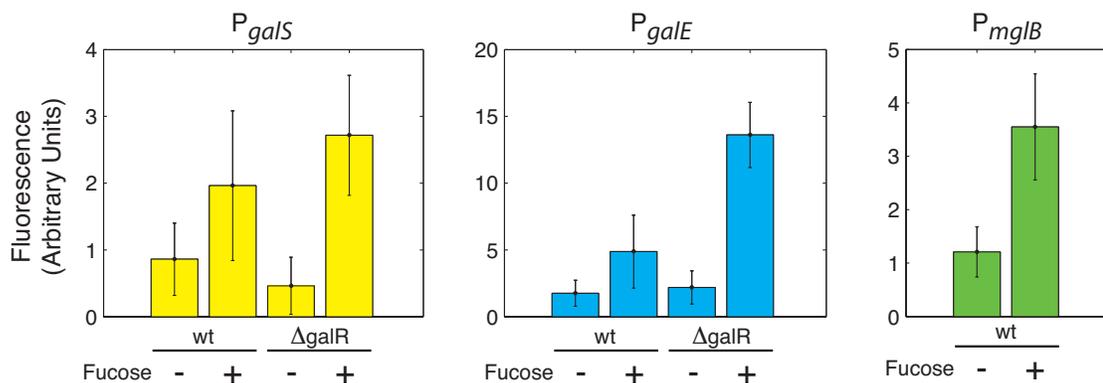


Figure 5.17: Expression levels of reporters for P_{galS} , P_{galE} , and P_{mglB} under different induction and genetic conditions. Data are background subtracted. Error bars show standard deviation over all the individual cells measured in snapshots (typically 100-200 cells). + Fucose is 10 mM, all cells are grown with mannose as a sugar source.

5.5.2 Methods

Promoter regions for $galS$, $galE$, and $mglB$ were taken from plasmids in Alon Zaslaver’s reporter library [65]. Promoter-fluorescent protein fusions were made with fusion PCR and verified by sequencing. The fusion PCR product was cloned into a vector with the kanamycin resistance marker and a low copy (SC101) origin of replication (pZS2 [42]). The region from kanamycin through the terminators following CFP was amplified using PCR with homology arms for $intC$

H1: 5′ – CCGTAGATTTACAGTTCGTCATGGTTCGCTTCAGATCGTTGACAGCCGCA – 3′

H2: 5′ – ATAGTTGTTAAGGTCGCTCACTCCACCTTCTCATCAAGCCAGTCCGCCCA – 3′,

and integrated into the MG1655 chromosome using recombineering [44].

$galR$ was deleted from the MG1655 strain with chromosomally integrated P_{galS} -YFP/ P_{galE} -CFP. The chloramphenicol marker from pKD3 was amplified using the PCR primers described in [67] (P1: 5′ – GTGTAGGCTGGAGCTGCTTC – 3′, P2: 5′ – ATGGGAATTAGCCATGGTCC – 3′) with homology arms for $galR$ deletion from [68]:

H1: 5′ – TCCGTAACACTGAAAGAATGTAAGCGTTTACCCACTAAGGTATTTTCATG – 3′

H2: 5′ – TACTGGCGCTGGAATTGCTTTAACTGCGGTTAGTCGCTGGTTGCATGATG – 3′.

Cells were grown overnight in MO (M9 salts with 1 mM $MgSO_4$, 0.1 mM $MgCl_2$, and 30 μ g/ml kanamycin) supplemented with 0.4% (w/v) glucose, 0.5% (v/v) glycerol, and 0.1% (w/v) Casamino

acids (called MON in [63]). Cultures were diluted back 1:50 in MO + 0.8% mannose (and 10 mM D-fucose, where applicable). After reaching OD 0.1–0.2, cells were further diluted and placed on a pad made of the same MO + 0.8% mannose (and 10 mM D-fucose) media as the original dilution. Cells were grown and imaged at 37°C.

Image acquisition and analysis methods were identical to those described in Chapter 3.

Chapter 6

Conclusion

6.1 Summary of Contributions

This thesis explores the idea of using noise as a tool for understanding gene regulation. Traditionally, cellular noise was considered detrimental because it introduces variability that may prevent a biological circuit from operating in a predictable fashion. As a result, most work focused on understanding how cells mitigate the effect of noise [69, 70, 71, 72]. However, recent work has shown that cells may use noise to their advantage: *B. subtilis* stochastically differentiate into cellular states that allow it to resist conditions where it would otherwise be deprived of nutrients [8]. Viruses use noise to create a bimodal population with two phenotypically distinct sub-populations as a mechanism to evade the host's response [73]. Yeast stress response genes are noisier than general housekeeping genes, allowing a subset of the population to react quickly to environmental stresses [74]. And theoretical work has shown that under certain conditions noisy signals may actually make the response of a gene network highly sensitive [75]. Clearly cells must have mechanisms for dealing with noise and there are many examples where stochastic properties have been exploited.

In this thesis we use noise as a tool for system identification, learning about the time scales, strength, and activity of gene regulation. Cellular noise has not been considered as an engineering tool before, but given its ubiquitous nature in biological systems it would be useful to exploit noise for the purpose of learning about gene regulation. Because the noise occurs naturally it can act as a minimally invasive form of perturbation, and allows for accurate measurements of the cellular conditions that the organism normally experiences.

Here we have shown that noise can be used to measure whether regulation is active, and its temporal properties, using the cross correlation function. We have developed a theoretical framework for calculating the cross correlation between two noisy signals and have used this to predict the properties of experimental systems, both synthetic and natural. We showed that a commonly occurring pattern of regulation, the feed-forward loop, can exhibit a variety of temporal responses that are dependent upon specific model parameters and cellular conditions. In two naturally occurring

feed-forward loops, it was shown that neither is actively regulating its target.

6.2 Future Work

6.2.1 Biological Persistence of Excitation

Engineering work in system identification has proven that to properly identify parameters, a system must be driven by a signal that is sufficiently rich [76]. Without perturbations there are properties of the system's behavior that may go undiscovered; with sufficiently rich inputs the full range of the dynamics are explored. Work on persistence of excitation has shown that for each system parameter to be identified there must be at least one unique frequency in the driving input [77]. White noise, because it excites at all frequencies, has been used as an input [76]. Using noise internal to the system has not been explored in an engineering context, but would be of great utility in understanding biological systems, where significant levels of noise are commonplace.

In this work we do not have direct control over the noisy input to the system and can only measure filtered versions of it. In addition, the frequency of these measurements is limited by practical experimental considerations. It would be interesting to develop a theoretical framework for understanding which model parameters can be identified given realistic sources of biological noise. Given the success of the linearized model, we could start with work on persistence of excitation in linear systems and develop a theory to describe when biological model parameters can be identified.

6.2.2 Network Identifiability

Work in control theory has developed methods for determining when systems are controllable and observable [78]. It would be useful to extend these ideas to a theory of network identifiability. Given a set of measurements and a network structure, an identifiable network is one where the measurements can uniquely determine network parameters. Such a theory would clarify which signals are important to measure and how many measurements are needed.

6.2.3 Monitoring Dynamic Changes

Measurements in this thesis were conducted in conditions where cells had equilibrated to their surrounding environment. It would be interesting to extend these methods to dynamic environments where protein levels are either switching between states, or exhibiting more complex behavior, such as oscillations. In principle, noise-based inference methods can still be applied to dynamically changing systems, though certain approximations about linearized systems may no longer be valid. Experimentally monitoring dynamic behavior may be challenging, too, since the noise-based method relies heavily on averaging across many sets of time-series data.

It may not be necessary to use noise as the perturbing force if the system is already changing dynamically. For example, a system that oscillates may explore enough regions of the state space that noise is not required as a perturbing force.

6.2.4 External Inputs for System Identification

Recent work in microfluidics has made it possible to monitor individual cells while exposing them to time-varying chemical stimuli [79, 80]. This is a more direct way of perturbing the system and measuring its response. Although the cells may not be exposed to purely natural signals, the response properties can be characterized more thoroughly. In addition, applying a known, prescribed input can be an efficient way of learning about the response of a system.

6.2.5 Context-Sensitive Maps of Gene Regulation

Databases of gene regulatory interactions are a good source of information about gene network topology, but lack information about the context in which regulatory elements are active. It will be interesting to explore other examples of regulation to learn when they are active. Although this presents an additional layer of information, the ultimate goal of reducing network structures to the parts that are actually active has the potential to significantly simplify network analysis.

Appendix A

Mathematica Code for Calculating Cross Correlations of a Cascade

```

Clear[ps, NN]
NN = 8; //cascade length, don't set less than 2
ps = {};
For[n = 2, n <= NN, n++,
Clear[fint, fext, z, w, t, p1, p2, RextL, RextG, RintL, RintG, fx1x1, fxn, Rx1x1, Rxn];
Clear[p1, p2];
gval = 0.02;
g = {-gval};
For[ii = 2, ii < n, ii++, gn = {gval}; g = Join[g, gn]];
g = Join[g, {1}];
Wi = 1; We = 0.005;
θ = 1; λ = 1;
β = Log[2]/60.0;
k = Log[2]/5.0;
fint[z_]:=1/((β+iz)^n(β-iz))1/(k+iz)(k-iz) (∏_{j=1}^n g[[j]]) λ^2Wie^{-izt};
fext[z_]:=∑_{i=1}^n 1/((β+iz)^{n-i+2}(β-iz)^2) (∏_{j=i}^n g[[j]]) θ^2Wee^{-izt};
RintG[t_]:= -i(Residue[fint[z], {z, -iβ}] + Residue[fint[z], {z, -ik}]);
RintL[t_]:=i(Residue[fint[z], {z, iβ}] + Residue[fint[z], {z, ik}]);
RextG[t_]:= -i(Residue[fext[z], {z, -iβ}] + Residue[fext[z], {z, -ik}]);
RextL[t_]:=i(Residue[fext[z], {z, iβ}] + Residue[fext[z], {z, ik}]);
RiG[t_]:=RintG[t]//FullSimplify;
RiL[t_]:=RintL[t]//FullSimplify;
ReG[t_]:=RextG[t]//FullSimplify;
ReL[t_]:=RextL[t]//FullSimplify;

```

```

fx1x1[z_]:=1/(beta-iz)1/(beta+iz) ( (theta^2We)/((beta-iz)(beta+iz)) + (lambda^2Wi)/((k-iz)(k+iz)) ) e^{-iz*tau};
fxnzn[z_]:=Sum_{i=1}^n (1/(beta-iz)1/(beta+iz))^{n-i+1} (Product_{j=i}^n g[[j]])^2 ( (theta^2We)/((beta-iz)(beta+iz)) + (lambda^2Wi)/((k-iz)(k+iz)) ) e^{-iz*tau};
Rx1x1 = i(Residue[fx1x1[z], {z, i*beta}] + Residue[fx1x1[z], {z, i*k}]);
Rxnzn = i(Residue[fxnzn[z], {z, i*beta}] + Residue[fxnzn[z], {z, i*k}]);
tau = 0;
normval = 1;
If[Im[Sqrt[Rx1x1 * Rxnzn]] < 0.0001, normval = Re[Sqrt[Rx1x1 * Rxnzn]]];
Print[normval];
p1 = Plot[(ReG[t] + RiG[t])/normval, {t, 0, 1000}, PlotStyle -> {Hue[n/NN]}];
p2 = Plot[(ReL[t] + RiL[t])/normval, {t, -1000, 0}, PlotStyle -> {Hue[n/NN]}];
ps = Join[ps, {p1, p2}];
]
Show[ps, PlotRange -> {{-1000, 300}, {-1, 1}}]

```

Appendix B

Synthetic Circuit Movie Frames

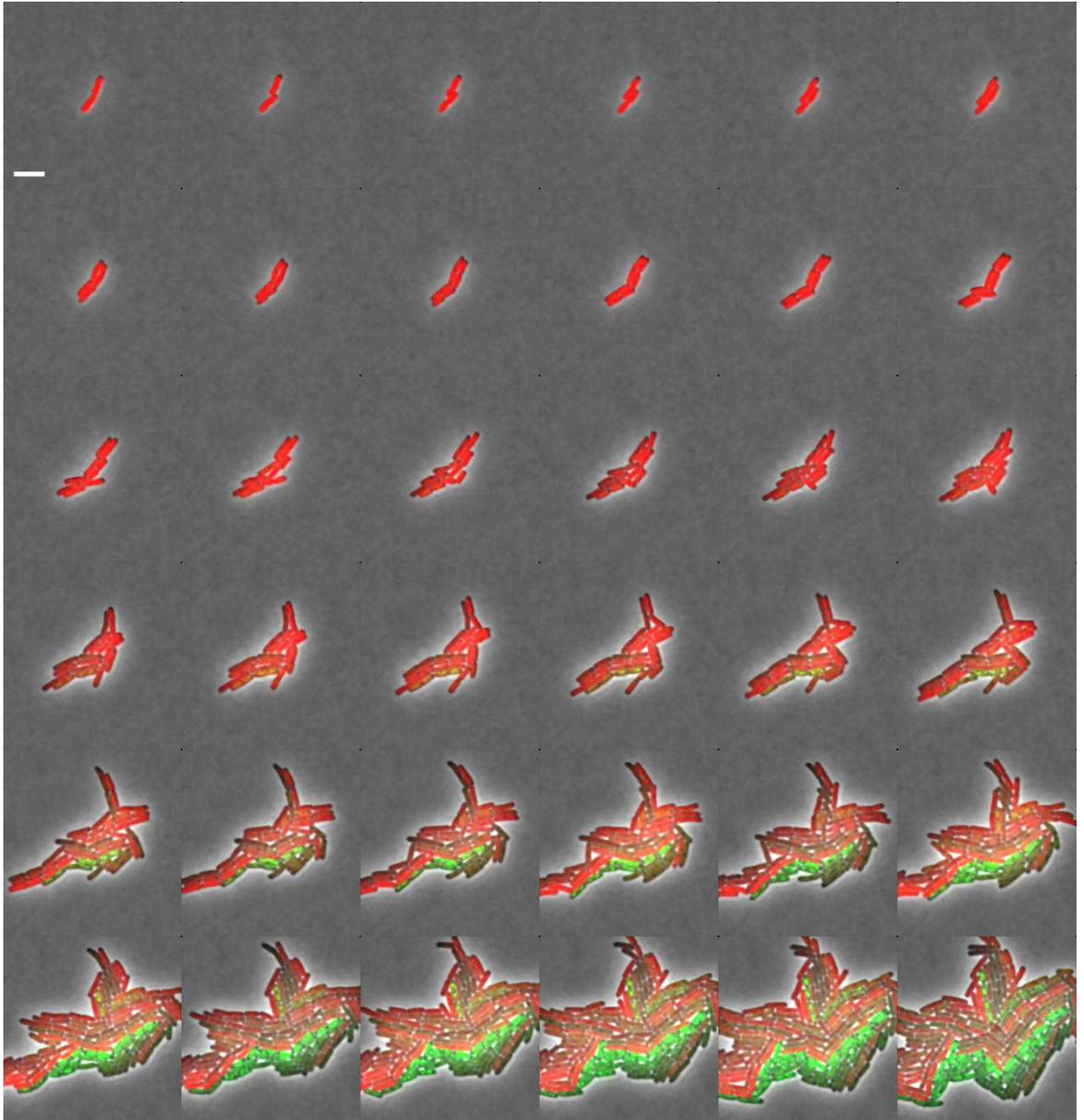


Figure B.1: Chromosome: YFP and RFP. 20 minutes between frames, starting with the first frame at 10 minutes. All labeling and notation is the same as in Fig. 3.5.

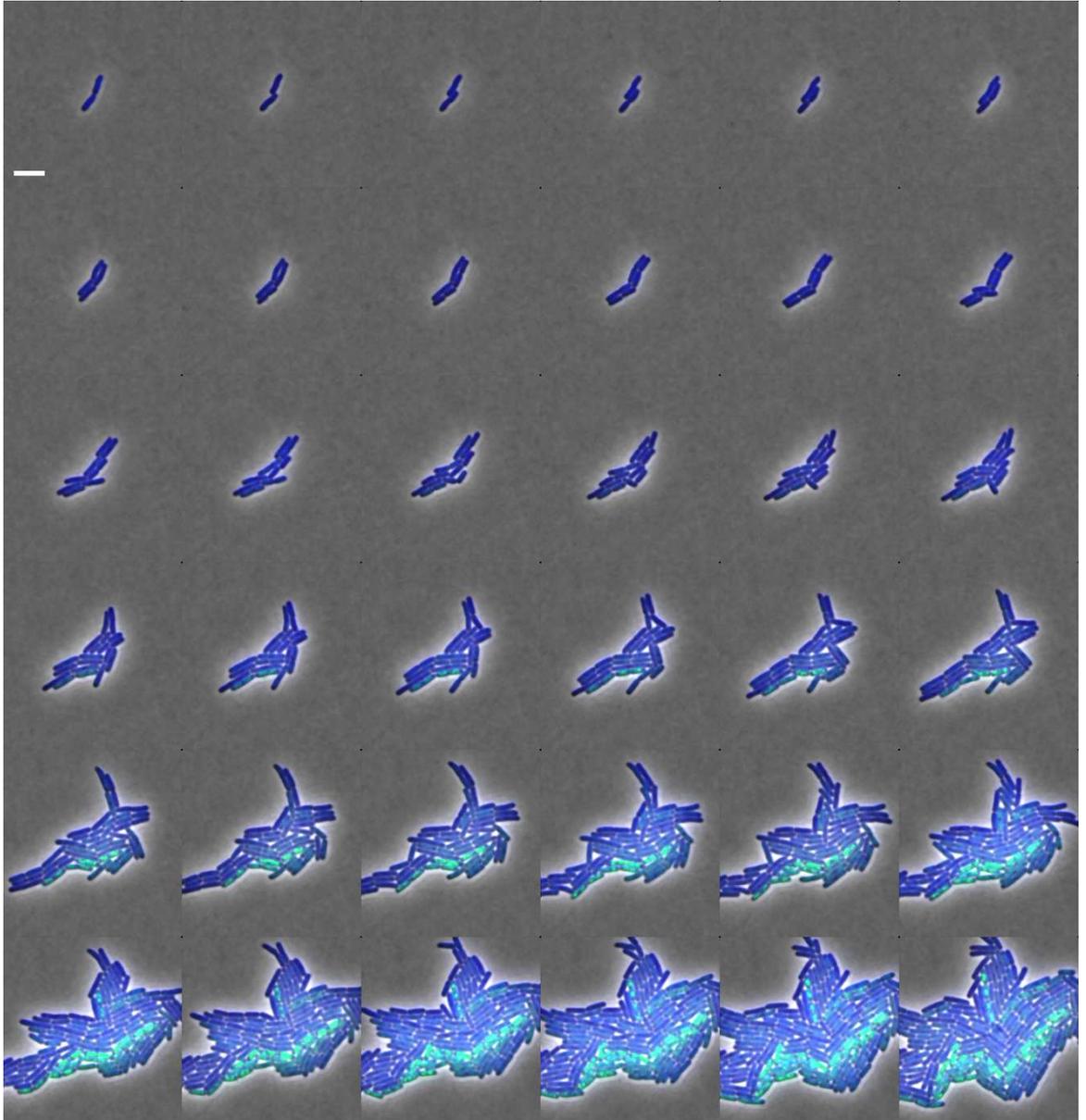


Figure B.2: Chromosome: YFP and CFP

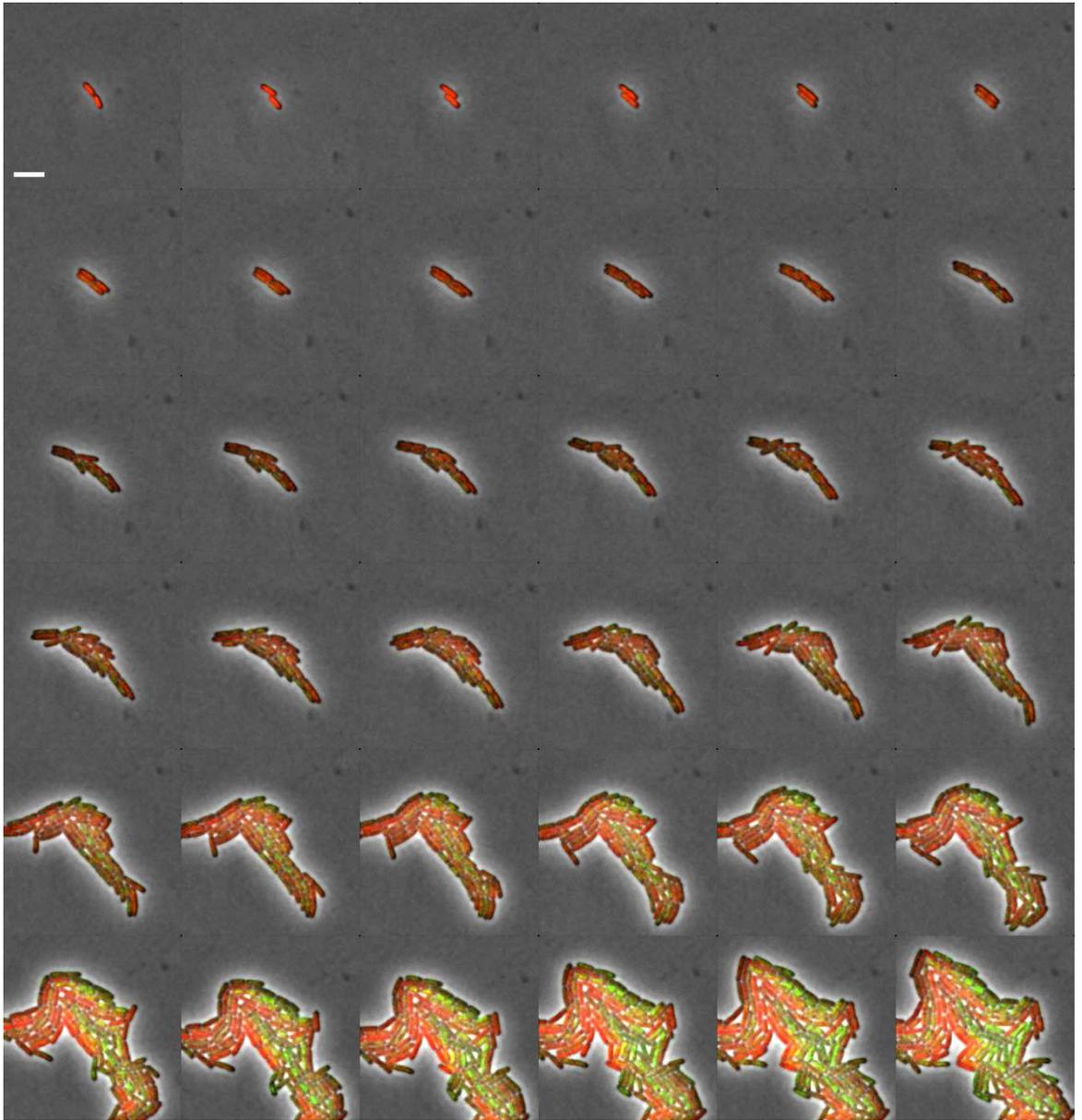


Figure B.3: Plasmid: YFP and RFP

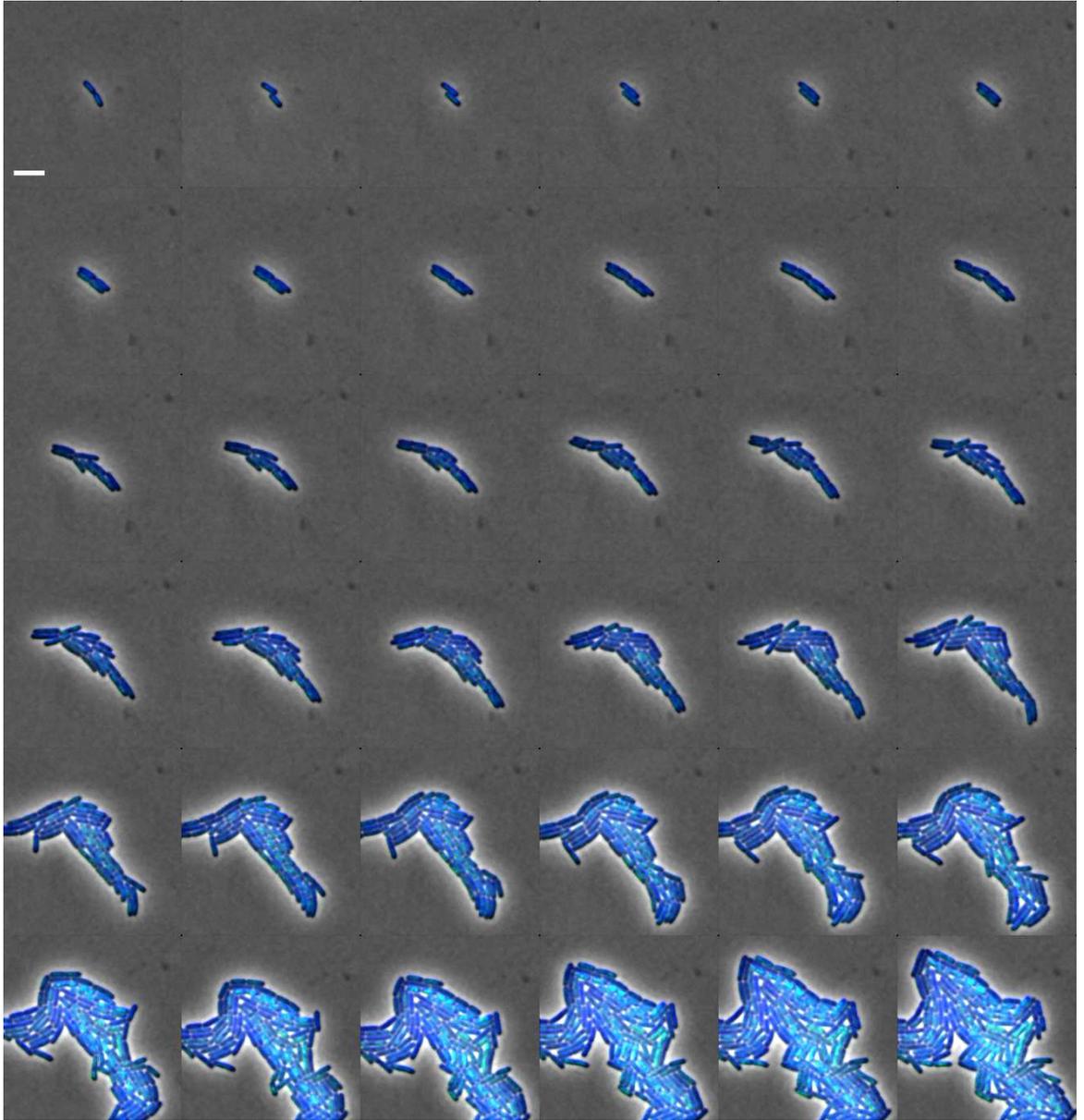


Figure B.4: Plasmid: YFP and CFP

Appendix C

Full Table of Feed-Forward Loop Cluster Percentages

			Signal	Signal	Signal																Entropy
X-YY-ZX-Z	X-Y	Y-Z	X-Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O			
act act act	+	+	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act act	+	+	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act act	+	+	-	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act act	+	0	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act act	+	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act act	+	0	-	0	0.11	0	0	0	0.56	0	0	0	0.06	0	0.28	0	0	0	0	1.57	
act act act	+	-	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act act	+	-	0	0	0.11	0	0	0	0.56	0	0	0	0.06	0	0.28	0	0	0	0	1.57	
act act act	+	-	-	0	0.11	0	0	0	0.56	0	0	0	0.06	0	0.28	0	0	0	0	1.57	
act act act	0	+	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act act	0	+	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act act	0	+	-	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act act	0	0	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act rep	+	+	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act rep	+	+	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act rep	+	+	-	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act rep	+	0	+	0	0	0.89	0	0	0	0	0	0	0.09	0	0	0	0	0	0.02	0.59	
act act rep	+	0	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act rep	+	0	-	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act rep	+	-	+	0	0.06	0	0	0	0.61	0	0	0	0.06	0	0.28	0	0	0	0	1.41	
act act rep	+	-	0	0	0.11	0	0	0	0.56	0	0	0	0.06	0	0.28	0	0	0	0	1.57	
act act rep	+	-	-	0	0.11	0	0	0	0.56	0	0	0	0.06	0	0.28	0	0	0	0	1.57	
act act rep	0	+	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act rep	0	+	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act rep	0	+	-	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
act act rep	0	0	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Table C.1: Percentage of Cluster Types Exhibited By Circuit Configurations (Page 1 of 4)

SignalSignalSignal				A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Entropy	
X-YY-ZX-Z	X-Y	Y-Z	X-Z																	
act rep act	+	+	+	0	0	0.80	0	0	0	0	0	0	0.09	0	0	0	0	0.11	0	0.93
act rep act	+	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep act	+	+	-	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep act	+	0	+	0	0	0.80	0	0	0	0	0	0	0.09	0	0	0	0	0.11	0	0.93
act rep act	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep act	+	0	-	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep act	+	-	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep act	+	-	0	0	0.11	0	0	0.06	0	0	0	0	0	0	0.50	0	0.33	0	0	1.61
act rep act	+	-	-	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep act	0	+	+	0	0.01	0.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05
act rep act	0	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep act	0	+	-	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep act	0	0	+	0	0.01	0.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05
act rep rep	+	+	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep rep	+	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep rep	+	+	-	0.41	0	0.51	0.02	0	0	0	0	0	0.01	0	0	0	0	0.06	0	1.41
act rep rep	+	0	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep rep	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep rep	+	0	-	0.38	0	0.51	0.03	0	0	0	0	0	0.01	0	0	0	0	0.07	0	1.49
act rep rep	+	-	+	0.90	0	0	0	0.09	0	0	0	0	0	0	0	0	0.01	0	0	0.52
act rep rep	+	-	0	0	0.11	0	0	0.06	0	0	0	0	0	0	0.50	0	0.33	0	0	1.61
act rep rep	+	-	-	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep rep	0	+	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep rep	0	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
act rep rep	0	+	-	0.48	0	0.52	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00
act rep rep	0	0	+	0.99	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05

Table C.2: Percentage of Cluster Types Exhibited By Circuit Configurations (Page 2 of 4)

SignalSignalSignal				A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Entropy
X-YY-ZX-Z	X-Y	Y-Z	X-Z																
rep act act	+	+	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	+	+	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	+	+	-	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	+	0	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	+	0	0	0.98	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0.17
rep act act	+	0	-	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	+	-	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	+	-	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	+	-	-	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	0	+	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	0	+	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	0	+	-	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act act	0	0	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	+	+	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	+	+	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	+	+	-	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	+	0	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	+	0	0	0.98	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0.13
rep act rep	+	0	-	0.29	0	0.50	0.02	0	0	0	0	0.05	0	0	0	0	0.14	0	1.74
rep act rep	+	-	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	+	-	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	+	-	-	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	0	+	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	0	+	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	0	+	-	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0
rep act rep	0	0	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table C.3: Percentage of Cluster Types Exhibited By Circuit Configurations (Page 3 of 4)

SignalSignalSignal				A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Entropy	
X-YY-ZX-Z	X-Y	Y-Z	X-Z																	
rep rep act	+	+	+	0	0.01	0.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05
rep rep act	+	+	0	0.96	0	0	0	0	0	0	0.04	0	0	0	0	0	0	0	0	0.23
rep rep act	+	+	-	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep act	+	0	+	0	0.01	0.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05
rep rep act	+	0	0	0	0.02	0.98	0	0	0	0	0	0	0	0	0	0	0	0	0	0.17
rep rep act	+	0	-	0.94	0.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.31
rep rep act	+	-	+	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep act	+	-	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep act	+	-	-	0.94	0.06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.31
rep rep act	0	+	+	0	0.01	0.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05
rep rep act	0	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep act	0	+	-	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep act	0	0	+	0	0.01	0.99	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05
rep rep rep	+	+	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep rep	+	+	0	0.96	0	0	0	0	0	0	0.04	0	0	0	0	0	0	0	0	0.23
rep rep rep	+	+	-	0.47	0	0.51	0	0	0	0	0.02	0	0	0	0	0	0	0	0.01	1.16
rep rep rep	+	0	+	0.52	0.03	0.28	0	0	0.02	0	0	0	0	0	0.01	0	0	0	0.14	1.74
rep rep rep	+	0	0	0	0.02	0.98	0	0	0	0	0	0	0	0	0	0	0	0	0	0.13
rep rep rep	+	0	-	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep rep	+	-	+	0.48	0	0.50	0	0	0	0	0	0	0	0	0	0	0	0	0.02	1.11
rep rep rep	+	-	0	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep rep	+	-	-	0	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep rep	0	+	+	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep rep	0	+	0	1.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rep rep rep	0	+	-	0.49	0	0.51	0	0	0	0	0	0	0	0	0	0	0	0	0	1.00
rep rep rep	0	0	+	0.99	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05

Table C.4: Percentage of Cluster Types Exhibited By Circuit Configurations (Page 4 of 4)

Bibliography

- [1] B. Alberts, D. Bray, K. Hopkin, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Essential Cell Biology*. Garland Science, 2004.
- [2] A. Nakabachi, A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, N. A. Moran, and M. Hattori. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science*, 314(5797):267–267, October 2006.
- [3] E. Pennisi. Genetics—working the (gene count) numbers: Finally, a firm answer? *Science*, 316(5828):1113, 2007.
- [4] A. Becskei and L. Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–593, June 2000.
- [5] N. Rosenfeld, M. B. Elowitz, and U. Alon. Negative autoregulation speeds the response times of transcription networks. *Journal Of Molecular Biology*, 323(5):785–793, November 2002.
- [6] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Accurate prediction of gene feedback circuit behavior from component properties. *Molecular Systems Biology*, 3:143, November 2007.
- [7] S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A. M. Huerta, L. Trevino-Quintanilla, and J. Collado-Vides. RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic Acids Research*, 36:D120–D124, January 2008.
- [8] G. M. Suel, J. Garcia-Ojalvo, L. M. Liberman, and M. B. Elowitz. An excitable gene regulatory circuit induces transient cellular differentiation. *Nature*, 440(7083):545–550, March 2006.
- [9] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, October 2002.

- [10] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, March 2004.
- [11] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1):64–68, May 2002.
- [12] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 100(21):11980–11985, October 2003.
- [13] M. Isalan, C. Lemerle, K. Michalodimitrakis, C. Horn, P. Beltrao, E. Raineri, M. Garriga-Canut, and L. Serrano. Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, 452(7189):840–U2, April 2008.
- [14] G. Tkacik, C. G. Callan Jr, and W. Bialek. Information flow and optimization in transcriptional control. arXiv:0705.0313v1, May 2007.
- [15] F. Toledo and G. M. Wahl. Regulating the p53 pathway: *in vitro* hypotheses, *in vivo* veritas. *Nature Reviews Cancer*, 6(12):909–923, December 2006.
- [16] P. J. Piggot and D. W. Hilbert. Sporulation of *Bacillus subtilis*. *Current Opinion In Microbiology*, 7(6):579–586, December 2004.
- [17] S. D. M. Santos, P. J. Verveer, and P. I. H. Bastiaens. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nature Cell Biology*, 9(3), March 2007.
- [18] J. M. Pedraza and A. van Oudenaarden. Noise propagation in gene networks. *Science*, 307(5717):1965–1969, March 2005.
- [19] 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, March 2005.
- [20] J. M. Raser and E. K. O’Shea. Noise in gene expression: Origins, consequences, and control. *Science*, 309(5743):2010–2013, September 2005.
- [21] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O’Shea, Y. Pilpel, and N. Barkai. Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643, June 2006.
- [22] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature*, 441(7095):840–846, June 2006.

- [23] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, August 2002.
- [24] A. Arkin and J. Ross. Statistical construction of chemical-reaction mechanisms from measured time-series. *Journal Of Physical Chemistry*, 99(3):970–979, January 1995.
- [25] A. Arkin, P. D. Shen, and J. Ross. A test case of correlation metric construction of a reaction pathway from measurements. *Science*, 277(5330):1275–1279, August 1997.
- [26] W. A. Schmitt, R. M. Raab, and G. Stephanopoulos. Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Research*, 14(8):1654–1663, August 2004.
- [27] N. Friedman, M. Linial, I. Nachman, and D. Pe’er. Using Bayesian networks to analyze expression data. *Journal Of Computational Biology*, 7(3-4):601–620, 2000.
- [28] 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, July 2003.
- [29] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, G. 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 Biology*, 5:54–66, 2007.
- [30] F. Geier, J. Timmer, and C. Fleck. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Systems Biology*, 2007.
- [31] E. Sontag, A. Kiyatkin, and B. N. Kholodenko. Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data. *Bioinformatics*, 20(12):1877–1886, August 2004.
- [32] M. J. L De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. In *Pacific Symposium on Biocomputing*, 2003.
- [33] W. Vance, A. Arkin, and J. Ross. Determination of causal connectivities of species in reaction networks. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 99(9):5816–5821, April 2002.
- [34] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC, 2007.
- [35] J.D. Murray. *Mathematical Biology I*. Springer Verlag, 3rd edition, 2002.

- [36] A. Sigal, R. Milo, A. Cohen, N. Geva-Zatorsky, Y. Klein, Y. Liron, N. Rosenfeld, T. Danon, N. Perzov, and U. Alon. Variability and memory of protein levels in human cells. *Nature*, 444(7119):643–646, November 2006.
- [37] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the Brownian motion. *Physical Review*, 36(5):0823–0841, September 1930.
- [38] D. T. Gillespie. Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral. *Physical Review E*, 54(2):2084–2091, August 1996.
- [39] S. Hooshangi, S. Thiberge, and R. Weiss. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 102(10):3581–3586, March 2005.
- [40] M. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338, 2000.
- [41] B. J. Meyer, R. Maurer, and M. Ptashne. Gene-regulation at the right operator (O_R) of bacteriophage-lambda II O_{R1} , O_{R2} , and O_{R3} —their roles in mediating the effects of repressor and cro. *Journal Of Molecular Biology*, 139(2):163–194, 1980.
- [42] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/ I_1 - I_2 regulatory elements. *Nucleic Acids Research*, 25(6):1203–1210, March 1997.
- [43] Y. Ohashi, A. Yamashiro, T. Washio, N. Ishii, H. Ohshima, T. Michishita, M. Tomita, and M. Itaya. In silico diagnosis of inherently inhibited gene expression focusing on initial codon combinations. *Gene*, 347(1):11–19, February 2005.
- [44] D. G. Yu, H. M. Ellis, E. C. Lee, N. A. Jenkins, N. G. Copeland, and D. L. Court. An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 97(11):5978–5983, May 2000.
- [45] R. Schleif. Regulation of the L-arabinose operon of *Escherichia coli*. *Trends in Genetics*, 16:559–565, 2000.
- [46] M. Wall, W. Hlavacek, and M. Savageau. Design of gene circuits: Lessons from bacteria. *Nature Reviews Genetics*, 5:34–42, 2004.
- [47] D.S. Hochbaum and D.B. Shmoys. A best possible approximation algorithm for the k-center problem. *Mathematics of Operations Research*, 10:180–184, 1985.

- [48] M. Wall, A. Rechtsteiner, and L. Rocha. *A Practical Approach to Microarray Data Analysis*. D. P. Berrar, W. Dubitzky, and M. Granzow, eds., pp. 91–109. Kluwer, Norwell, MA, 2003.
- [49] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403(6767):339–342, January 2000.
- [50] M. R. Atkinson, M. A. Savageau, J. T. Myers, and A. J. Ninfa. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. *Cell*, 113(5):597–607, May 2003.
- [51] M.A. Savageau. Design of molecular control mechanisms and the demand for gene expression. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 74:5647–5651, 1977.
- [52] S. Semsey, S. Krishna, K. Sneppen, and S. Adhya. Signal integration in the galactose network of *Escherichia coli*. *Molecular Microbiology*, 65(2):465–476, July 2007.
- [53] M. J. Weickert and S. Adhya. The galactose regulon of *Escherichia coli*. *Molecular Microbiology*, 10(2):245–251, October 1993.
- [54] S. Chatterjee, Y. N. Zhou, S. Roy, and S. Adhya. Interaction of gal repressor with inducer and operator: Induction of gal transcription from repressor-bound DNA. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 94(7):2957–2962, April 1997.
- [55] M. P. Brown, N. Shaikh, M. Brenowitz, and L. Brand. The allosteric interaction between D-galactose and the *Escherichia coli* galactose repressor protein. *Journal Of Biological Chemistry*, 269(17):12600–12605, April 1994.
- [56] M. Geanacopoulos and S. Adhya. Functional characterization of roles of GalR and GalS as regulators of the gal regulon. *Journal Of Bacteriology*, 179(1):228–234, January 1997.
- [57] M. J. Weickert and S. Adhya. Control of transcription of gal repressor and isorepressor genes in *Escherichia coli*. *Journal Of Bacteriology*, 175(1):251–258, January 1993.
- [58] H. E. Choy and S. Adhya. Control of gal transcription through DNA looping—inhibition of the initial transcribing complex. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 89(23):11264–11268, 1992.
- [59] S. Chatterjee, K. Ghosh, A. Dhar, and S. Roy. Ligand specificity and ligand-induced conformational change in gal repressor. *Proteins-Structure Function And Genetics*, 49(4):554–559, December 2002.

- [60] T. Aki, H. E. Choy, and S. Adhya. Histone-like protein HU as a specific transcriptional regulator: co-factor role in repression of *gal* transcription by GAL repressor. *Genes To Cells*, 1(2):179–188, 1996.
- [61] K. Virnik, Y. L. Lyubchenko, M. A. Karymov, P. Dahlgren, M. Y. Tolstorukov, S. Semsey, V. B. Zhurkin, and S. Adhya. “Antiparallel” DNA loop in gal repressosome visualized by atomic force microscopy. *Journal Of Molecular Biology*, 334(1):53–63, November 2003.
- [62] M. J. Weickert and S. Adhya. Isorepressor of the gal regulon in *Escherichia coli*. *Journal Of Molecular Biology*, 226(1):69–83, 1992.
- [63] S. Mangan, S. Itzkovitz, A. Zaslaver, and U. Alon. The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*. *Journal Of Molecular Biology*, 356(5):1073–1081, March 2006.
- [64] M. E. Wall, M. J. Dunlop, and W. S. Hlavacek. Multiple functions of a feed-forward-loop gene circuit. *Journal Of Molecular Biology*, 349(3):501–514, June 2005.
- [65] A. Zaslaver, A. Bren, M. Ronen, S. Itzkovitz, I. Kikoin, S. Shavit, W. Liebermeister, M. G. Surette, and U. Alon. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nature Methods*, 3:623–628, 2006.
- [66] S. Kaplan, A. Bren, A. Zaslaver, E. Dekel, and U. Alon. Diverse two-dimensional input functions control bacterial sugar genes. *Molecular Cell*, 29(6):786–792, 2008.
- [67] K. A. Datsenko and B. L. Wanner. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 97(12):6640–6645, 2000.
- [68] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, 2006.
- [69] 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, June 2005.
- [70] T. S. Gardner and J. J. Collins. Gene regulation—neutralizing noise in gene networks. *Nature*, 405(6786):520–521, June 2000.
- [71] S. Hooshangi and R. Weiss. The effect of negative feedback on noise propagation in transcriptional gene networks. *Chaos*, 16(2):026108, June 2006.

- [72] J. M. G. Vilar, H. Y. Kueh, N. Barkai, and S. Leibler. Mechanisms of noise-resistance in genetic oscillators. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 99(9):5988–5992, April 2002.
- [73] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics*, 149(4):1633–1648, August 1998.
- [74] W. J. Blake, G. Balazsi, M. A. Kohanski, F. J. Isaacs, K. F. Murphy, Y. Kuang, C. R. Cantor, D. R. Walt, and J. J. Collins. Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular Cell*, 24(6):853–865, December 2006.
- [75] J. Paulsson, O. G. Berg, and M. Ehrenberg. Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 97(13):7148–7153, June 2000.
- [76] L. Ljung. *System Identification: Theory for the User*. Prentice Hall, Upper Saddle River, N.J., 1987.
- [77] S. Boyd and S. Sastry. On parameter convergence in adaptive-control. *Systems & Control Letters*, 3(6):311–319, 1983.
- [78] K. J. Astrom and R. M. Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2008.
- [79] J. T. Mettetal, D. Muzzey, C. Gomez-Urbe, and A. van Oudenaarden. The frequency dependence of osmo-adaptation in *Saccharomyces cerevisiae*. *Science*, 319(5862):482–484, January 2008.
- [80] F. K. Balagadde, L. C. You, C. L. Hansen, F. H. Arnold, and S. R. Quake. Long-term monitoring of bacteria undergoing programmed population control in a microchemostat. *Science*, 309(5731):137–140, July 2005.