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 of orders of magnitude, from the bacterium *C. ruddii* with 160 thousand base pairs of DNA to the ameoba *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

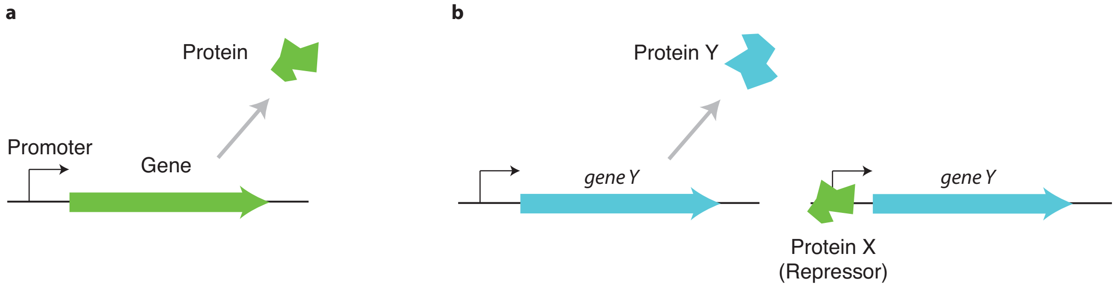

Figure 1.1: Control of gene expression. (a) This diagram depicts a small section of the DNA that is part of the larger genome. A promoter region upstream of the gene is used to control expression of the gene. When gene expression is ON, the gene is transcribed and then translated into a protein. (b) 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

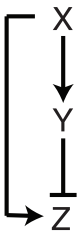

Figure 1.2: Feed-forward loop network motif. X and Y are transcription factors that control Z. Regular arrows indicate activation; T-shaped arrows indicate repression.

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?

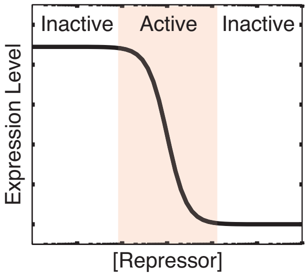

Figure 1.3: Target gene expression versus repressor concentration (schematic). Regions where changes in repressor concentration cause changes in target gene expression are active. When repressor levels saturate (right) or are insufficient to repress (left), then the link is inactive. Regulatory links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

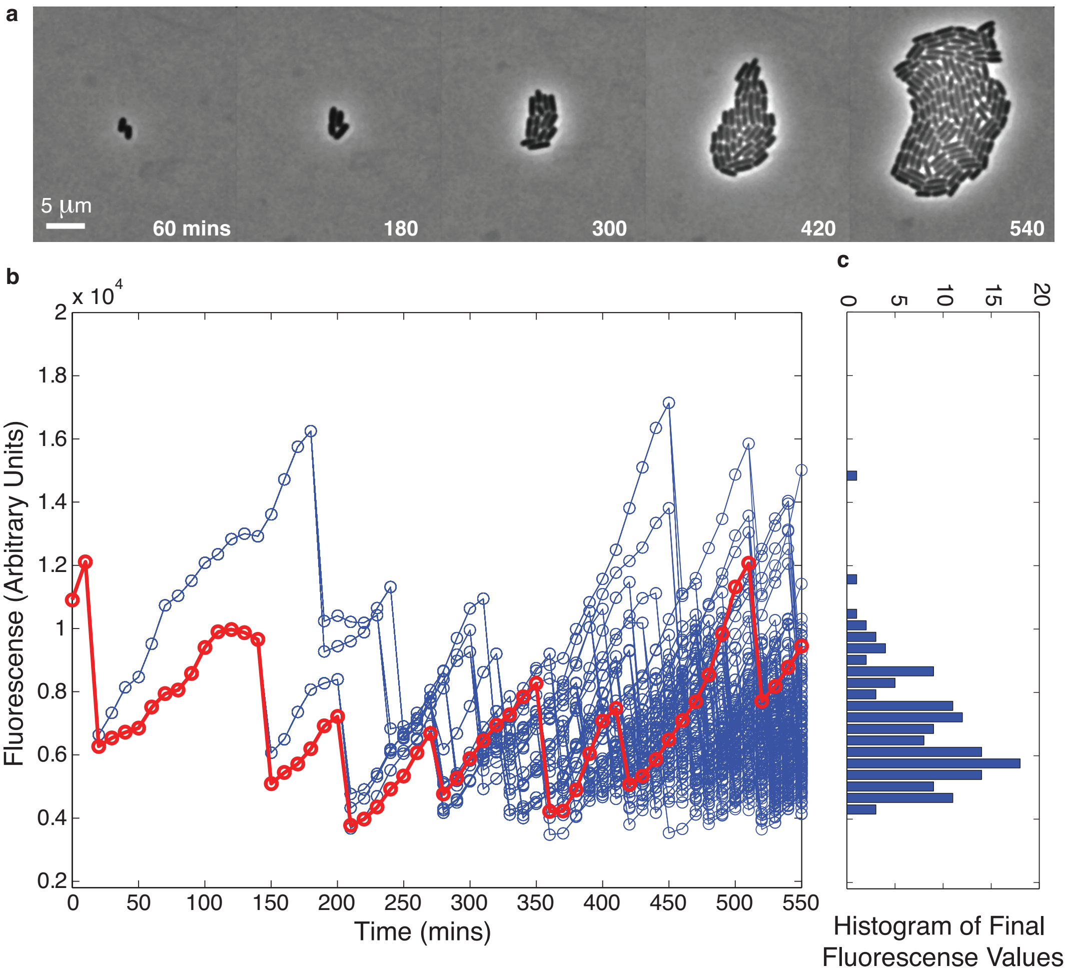

Figure 1.4: Cell-cell variability in *E. coli* gene expression. (a) Snapshots of single cells. The microcolony originated from a single cell. Time and length scales are indicated on the figure. (b) Fluorescent protein expression levels versus time. Blue lines are data for all cells. The red trace follows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

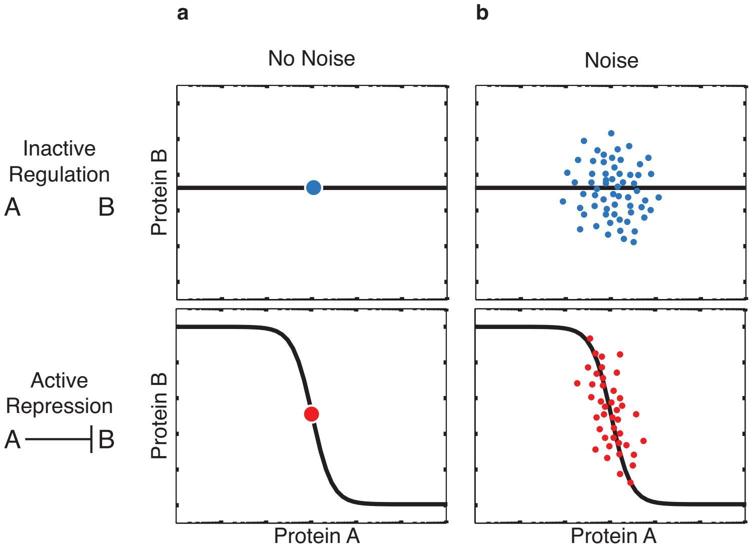

Figure 1.5: Correlations in noise distributions show active regulation. Each dot represents a single cell. Gene regulation function is shown as a solid black line. (a) Without noise, all cells show the same levels of expression. (b) With noise, correlations can suggest active regulation. T-shaped arrow 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.

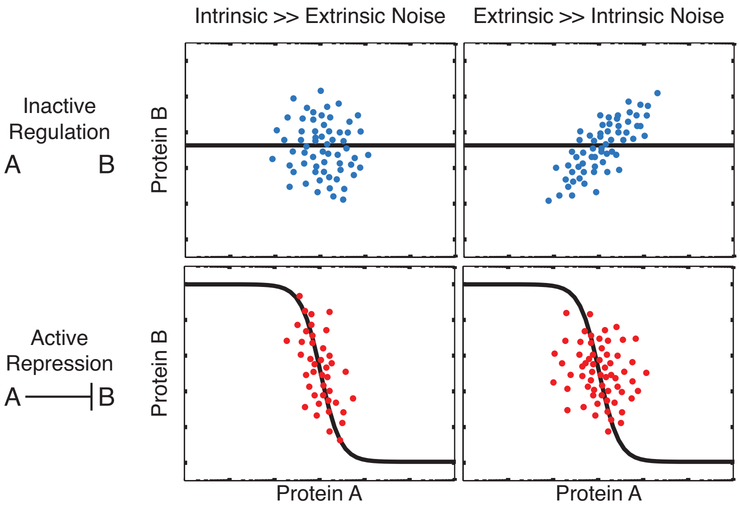

Figure 1.6: Noise can produce different types of static correlations. In each plot, dots represent individual cells from a hypothetical population. Top plots show correlations without an active regulatory link, while the bottom show correlations with active repression. We consider noise regimes in which 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.

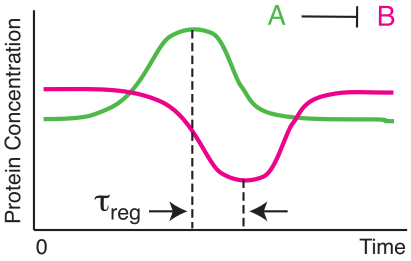

Figure 1.7: Temporal gene expression patterns for a repressor, A, (green line) and its target, B, (magenta line) are anti-correlated at a delay time denoted τ_{reg} (schematic).

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].

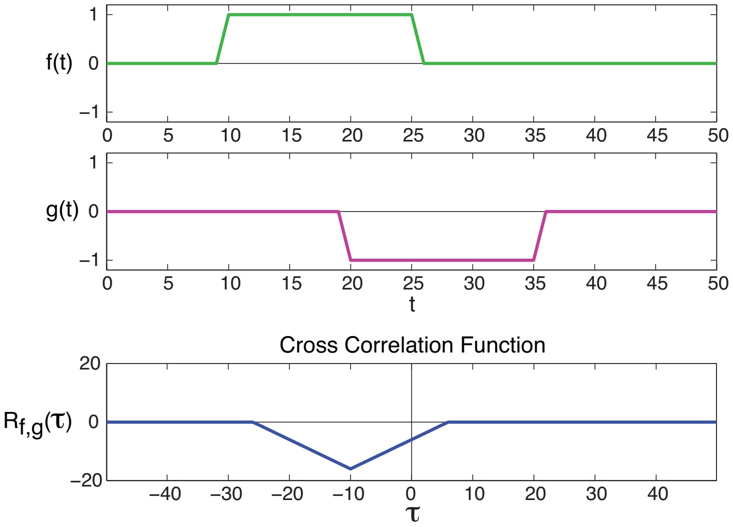

Figure 1.8: Two signals $f(t)$ and $g(t)$ are compared using the cross correlation function $R_{f,g}(\tau)$.

Infering 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.