Quantifying Gene Regulatory Networks

by

Shangying Wang

Department of Physics
Duke University

Date: _____

Approved:

_____
Sridhar Raghavachari, Co-Supervisor

_____
Nicolas Buchler, Co-Supervisor

_____
Joshua Socolar

_____
Kate Scholberg

_____
Anne West

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Physics
in the Graduate School of Duke University
2014

Abstract

Quantifying Gene Regulatory Networks

by

Shangying Wang

Department of Physics
Duke University

Date: _____

Approved:

_____
Sridhar Raghavachari, Co-Supervisor

_____
Nicolas Buchler, Co-Supervisor

_____
Joshua Socolar

_____
Kate Scholberg

_____
Anne West

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Physics in the Graduate School of Duke University
2014

Abstract

Transcription and translation describe the flow of genetic information from DNA to mRNA to protein. Recent studies show that at a single cell level, these processes are stochastic, which results in the variation of the number of mRNA and proteins even under identical environmental conditions. Because the number of mRNA and protein in each single cell are actually very small, these variations can be crucial for cellular function in diverse contexts, such as development, stress response, immunological and nervous system function. Most studies examine the origin and effects of stochastic gene expression using computer simulations. My goal is to develop a theoretical framework to study activity-dependent gene expression using simplified models that capture essential features.

I have examined the dynamics of stochastic gene regulation in three contexts. First, I examine how fluctuations in promoter accessibility lead to bursty transcription, during which genes are turned on or off stochastically. I describe a mathematical formalism to represent bursty gene expression in a coarse-grained manner as a Markov process and derive a master equation for the time evolution of the probability distribution of the number of mRNA molecules. This allows us to examine how transcript number responds to time varying stimuli. This model forms a basic building block for understanding the signal transmission and noise of the transcription process to time varying inputs as would be sensed by cells in dynamic environments. In addition to synthesis, gene expression is subject to additional modes of regulation.

One such mechanism that controls transcript numbers is by microRNAs (miRNAs), which pair with target mRNAs to repress protein production following transcription. Although hundreds of miRNAs have been identified in mammalian genomes, the function of miRNA-based repression in the context of gene regulation networks still remains unclear. I explore the functional roles of feedback regulation by miRNAs and show that protein fluctuations strongly depend on the mode of miRNA-mediated repression. I discuss the functional implications of protein fluctuations arising from miRNA-mediated repression on gene regulatory networks. Finally, I examine the impact of fluctuations on alternative splicing, which is a major source for proteomic complexity in higher eukaryotes. Although the proteins regulating alternative splicing have been extensively studied, little is known about how noise arising from the stochastic nature of alternative splicing contributes to the entire gene expression process. I explore the functional roles and noise properties of alternative splicing, focusing on the case of exon skipping and intron retention. I show that while the overall counts of the mRNAs of the two isoforms are independent and Poisson distributed, diffusion and binding of the splicing factors contributes to the variance in the abundance of the isoforms.

Noise in gene expression may be of particular relevance in the nervous system. Environmental stimuli drive the rapid remodeling of neural circuitry in part by inducing the activation of genes to make proteins that modify neuronal excitability and connectivity, ultimately influencing higher order brain function. Finally, I examine the implications of our studies for activity dependent gene expression in the nervous system.

To my parents and the memory of my grandparents.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

For Chapter 2

D_I	Gene in activated state.
D_A	Gene in inactivated state.
λ	Gene activation rate.
μ	Gene inactivation rate.
ν	transcription rate.
δ	mRNA degradation rate.
$\langle m \rangle$	average number of mRNA per cell.
F	Fano factor, which is defined as the variance over the mean. For Poisson process, $F = 1$.
$G(z, t)$	Moment generating function for the total mRNA number at time t.
$p_m(t)$	Probability there are m number of mRNA at time t per cell.

For Chapter 3

α_m	mRNA synthesis rate.
α_μ	miRNA synthesis rate.
k_p	translation rate of TF from mRNA template.
κ	rate miRNA binds to mRNA to forms a RISC complex.

k_d	Dissociation constant of transcription factor complex from the promoter region of miRNA gene.
γ_m,γ_μ,γ_p	mRNA, miRNA, TF degradation rate respectively.

For Chapter 4

u	number of pre-mRNA without splicing factor binds to its upstream weak splicing site (PM1).
b	number of pre-mRNA with splicing factor binds to its upstream weak splicing site (PM2).
m_1	mRNA isoform produced from PM1.
m_2	mRNA isoform produced from PM2.
c	splicing factor concentration in nucleus.

Abbreviations

RISC	RNA-induced silencing complex.
TF	Transcription factor.
snRNPs	small nuclear ribonucleic particles.
Pol II	RNA polymerase II, an enzyme found in eukaryotic cells. It catalyzes the transcription of DNA to synthesize precursors of mRNA and most snRNA and microRNA.

Acknowledgements

Scientific research is filled with frustration and difficulty, without love and encourage from people beside me, I cannot achieve what I have had today. I would like to take this opportunity to thank everyone who has helped me.

First of all, I want to thank my supervisor Prof. Sridhar Raghavachari. His passion and integrity in science has been my constant source of inspiration. He introduced me to the frontier of research on biophysics. He also taught me how to be an excellent scientist. I greatly appreciate his patient and care when I was going through personal difficulties. I am extremely fortunate to have him as my supervisor. Not only he always cares about us, he also supports us financially. He sponsored me to go to Cold Spring Harbor Laboratory for computational cell biology summer school. He also sponsored me to a lot of conferences, helped me travel across the country: Washington DC, New Orleans, Santa Fee. I even went to Hawaii twice for winter q-bio conferences! He is the best supervisor in the world!

I also appreciate the help and support from Prof. Nicolas Buchler, who supervises my PhD study, from Prof. Joshua Socolar, Prof. Anne West and Prof. Kate Scholberg, who serve as my dissertation committee members.

A special thank-you to Prof. Glenn Edwards for leading me on my track at the beginning of my Ph.D. pursuit, to Prof. Henry Greenside for giving me a lot of help and coaching me pass my preliminary exam.

Secondly, I want to thank physics department of Duke University. I am so hon-

ored to be a part of it. Everyone in physics is so kind and helpful. Also you can choose your research group based on you research interests, even research group in other departments. This gives me a chance to do interdisciplinary research with the guidance of two great advisors, one from physics department and the other from neurobiology department.

Last but not least, I want to take this opportunity to thank my parents and all my friends. My parents always encourage me to do what most interests me. They never forced me to do anything I don't want. Furthermore, because of them, I understand the meaning of love. I feel so grateful for them to have such a faith in me and support me with all their heart and soul. I want to thank all the members in the Laboratory, Xiaohua Xu, Yu Song and Samuel Ramirez. They always supported me and lent me a hand whenever I asked for help. I want to thank my friend Jun Yan, Ying Wang, Meizhen Shi, Yu Song and Jia Li and more who took great care of me during my worst period of time. That's a memory I want to forget, however, I am indebted to all of you and I will always remember how you helped me and supported me. I also want to say special thanks to Shutong Jiang. I am sorry it didn't work out between us. However, I cherish every moment we spent together since my undergraduate. There are a lot of memories that I will never forget in my whole life. I hope you will be happy in the future even without my accompany.

1

Introduction

1.1 General Introduction of Gene Expression

Transcription and translation are the two main steps of gene expression. A typical gene is first transcribed into mRNA by an enzyme called RNA polymerase and that mRNA is then translated into proteins. Transcription process can be further divided into initiation, elongation and termination. During initiation process, RNA polymerase (RNAP) binds to the promoter region of a gene with the help of transcription factors (TFs), which initiate transcription. After RNAP binds to DNA at the transcription start site, it traverses the template strand and creates an mRNA strand. This is called transcription elongation. In eukaryotes, the elongation process is also coupled with mRNA processing, such as capping, splicing, etc. Transcription proceeds until RNAP transcribes a DNA sequence called terminator. Then Polymerase and RNA are released from DNA. After transcription terminates, mRNA moves out of the nucleus, binds to ribosome complex and produces polypeptide, which will later fold into an active protein.

However, although there are thousands of genes in the genome of a cell, only a

fraction is expressed at a given time. Gene regulation is the reason why not all genes are expressed all the time. Gene regulation ensures that cells will only express protein products when needed. Lac operon in E. coli was the first gene regulation system that was fully understood. It was worked out by Jacob and Monod in 1961 [1]. It has both the transcription repressor and transcription activator to regulate transcription. The amount of transcription repressor and transcription activator is influenced by environmental factors, i.e., the concentrations of glucose and lactose. Many other prokaryotic genes are regulated in a similar fashion, and the basic principles carry over into eukaryotes. Beginning with the work of Jacob and Monod, a qualitative picture of gene regulation by environmental factors has emerged. Cells encode the environmental state in the concentration of transcription factors, special proteins that bind particular sites on DNA and enhance or inhibit the rate at which nearby genes are read (transcribed) into mRNA, which is then translated into proteins.

Genes can be regulated at many stages along the path from DNA to protein. This is what people have been working on for decades. They want to know how every stage influence gene expression and most recently, they've found even cells or organisms with the same genes, in the same environment, with the same history, display variations in form and behavior that can be subtle or dramatic. Transcriptional control is one of the most important mechanisms and will be the focus of our discussion. The most pervasive form of gene regulation involves the initiation of transcription, in which proteins called transcription factors bind to a DNA sequence in response to environmental signals such as temperature, energy sources, hormones or developmental programs, and help or inhibit the binding of RNAP.

Cells live in a complex and dynamic environment, where physical parameters such as temperature, osmotic pressure and mechanical stimuli as well as chemical parameters such as nutrients, harmful chemicals and signaling molecules change over a wide variety of time-scales. Both single-cell and multi-cellular organisms have de-

veloped the ability to detect and respond to these environmental changes, enabling them to survive and propagate. One kind of an adaptive response is the immediate and transient modification of proteins (millisecond to minute time scales) so as to alter cellular morphology or intracellular molecular concentrations. Environmental changes also trigger long-lived and possibly slower responses by producing new proteins that modify the state of the cell in an appropriate manner.

One can view the cell as a computational device, which represents the external environment by the concentrations of molecules and the cellular response by the change of the concentrations. Framing the question in this manner highlights several quantitative issues – namely, the precision with which environmental changes can be detected and discriminated, the speed with which the cell can respond and the reliability of these responses. Extensive theoretical and experimental work has focused on how biochemical signaling networks solve this problem [2, 3]. However, it is far from clear how cells measure, process and integrate environmental stimuli (or internal states for that matter) in order to produce appropriate amounts and types of proteins by transcribing genes.

1.2 Stochastic Model of Gene Expression

The single gene is the fundamental module of gene regulatory circuits. A simple case of single gene expression is where mRNA molecules are transcribed and are translated at constant rates. We can describe this system as a Markov process. A Markov process is a mathematical model for the random evolution of a memoryless system, that is, the future state of the system, at any given moment, depends only on its present state, and not on any past states. A simple, intuitive model of gene expression that contains the essential features and successfully reproduces experimental observations is:

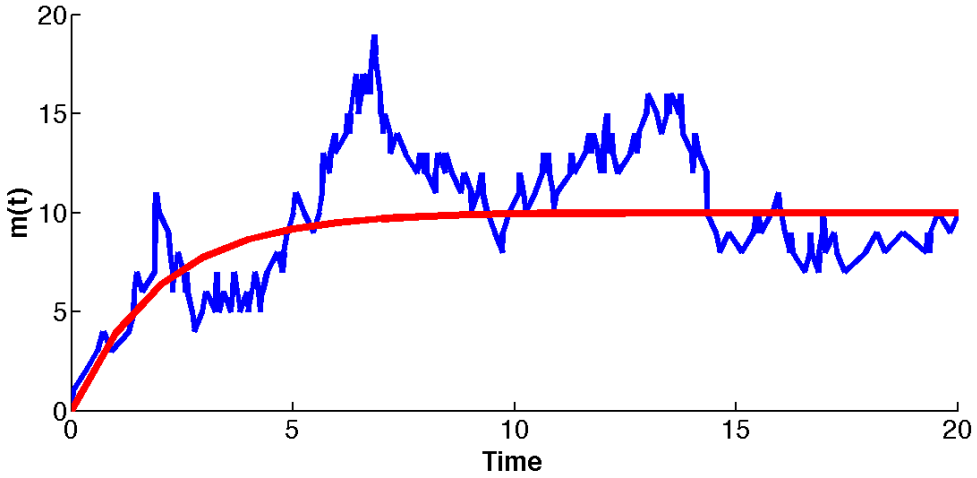


FIGURE 1.1: mRNA evolution based on Mass Action equation and simulation. The smooth red line corresponds to Equation 1.3 and blue curve is the simulation results by assuming each process is Poisson. The parameters are chosen as $k_m = 5.0$ and $\gamma_m = 0.5$.

$$\begin{aligned}
\text{DNA} &\xrightarrow{k_m} \text{DNA} + \text{mRNA} && \text{transcription},\\
\text{mRNA} &\xrightarrow{k_p} \text{mRNA} + \text{protein} && \text{translation},\\
\text{mRNA} &\xrightarrow{\gamma_m} \phi && \text{mRNA degradation},\\
\text{protein} &\xrightarrow{\gamma_p} \phi && \text{protein degradation}.
\end{aligned} \tag{1.1}$$

with rate constants k_m,k_p,γ_m and γ_p respectively. For a single gene regulatory, we can assume the number of DNA is a unit number(i.e. 1) and m and p denote the number of mRNA and protein.

The simplest approach for analyzing this reaction scheme is the law of Mass Action. Let $m(t)$ denote the number of mRNAs in the cell at time t. Then the Mass Action equation describing the evolution of mRNA is:

$$\frac{\mathrm{d}m(t)}{\mathrm{d}t} = k_m - \gamma_m m(t). \tag{1.2}$$

By assuming the initial condition that m(0) = 0, the solution to this equation can

be found as:

$$m(t) = \frac{k_m}{\gamma_m}(1 - e^{-\gamma_m t}). \tag{1.3}$$

Equation 1.3 is the smooth line as shown in Figure 1.1. The smoothness is due to the fact that the Law of Mass Action considers the reaction to be macroscopic and deterministic. However, the simulation results show each realization of the evolution will randomly fluctuate around the smooth line. The Mass Action solution can only describe the ensemble average of the system evolution. In sum, the whole process of gene expression is better to describe as a random process because biochemical reactions are intrinsically noisy and the randomness is a natural consequence of the discreteness of molecules. If the number of transcript is very large, it will approach the deterministic solution that was described in this Mass Action equation. However, If the number of molecules is very small, for example, there are usually only one or two copies of DNA at any given point in the cell cycle and the mRNA level is usually low due to low transcription rate and short mRNA half-life[4], the system will be very noisy and we cannot neglect the effect of randomness. We have to describe gene expression as a stochastic process.

The stochasticity in gene expression, far from being a nuisance, can have great value in a number of circumstances. For instance, stochasticity can induce diversity in a population or allow organisms to optimize their metabolism in fluctuating environments [5]. However, stability against fluctuations is essential when the protein product of one gene controls another gene, as occurs in many cases. Thus understanding the origins of these fluctuations, and their dependence on parameters characterizing gene expression requires theoretical modeling, which can be used to make predictions and guide experiments.

The simplest stochastic model of gene expression was originally proposed in 1977 [6]. By assume that all these fundamental processes (e.g. transcription, transla-

tion and decay of molecules) are Poisson processes, the probability that one reaction with rate k for each of the molecule will take place during the small time interval t to $t + \Delta t$ is kt.

If we choose the time step to be small enough, then for any time step only one process can be realized. Also the process picked up at a certain time doesn't depend on the history of the processes to be chosen. The dynamics of these genetic expressions can be described by coupled birth-death processes. Birth corresponds to synthesis while death occurs via degradation. So this set of transitions defines a Markov system. Let P_m be the probability that the number of mRNA in the system is m, P_p denote the probability that the number of protein in the system is p. So the system can be described by Master Equations as follows:

$$\begin{aligned} \frac{dp_m(t)}{dt} &= k_m P_{m-1}(t) + (m+1)\gamma_m P_{m+1}(t) - (k_m + m\gamma_m)P_m(t) \\ \frac{dP_p(t)}{dt} &= mk_p P_{p-1}(t) + (p+1)\gamma_p P_{p+1}(t) - (mk_p + p\gamma_p)P_p(t) \end{aligned} \tag{1.4}$$

In the study of probability, given two random variables X and Y, the joint probability of X and Y is the probability of the intersection of the events X and Y, i.e., both events X and Y occurring together. If we use the joint probability $P_{(m,p)}(t)$ to describe the probability that at certain time t, the number of mRNA is m and the number of protein is p. the system can be described as:

$$\begin{aligned} \frac{dP_{m,p}(t)}{dt} =& k_m P_{m-1,p}(t) + (m+1)\gamma_m P_{m+1,p}(t) \\ &+ mk_p P_{m,p-1}(t) + (p+1)\gamma_p P_{m,p+1}(t) \\ &- (k_m + m\gamma_m + mk_p + p\gamma_p)P_{m,p}(t) \end{aligned} \tag{1.5}$$

The mean (or first moment) of a function $f(x)$:

$$\langle f(x) \rangle = \int f(x)P(x)\,dx \tag{1.6}$$

Because m and p are discrete integer numbers, the mean should be:

$$\begin{aligned} \langle m \rangle &= \sum m P_m \\ \langle p \rangle &= \sum p P_p \end{aligned} \tag{1.7}$$

Also the probability should satisfy: $\sum P(m) = 1$ and $\sum P(p) = 1$. If we multiply both side of Eqn. 1.4 by either m or p and summing over all possible value of m or p, we can get the following differential equation for the mean value of mRNA and protein:

$$\begin{aligned} \frac{d\langle m \rangle}{dt} &= k_m - \gamma_m \langle m \rangle \\ \frac{d\langle p \rangle}{dt} &= k_p \langle m \rangle - \gamma_p \langle p \rangle \end{aligned} \tag{1.8}$$

For a simple case of single gene expression, the synthesis and degradation rate of both mRNA and protein are constant. From Eq. 1.8, in steady state, the mean and Fano factor[1] are:

$$\begin{array}{rclcrcl} \langle m \rangle & = & \frac{k_m}{\gamma_m} & & \frac{\delta m^2}{\langle m \rangle} & = & 1 \\ \langle p \rangle & = & \frac{k_m b}{\gamma_p} & & \frac{\delta p^2}{\langle p \rangle} & = & \frac{b}{1+\eta} + 1 \end{array}$$

where $b = \frac{k_p}{\gamma_m}$ and $\eta = \frac{\gamma_p}{\gamma_m}$. So the mRNA distribution is Poisson. Usually the lifetime of protein is much long than the lifetime of mRNA, i.e., $\eta \ll 1$, so the Fano factor $\frac{\delta p^2}{\langle p \rangle} \cong b + 1$.

Although this calculation has been proposed in 1977, until relatively recently, experiments studying the properties of gene expression focused on entire populations of cells rather than individual cells because of the technical challenges involved in making single-cell measurements. The first experimental observations, which suggest this model is capable of explaining the properties of gene expression in prokaryotes, was done by Ozbudak et al. [7] and Elowitz et al. [8] in bacteria in 2002.

[1] Fano factor is a measure of noise strength. The definition of Fano factor is the variance over the mean. For Poisson process, Fano factor equals 1.

The development of techniques using fluorescent proteins, such as green fluorescent protein (GFP), led to an explosion in experimental work which showed that gene expression is inherently stochastic in individual cells that continues until now. Technological advances increase the temporal and spatial resolution with which RNA can be studied, down to the single-cell and even single-molecule level [7, 9, 10]. A lot of recent papers illustrate how high-resolution methods can provide insight into the rate and regulation of distinct stages of transcription. For Example, Rabani et al. provide high temporal resolution of RNA regulation by using metabolic labeling of RNA coupled to high-throughput RNA sequencing [11]; Larson et al. can monitor transcription of single molecules over time in individual, living yeast cells [12]; Harper et al. explored real-time dynamics in rat cells by using fluorescence microscopy to monitor expression of reporter genes driven by prolactin promoter [13]; So et al. quantified the number of copies of various mRNAs by single-molecules fluorescence in situ hybridization (smFISH) in Escherichia coli [14].

Recent experimental advances have enabled an expansion of this qualitative picture into the quantitative measurement of the input (environmental changes)/output (gene expression levels) relationships in single cells. A fundamental point that has emerged from these studies is that gene expression at the single-cell level is intrinsically stochastic owing to the fact that the number of molecules involved in gene expression (DNA, mRNA, transcription factors, etc.) in each cell is small. For example, The small numbers of these molecules implies that the fluctuations can - not, in general, be averaged away. Also DNA/transcription factor interactions are $\sim \mathcal{O}(k_B T)$ and hence subject to thermal fluctuations and diffusion is a major randomizing influence on all reactions. These observations raise the following questions – does noise play an important role in the cells' ability to detect, encode and respond to environmental changes, and if so, how do genetic networks work reliably in the presence of this noise? Recent experimental [7, 8, 15–21] and theoretical [22, 23]

work has begun addressing these questions from a quantitative viewpoint.

Since the model is essentially stochastic, the relevant quantities are the probability distributions of the variables m and p. A number of techniques, both numerical and analytic, can be used to calculate the distribution as well as their low order moments, the mean and the variance. Experiments are often limited in their ability to estimate the entire distribution and thus restricted to estimating the mean and the variance.

Most, if not all, of these studies have examined these questions in single-celled organisms (yeast or bacteria) acting in response to quasi-static stimuli, such as steady-state levels of nutrients, medium osmolarity or signaling molecules. This leaves open the question how gene networks encode and respond to temporally varying stimuli. Quantitatively speaking, what is the input-output relation between activity and transcription? How does intrinsic noise in transcription influence the protein output? Is gene expression optimized for monitoring mean levels of activity or tracking variations in activity instead? These are some of the many questions that arise. In order to deduce this relationship, we need to understand the statistics of noisy gene expression.

However, the simple model is insufficient to explain the experimental observations in eukaryotes [16, 19, 20, 24, 25], where mRNA transcripts appear to be produced in bursts. While the number of transcripts within a burst is Poisson-distributed, the overall distribution is not. However, these observations are compatible with a more sophisticated model in which the gene itself randomly transitioned between transcriptionally active state D_I and inactive state D_A. Although the mechanism behind such burst behavior is not known, the most likely explanation is that in eukaryotes, this burst behavior is due to chromatin remodeling [13]: In eukaryotes, DNA is wrapped around proteins called histones and packed into a very condensed structure called chromatin which limits the access to transcription factors and RNAP. Chromatin remodeling is the enzyme-assisted movement of nucleosomes on DNA. By moving

nucleosomes, proteins like transcription factors can get access to DNA that was previously unavailable, wrapped around nucleosome cores. Remodelers are thus necessary to provide access to the underlying DNA to enable transcription. Experiments that have managed to observe the production of mRNA transcripts of synthetic genes suggest that transcription coincides with the opening of chromatin [19]. Moreover, the decondensation of chromatin appears to be stochastic, consequently changes the accessibility of the gene in a stochastic manner.

Previous studies [15, 19] have considered a model in which the fluctuations of transcription rates are caused by overall dynamics that can be described by a "random telegraph process", where the gene switches between an active and an inactive state:

$$D_I \underset{k_-}{\overset{k_+}{\rightleftharpoons}} D_A \tag{1.9}$$

$T_A = 1/k_-$ and $T_I = 1/k_+$ are the mean residence time for the active and inactive states respectively. The probability distribution for the duration time a gene stays at active or inactive state is exponential. The main properties of exponential distribution is memoryless, so the time already spent waiting in that state would not affect how much longer one would have to wait until the next switch.

In Chapter 2, we describe a mathematical formalism that we use to represent stochastic gene expression. We represent the elementary steps in a coarse-grained manner as a Markov process and derive a master equation for the probability distribution of the number of mRNA molecules. We then derive a time dependent solution of mRNA distribution and how this distribution changes with time in response to the changes of the rate corresponding to external stimuli. This model forms a basic building block for understanding the signal transmission and noise property of transcription process.

1.3 Post-transcriptional Function of Micro-RNAs

miRNAs are short (on average of 22 nucleotides long), non-coding RNA molecules that act as post-transcriptional regulators [26]. miRNAs regulate gene expression by base-pairing to target mRNA molecules at conserved sites in the $3'$ untranslated regions of mRNAs, ultimately leading to a reduction in the levels of protein encoded by the target mRNA [27]. Extensive evidence suggests that this suppression can occur by either the repression of translation or induction of mRNA degradation. In the former, miRNAs act as catalytic factors, preventing the initiation of translation, suppressing the production of proteins. In the latter, miRNAs act in a non-catalytic fashion, leading to the degradation of the target mRNA and the miRNA itself. Through either mechanism, miRNAs can keep gene products at extremely low copy numbers. Although thousands of mammalian genes are potentially targeted by miRNAs [27] and miRNAs have been identified as the primary negative regulators of gene expression, the functions of miRNAs in the context of gene networks are still not well understood [28–31].

Of particular relevance is the accumulating evidence that small non-coding RNAs combine with transcriptional activators and repressors to regulate key developmental events [29, 32–41]. Bioinformatics analysis have identified an abundance of negative feedback motifs involving miRNAs and transcriptional activators and repressors that control differentiation [28, 29, 31, 42]. These observations imply the existence of considerable crosstalk between the miRNA-mediated posttranscription layer, and the transcriptional regulation layer, whose dominant players, the transcription factors (TFs), regulate the production of protein-coding mRNAs. Analysis of transcription factor mediated feedback loops suggests that they serve to maintain protein expression at a fixed level. In this way, negative feedback loops buffer against fluctuations arising from environmental variations as well as intrinsic stochasticity of biochemical

reactions, imparting precision and robustness to regulation of gene expression. However, it is not clear whether miRNA-mediated negative feedback regulation similarly acts to suppress fluctuations in TF numbers. Moreover, it is not known how the feedback regulation of TF levels by miRNAs impact the activation of other genes regulated by the same TFs.

To address these questions, we study the dynamics of a negative feedback circuit consisting of a TF that activates the production of a miRNA, which in turn acts as a translational repressor of the transcription factor. This model circuit is motivated by a recent study suggesting that transcription factor, *pitx3* and the microRNA, miR-133b, form a negative feedback circuit in midbrain dopamine neurons [29]. *Pitx3* is a transcription factor for genes that mark the differentiation of precursor cells into dopaminergic neurons in the mammalian midbrain. These neurons release dopamine, an important neuromodulator involved in motivated behavior, learning and memory, and the loss of these neurons results in Parkinson's disease. miR-133 suppresses the translation of *pitx3* mRNA while *pitx3* induces transcription of miR-133b. Thus, the control of *pitx3* levels by this feedback circuit may play a vital role in the maintenance and survival of dopaminergic neurons.

We propose two simplified models which implement the non-catalytic and catalytic mode of translational repression by miRNA: which we term the sequestration model and the kinetic suppression model respectively. We show that these arise as limiting cases of a more complete model of miRNA based repression. We characterize and compare the steady-state behavior and noise properties of the two different modes of action of miRNA in this circuit. Specifically, we ask 1) how is the intrinsic noise of a gene network influenced by miRNA regulation and 2) whether miRNA: mRNA degradation and degradation-independent translational repression have a similar effect on the noise properties of the network. Finally, we show that these two modes of translational repression have distinct effects on genes controlled

by the common transcription factor.

1.4 Alternative Splicing

Eukaryotic genes were found to be broken up into small pieces of exons and introns. RNA splicing removes introns from newly transcribed pre-mRNAs. This process is catalyzed by spliceosome, including five small nuclear ribonucleoprotein particles and a large number of additional proteins. mRNA splicing mechanism is a prerequisite mechanism for gene expression.

Alternative splicing of pre-mRNA is the prominent source of protein diversity . During the process to remove intron from pre-mRNA to produce mature mRNA, which occurs in most eukaryotic genes, particular exons may be included within, or excluded from, the final, processed messenger RNA (mRNA) produced from that gene. These alternatively spliced mRNA isoforms then can encode functionally distinct protein products. Of course, sometimes, mRNA isoforms don't encode for any protein.

Alternative splicing is more important than expected. In humans, about 95% of multiexonic genes are alternatively spliced [43]. This is potentially a large noise source for gene expression variability between individual cells. Changing alternative splicing patterns in response to an stimulus seems to be a physiological process performed by many cells. Sometimes, misregulation of alternative splicing is important cause and indication of human genetic diseases[44, 45] as well as cancer [46, 47]. Although extensive researches have been done to establish the noise properties of gene expression [7, 8, 48–52], little is known how splicing process itself influence the noise properties of gene expression. Character this variability is also the key to understand how cells regulate the alternative splicing process in response to stimuli to adjust levels of distinct functioned isoforms.

Diffusion of RNA molecules and proteins within the cell nucleus is central to

genome function. According to the kinetic coupling model, owing to the high mobility of splicing factors, they can directly assemble on the nascent intron splicing site to facilitate splicing as soon as they are synthesized by RNA polymerase II and exposed [53]. The splice-site consensus sequences can attract splicing regulatory proteins to suppress or promote splicing [54]. Thus organisms can regulate alternative splicing site selection by changing the concentration and activity of splicing regulatory proteins.

Previous studies [55] have demonstrated that by assuming a binomial outcome of splicing due to fluctuation in splicing factors, mammalian cells minimize fluctuations in mRNA isoform ratios by tightly regulating the splicing machinery. However, this model is simply intuitive and doesn't include analysis of any parameters other than transcriptional rate. More work regarding diffusion of splicing factors is needed.

However, how diffusion of splicing factors affect alternative splicing process as well as the noise properties in this system remains unknown. In Chapter 4, we fully investigate the functional role of splicing factor diffusion as well as how it affect the noise propagation in exon skipping. This work can also be extended to intron retention and other kinds of alternative splicing.

1.5 Activity-dependent transcription in Neurons

Activity-dependent transcription is also very important in many aspect for nervous system development as well as synaptic plasticity. Neurons, the core components of the brain, are excitable cells, which signal to each other by rapid changes in their membrane electrical potential (the action potential). Action potentials can then cause the release of chemicals at connections between neurons (synapses), which in turn leads to changes in membrane potentials of the recipient cells by means of specialized proteins in the cell membrane. Action potentials not only mediate signaling between neurons, but also control transcription of genes important for

neuronal function. Among the many functions of activity dependent gene expression is the alteration of the levels of key proteins at synapses, thereby altering the synaptic connection between neurons for days or longer in response to changes of sensory experience [56–60].

Neuronal activity opens ion channels, cause Influx of Ca^{2+}. Ca^{2+} then may enter into nucleus and activate transcription factors (TFs). TFs bind to promoter region of DNA and initiate transcription. So gene can be at silent state or at activated state. The motivation of the first project is to investigate How neuronal gene expression can respond to time varying signal. The second project studied a generalized negative feedback regulation between microRNA and transcription factor, which may exists in nervous system development where dopamine neurons are specified from precursor cell. MicroRNA and transcription factor network are generally found in many cases. Therefore, we want to understand the function role of noise in this negative feedback loop. The motivation of the third project is the observation that splice isoform ΔFosB seems to play a very important role in the development of addictive behavior. Trying to connect FosB and ΔFosB all the way to synaptic plasticity is difficult because many of the intermediate steps are not well understood. What we can try is to explore is how alternative splicing itself will contribute to the noise propagation in gene expression. This is motivated by neurobiology, however, it can also being applied to many other biological systems since alternative splicing occurs in around 90% of multi-exonic pre-mRNAs. Finally, I will conclude what we have learnt from these projects and have a future outlook of what we can do next.

2

Bursty Gene Expression Analysis

2.1 Model Description

While in most cases prokaryotic transcription can be explained by a Poisson model, it is becoming increasingly clear that the number of transcripts when counted at a single cell level show highly non-Poisson behavior [16, 19, 20, 24, 25] . These observations are thus compatible with a more sophisticated model wherein the transcriptional apparatus transitions between multiple states, all but one being inactive for transcription. This may include, for instance, an RNA polymerase that is paused at the promoter, and commences transcription when the transcription factor binds to the promoter and allows polymerase escape. Another possibility is chromatin remodeling. In eukaryotes, DNA is wrapped around proteins called histones and packed into a very condensed structure called chromatin, which limits the access to transcription factors and RNAP. Experiments that have managed to observe the production of mRNA transcripts of synthetic genes suggest that transcription coincides with the opening of chromatin. Moreover, the decondensation of chromatin appears to be stochastic, consequently changing the accessibility of the gene in a stochastic

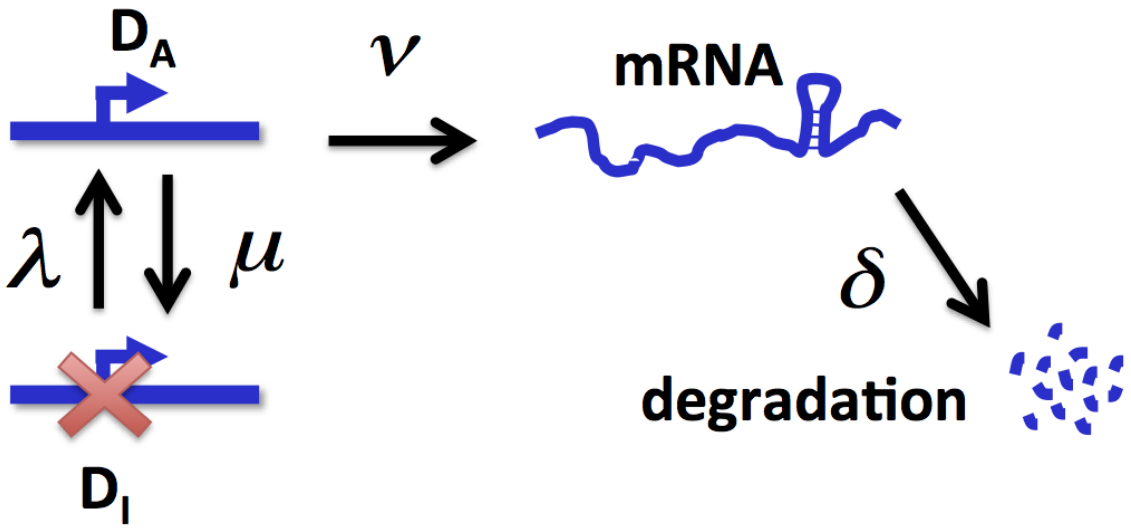


FIGURE 2.1: Scheme of bursty transcription.

manner.

Given the observations of burstiness, a conceptual model of the most elementary multi-state promoter, namely the two state promoter, as described in Figure 2.1, can be constructed. The gene itself randomly transition between transcriptionally inactive state D_I and active state D_A. This model contains four parameters: $\boldsymbol{\lambda}$, the rate at which the gene transitions from the inactive to the active state, which determine the rate of transcription bursts; $\boldsymbol{\mu}$, the rate at which the gene transitions from the active to inactive state, which determines the duration of transcription bursts; $\boldsymbol{\nu}$, the rate of transcription when the gene is in the active state, which determine the burst size, i.e., how many mRNAs are produced during each burst; and $\boldsymbol{\delta}$, the rate of mRNA degradation. Scheme 2.1 describes the mathematical simplification of this model.

Such a model was first analyzed by Peccoud and Ycart[61]. While the steady state solutions of the mRNA distribution was obtained, the time-varying solution, essential for our understanding of the dynamic control of gene expression by time-varying stimuli, is not known. The aim of the following study is to try to understand the

properties of bursts in gene expression. First, an expression for temporal evolution of the mRNA distribution in bursty gene expression is obtained and analyzed. Then the analysis is compared with computer simulation.

$$\begin{aligned} D_I &\xrightarrow{\lambda} D_A && \text{activated gene,} \\ D_A &\xrightarrow{\mu} D_I && \text{inactivated gene,} \\ D_A &\xrightarrow{\nu} D_A + \text{mRNA} && \text{transcription,} \\ \text{mRNA} &\xrightarrow{\delta} \phi && \text{mRNA degradation.} \end{aligned} \tag{2.1}$$

For simplicity, assuming all the rates are constant. The phase space of this system can be written as

$$S = \{(i, m)\ i \in \{0, 1\}, m \in \mathcal{N}\}\,. \tag{2.2}$$

In any state, $i = 0$ denotes that the gene is inactive or cannot transcribe mRNA while $i = 1$ implies that the gene is active and can transcribe mRNA. m denotes the number of mRNA which are produced at a rate ν and degraded at a rate $\delta \times m$. The gene remains activated during a random time distributed exponentially with parameter μ on average. The model is a Markov model in which the gene transitions randomly between an active state and an inactive state.

Let $p_{0,m}$ denote the probability that the gene is in the inactive state and m the number of mRNA and $p_{1,m}$ be the analogous probability when the gene is in the active state. So we have the following system of equations that govern these probabilities:

$$\begin{aligned} \frac{dp_{0,m}}{dt} &= -(\lambda + m\delta)p_{0,m}(t) + \delta(m+1)p_{0,m+1}(t) + \mu p_{1,m}(t), \\ \frac{dp_{1,m}}{dt} &= -(\mu + \nu + m\delta)p_{1,m}(t) + \delta(m+1)p_{1,m+1}(t) + \nu p_{1,m-1}(t) + \lambda p_{0,m}(t), \\ \frac{dp_{1,0}}{dt} &= -(\mu + \nu)p_{1,0}(t) + \delta p_{1,1}(t) + \lambda p_{0,0}. \end{aligned} \tag{2.3}$$

The generating functions for this process are:

$$G_0(z,t) = \sum_{m=0}^{\infty} z^m p_{0,m}(t), \tag{2.4}$$

$$G_1(z,t) = \sum_{m=0}^{\infty} z^m p_{1,m}(t). \tag{2.5}$$

The sum $G(z,t) = G_0(z,t) + G_1(z,t)$ is the generating function for the total mRNA probability distribution at time t.

2.2 Steady State Analysis of Bursty Transcription

We can get the exact solution of the steady state generating function (From Appendix Equation A.21):

$$g(z) \equiv \lim_{t\to\infty} G(z,t) = {}_1F_1(a,b;\frac{\nu}{\delta}(z-1)) = \sum_{n=1}^{\infty} \frac{(a)_n}{(b)_n n!}(\frac{\nu}{\delta})^n (z-1)^n \tag{2.6}$$

where $a \equiv \frac{\lambda}{\delta}$, $b \equiv \frac{1}{\delta}(\lambda+\mu)$.

${}_1F_1(a,b;y) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n n!} y^n$ is the confluent hypergeometric function of the first kind, where $(a)_0 \equiv 1$ and $(a)_m \equiv a(a+1)(a+2)\ldots(a+m-1)$.

From the steady state generating function, we can get all the moments of the mRNA and then derive the steady state mean and Fano factor of mRNA and protein molecules as follows:

$$\langle m \rangle = \frac{\nu\lambda}{\delta(\mu+\lambda)}, \tag{2.7}$$

$$\frac{\delta m^2}{\langle m \rangle} = \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle} = 1 + \frac{\nu\mu}{(\mu+\lambda)(\mu+\lambda+\delta)}. \tag{2.8}$$

Our expression for the mean and variance are identical to those obtained earlier by Peccoud and Ycart [61]. The mean and variance of the steady state mRNA level is not related to the initial state (gene active or inactive, initial mRNA number in the nucleus). It only depends on the rate parameters and for steady state solution, only three dimensionless parameters will affect the outcome: $\frac{\lambda}{\delta}$,$\frac{\mu}{\delta}$ and $\frac{\nu}{\delta}$. Experimental observations have indicated that the half-life of the mRNA is on the order of 5-10 minutes in bacteria [62], around 10 hours in human [63]. Any scaling will result in the same stationary distribution, as such distributions are, by definition, independent of time.

From the expression of the mean mRNA level (Equation 2.7), we can see that changes in mRNA level can be obtained by modifying any of the four kinetic parameters characterizing mRNA production. Different external stimuli can modify one or more parameters. For example, enzymes, such as histone acetyltransferases (HATs), deacetylases, methyltransferases, and kinases can change chromatin remodeling process, thus change the rate gene turns on and off (λ or μ). Transcription factors mainly changes the burst size, i.e., ν; the stability of mRNA lifetime ($\frac{1}{\delta}$) can also be altered by level of decapping enzyme or micro-RNAs, etc.

As shown in figure 2.2, we analyze how the steady state mean and variance of mRNA number altered in response to the changes of one of these parameters, μ. As shown in Figure 2.2 (A) , the mean for bursty gene expression is always lower than that for the gene which is always active (μ=0); If $\mu = \lambda$, the mean for bursty gene is half the value of that for activated gene. Also, the mean value of mRNA is linearly proportional to the value of $\frac{\nu}{\delta}$. Figure 2.2 (B) indicates the Fano factor of bursty gene is always larger than that of active gene which obeys Poisson distribution. While the inactivation rate μ increasing, the level of the mean mRNA number per cell decreases, however, Fano factor first increase, then decrease. Especially when $\mu = \sqrt{\lambda(\lambda + \delta)}$,

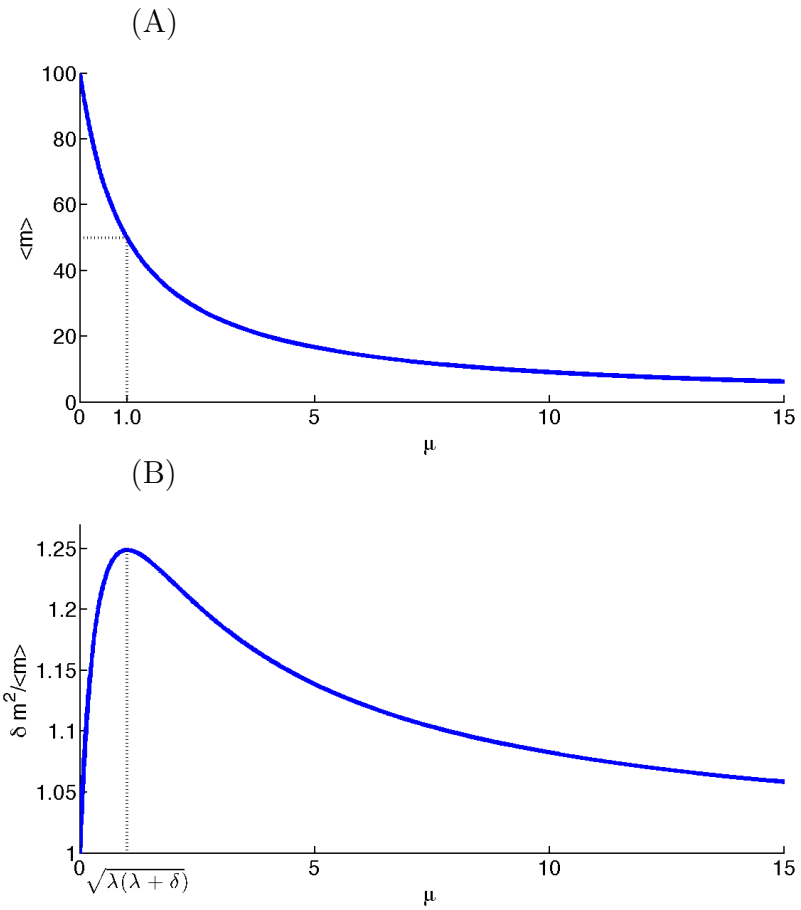


FIGURE 2.2: The mean and Fano factor of mRNA as a function of gene inactivation rate, u. $\lambda = 1.0; \nu = 10.0; \delta = 0.5$. While the inactivation rate increasing, the level of the mean mRNA number per cell decreases, however, Fano factor first increases then decreases. Especially when $\mu = \sqrt{\lambda(\lambda+\delta)}$, Fano factor achieves a maximum: $F = 1 + \frac{\nu}{(\sqrt{\lambda}+\sqrt{\lambda+\delta})^2}$.

Fano factor achieve maximum: $F = 1 + \frac{v}{(\sqrt{\lambda}+\sqrt{\lambda+\delta})^2}$.

We also plot Fano factor versus mean in response to different parameter changes. As shown in Figure 2.3, in each of the subfigures, we vary one parameter and keep the other three constant. We can see, by varying different parameters, that the relations between the mRNA mean and Fano factor are quite different. This has practical uses for experiments. By monitoring the relation between the mean and Fano factor of mRNA with changes of external stimuli, we can have a sense which parameter (parameters) altered in response to those stimuli.

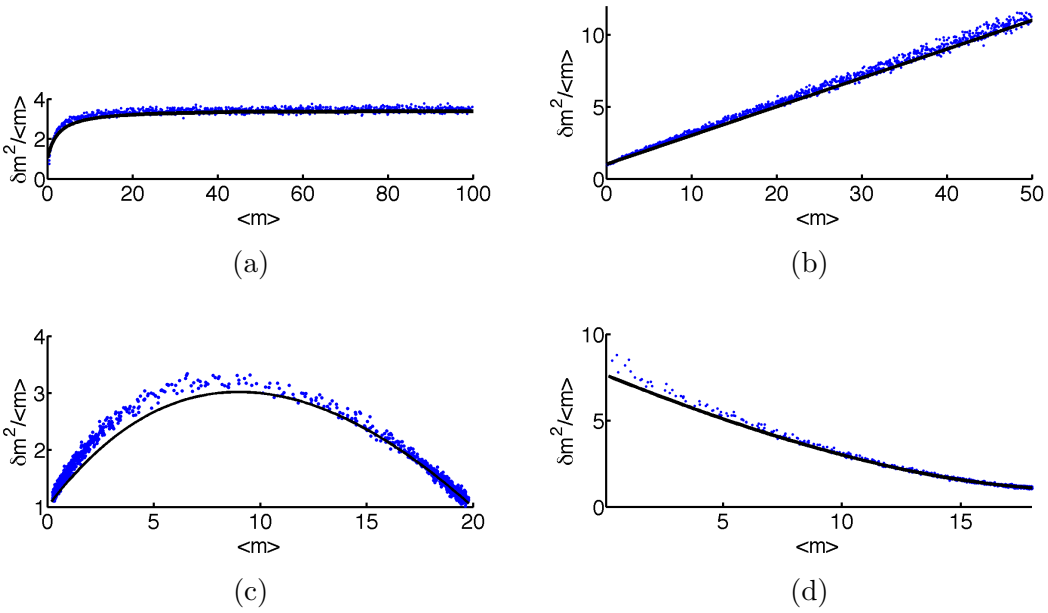


FIGURE 2.3: The noise properties of mRNA: The Fano factor $F = (\langle m^2 \rangle - \langle m \rangle^2)/\langle m \rangle$, is plotted as a function of the mean mRNA number $\langle m \rangle$. Blue dots represent computer simulation results, black line is from theoretical calculation. (a) varying δ. $\lambda = \mu = 1.0$, $\nu = 10$. (b) varying ν. $\lambda = \mu = 1.0$,$\delta = 0.5$; (c) varying μ, $\lambda = 1.0$, $\nu = 10$, $\delta = 0.5$. (d) varying λ. $\mu = 1.0$, $\nu = 10$, $\delta = 0.5$.

2.3 Complex Analysis to get Steady State Probability Distribution

Below we will use the complex analysis to get the probability distribution for different mRNA numbers m for steady state. First, we need to generalize the variable z to be a complex variable.

If a function $f(z)$ can be represented by a Laurent series of the form:

$$f(z) = \sum_{n=0}^{\infty} an(z-a)^n + \sum_{n=0}^{\infty} \frac{b_n}{(z-a)^n}. \tag{2.9}$$

Then

$$a_n = \text{Residue}[\tfrac{f(z)}{(z-a)^{n+1}}] \tag{2.10}$$

$$b_n = \text{Residue}[\tfrac{f(z)}{(z-a)^{-n+1}}]. \tag{2.11}$$

Thus

$$p_m(t \to \infty) = \text{Residue}[\frac{g(z)}{z^{m+1}}]. \tag{2.12}$$

$p_m(t \to \infty)$ represents the probability to have m mRNA at steady state. The full analytical expression of the steady state mRNA probability $p_m(t \to \infty)$ can be calculated as following:

$$p_m(t \to \infty) = \frac{1}{m!}\frac{(a)_m}{(b)_m}(\frac{v}{\delta})^m {}_1F_1(m+a, m+b; -\frac{v}{\delta}). \tag{2.13}$$

where $(a)_0 \equiv 1$ and $(a)_m \equiv a(a+1)(a+2)\dots(a+m-1)$, as defined previously in confluent hypergeometric series.

Because $(a)_m = \frac{\Gamma(m+a)}{\Gamma(a)}$, we can also write Equation 2.13 in the following form:

$$p_m(t \to \infty) = \frac{1}{m!}\frac{\Gamma(m+a)\Gamma(b)}{\Gamma(m+b)\Gamma(a)}(\frac{v}{\delta})^m {}_1F_1(m+a, m+b; -\frac{v}{\delta}), \tag{2.14}$$

So we not only know the first few moments of mRNA distribution, but also know the exact probability distribution of mRNA. As shown in Figure 2.4, we plotted some probability distribution with different parameters. Figure 2.4a, 2.4c show that if transcription rate is much less than mRNA degradation rate, no matter how fast the transition between active and inactive state, the mRNA number will be near zero. Figure 2.4a shows that if transcription rate is fast, activation/inactivation rate is low, then there will be two stable fixed point in the system. i.e., if the system is inactive, it will remain inactive for a long time; if it switches to active state, then it will remain active for a long time and produce mRNA at a relatively fast rate. Figure 2.4a shows us that if both transcription rate and activation/inactivation rate are high, then there will be one stable fixed point in the system and the mRNA distribution is bell shaped. Because we have an analytical expression of probability distribution of mRNA (Equation 2.13), as long as we know the value of these four rates, we can get

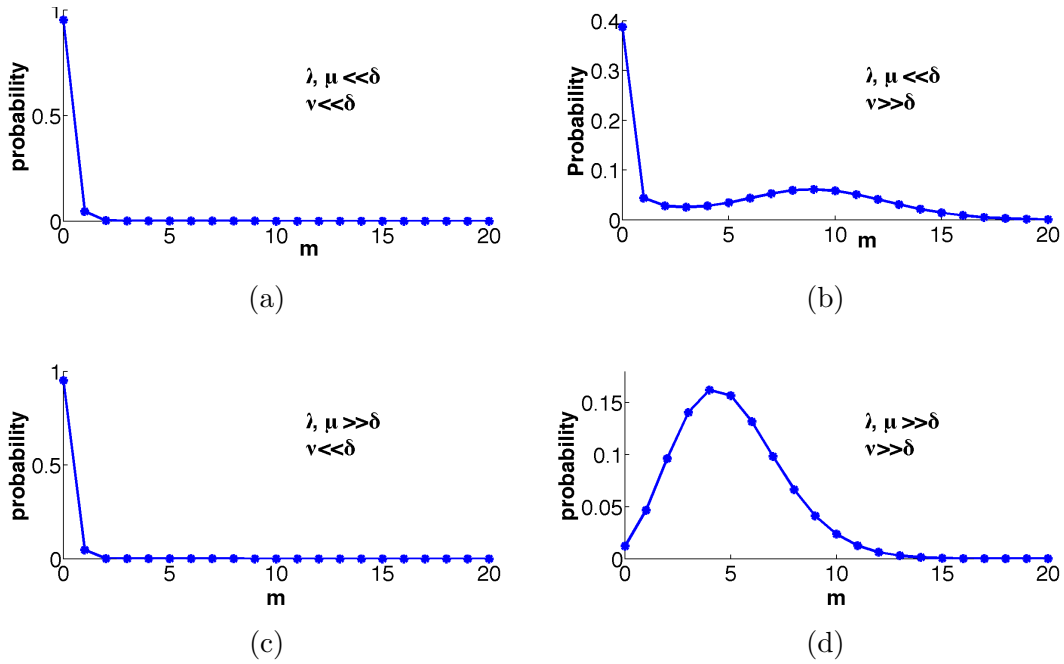
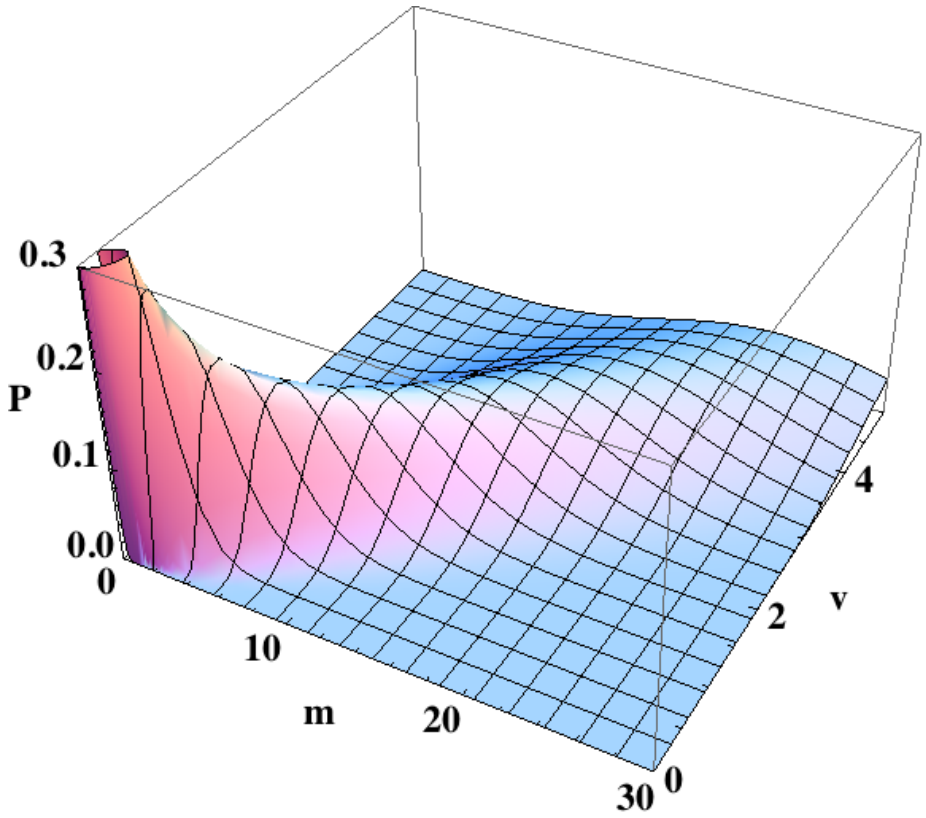


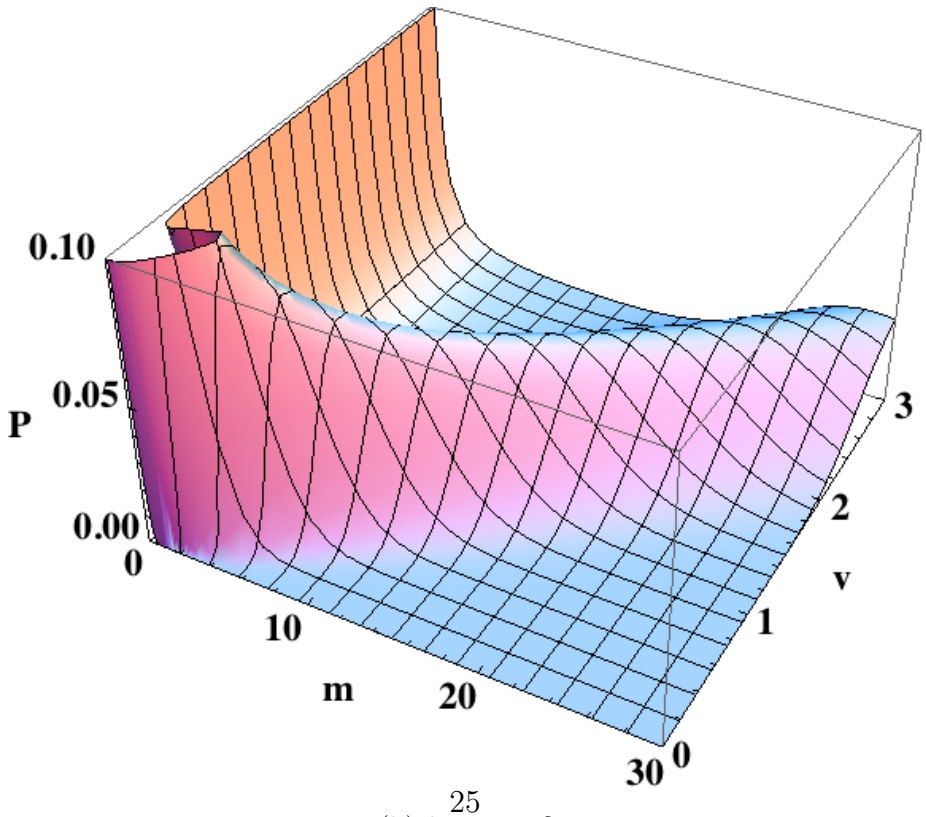
FIGURE 2.4: mRNA probability distribution characteristic diagram: (a) $\lambda = \mu = 0.1\delta$, $\nu = 0.1\delta$, (b) $\lambda = \mu = 0.1\delta$, $\nu = 10\delta$, (c) $\lambda = \mu = 10\delta$, $\nu = 0.1\delta$, (d) $\lambda = \mu = 10\delta$, $\nu = 10\delta$.

the probability distribution of mRNAs. Different values of parameters can result in different mRNA distribution, which can be used to extract parameters from mRNA-counting experiments, potentially revealing new information about which parameters are subjected to regulation [14].

Figure 2.5 shows how probability distributions change with the change of transcription rate. We plot in two cases: 1) When activation rate and inactivation rate are equally high: in this case, if transcription rate is low, mRNA is almost zero, while transcription rate ν increases, the peak of the mRNA distribution increases; 2) When activation rate and inactivation rate are equally low: in this case, at first, it's the same as in previous case. If transcription rate ν is low, mRNA is almost zero, while ν increases, the peak of the mRNA distribution increases. However, while transcription rate continues increasing, mRNA will have two peaks, one near zero, and the other increases with the increase of transcription rate.



(a) $\lambda = \mu \gg \delta$



(b) $\lambda = \mu \ll \delta$

FIGURE 2.5: How mRNA probability distribution evolve with the change of transcription rate: (a) $\lambda = \mu = 1.0$, $\delta = 0.1$; (b) $\lambda = \mu = 0.01$, $\delta = 0.1$

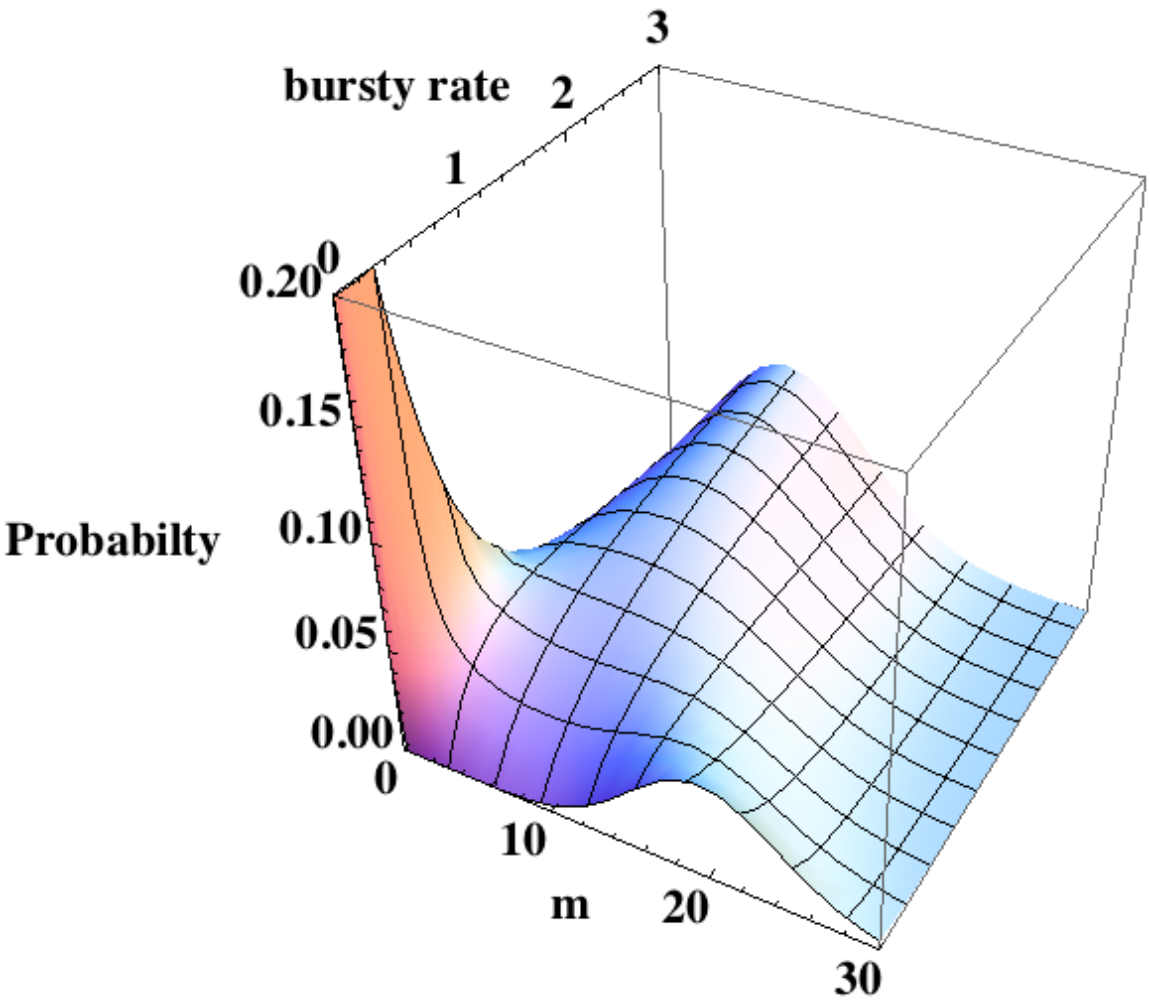


FIGURE 2.6: How mRNA probability distribution evolve with the change of bursty rate. $\nu = 1.0$, $\delta = 0.1$

Figure 2.6 shows how the probability distribution changes with the change of activation rate λ and inactivation rate μ (assume $\lambda = \mu$ changes simultaneously). We can see that the probability distribution will almost remain unchanged while $\lambda = \mu$ becomes big enough. That may indicate that if the system transitions between activated state and inactivated state is fast enough, it will not alter the mean value as well as the noise properties of mRNA.

2.4 Time Evolution Analysis of Bursty Transcription

While the steady state distribution is informative, it does not tell us anything about how the transcriptional process behaves in response to changes in parameters. Cells have to respond to varying environmental conditions by altering their complement of expressed genes. The changes in environmental parameters can manifest themselves as time-varying parameters. Understanding how these affect the evolution of the mRNA distribution is therefore of utmost importance.

For simplicity, we first assume initially $m(t=0)=0$. Then we can also get the time-dependent solution of the generating function (From Appendix Equation A.58):

$$G(z,t) = \frac{{}_1F_1(a,b;\frac{\nu}{\delta}(z-1))\theta(z,t)}{{}_1F_1(a,b;\frac{\nu}{\delta}(z-1)e^{-\delta t})} + \frac{{}_1F_1(1+a-b,2-b;\frac{\nu}{\delta}(z-1))(1-\theta(z,t))}{{}_1F_1(1+a-b,2-b;\frac{\nu}{\delta}(z-1)e^{-\delta t})}e^{(1-b)\delta t}, \tag{2.15}$$

where

$$\theta(z,t) = \big(1+\frac{aX}{b(b-1)}\frac{{}_1F_1(a+1,b+1;X){}_1F_1(1+a-b,2-b;X)}{{}_1F_1(a,b;X){}_1F_1(1+a-b,1-b;X)}\big)^{-1}. \tag{2.16}$$

$X \equiv X(z,t) = \frac{\nu}{\delta}(z-1)e^{-\delta t}$. It's defined here to shorten the expression.

From the time dependent generating function, we can further get the time dependent mean and Fano factor as below (from Appendix Equation A.62 and Equation A.64):

$$\langle m(t)\rangle = \frac{\nu a(b-1-be^{-\delta t}+e^{-b\delta t})}{\delta b(b-1)}, \tag{2.17}$$

$$\frac{\delta m^2}{\langle m\rangle} = 1-\langle m(t)\rangle \tag{2.18}$$

$$+\frac{\nu}{\delta}\frac{\{\frac{2(a-b)}{b+1}e^{-(b+1)\delta t}+\frac{2(b-a-1)}{b-2}e^{-b\delta t}+\frac{(a-1)b}{b-2}e^{-2\delta t}-2ae^{-\delta t}+\frac{(a+1)(b-1)}{b+1}\}}{b(1-e^{-\delta t})-(1-e^{-b\delta t})}, \tag{2.19}$$

where $a=\frac{\lambda}{\delta}$, $b=\frac{\lambda+\mu}{\delta}$. We can also get the time-dependent probability distribution. Due to the complexity of the expression, we didn't list it here.

This is a very meaningful discovery. So now we can know exactly how the system evolves with time. We can also know the mean and variance of a transcription system at any specific time. It's a powerful tool allowing us to investigate more deeply how every parameter affect transcription process. For example, Figure 2.7 shows that at a specific time point, how change of activation rate λ from one value to another will change the mean and Fano factor of the system and how they evolve with time.

Also, due to changes of different parameters at a specific time t_0, even after sometime, the steady state mean value will be the same, the time evolution process will be different and the noise properties will also be different. Assuming external signal increases transcription script level of a bursty gene, this can be done in four different ways : 1) increase transcription rate; 2) increase activation rate or 3) decrease inactivation rate, 4) decrease degradation rate. We analysis how these four ways can affect the time evolution process as well as the noise properties of the system. As shown in Figure 2.8,

Also to be noted, if $b=\frac{\lambda+\mu}{\delta}>1$, the characteristic time $\tau_c=\frac{1}{\delta}$. If $b=\frac{\lambda+\mu}{\delta}<1$, the characteristic time $\tau_c=\frac{1}{b\delta}=\frac{1}{\lambda+\mu}$. In this case, the characteristic time is only related to the bursty rates λ and μ. Transcriptional rate ν and degradation rate δ will change the amplitude of $\langle m\rangle$, but not the characteristic time. The characteristic

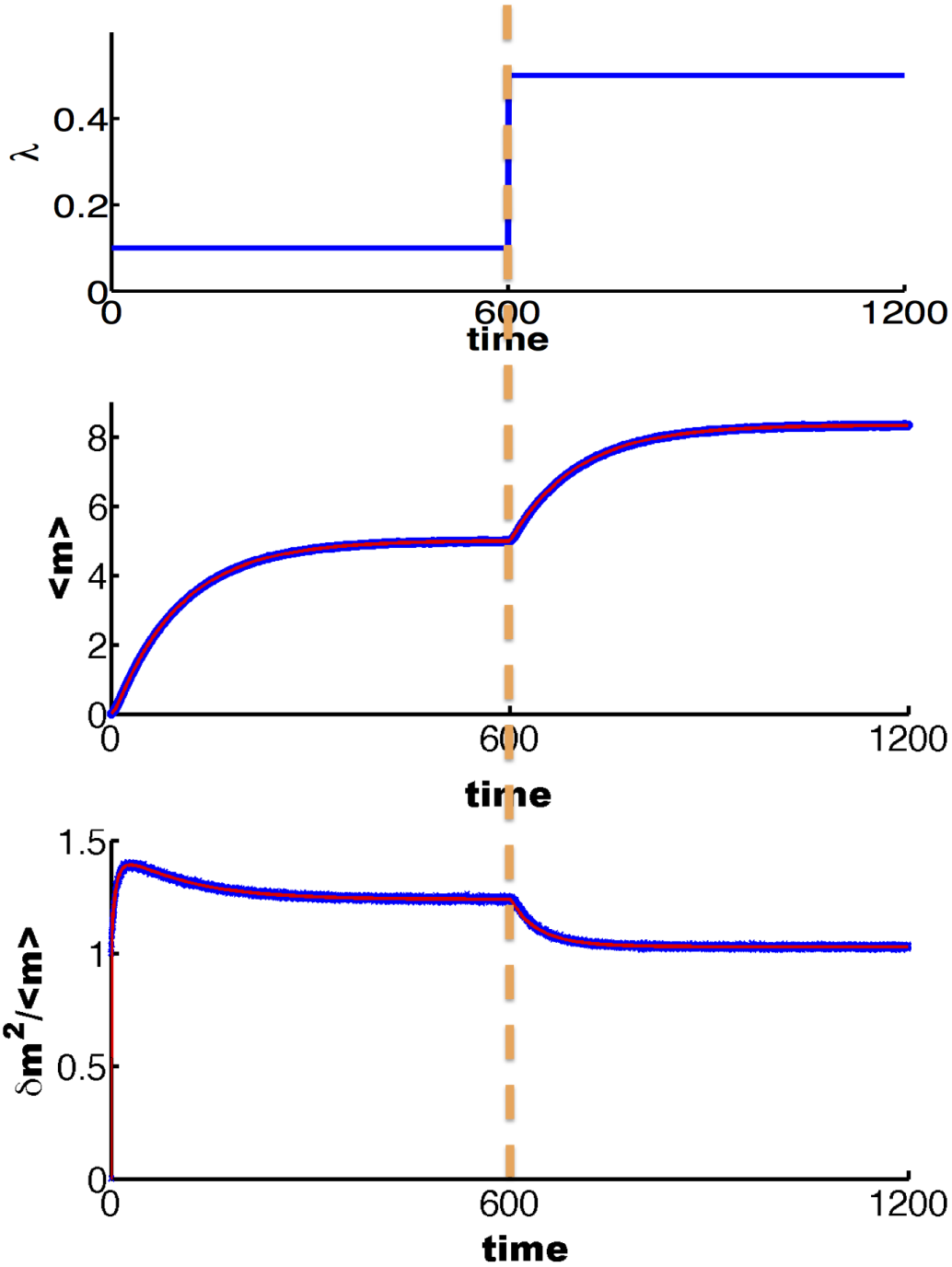
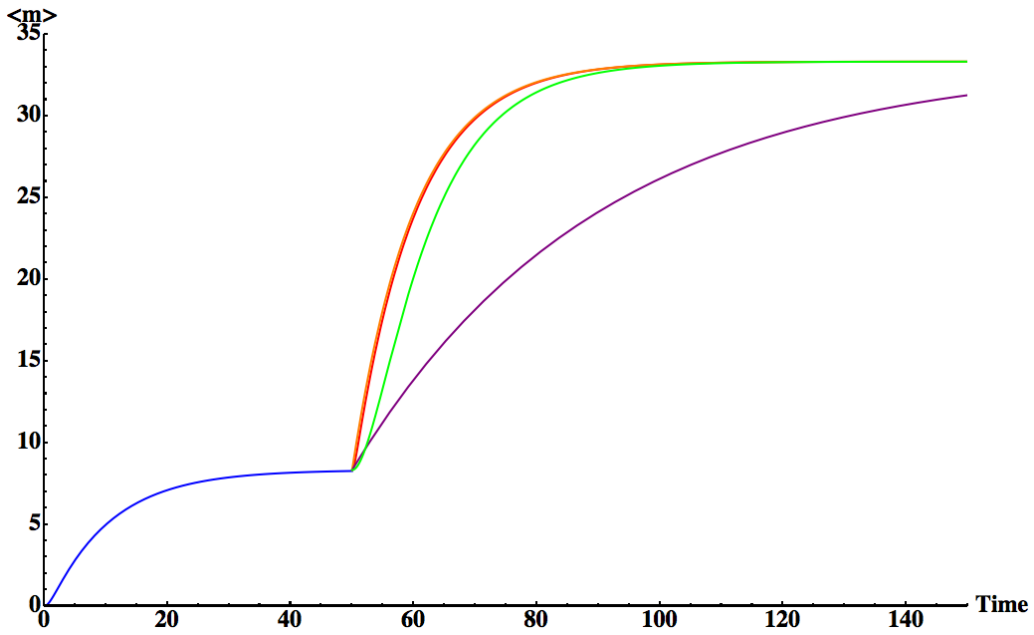
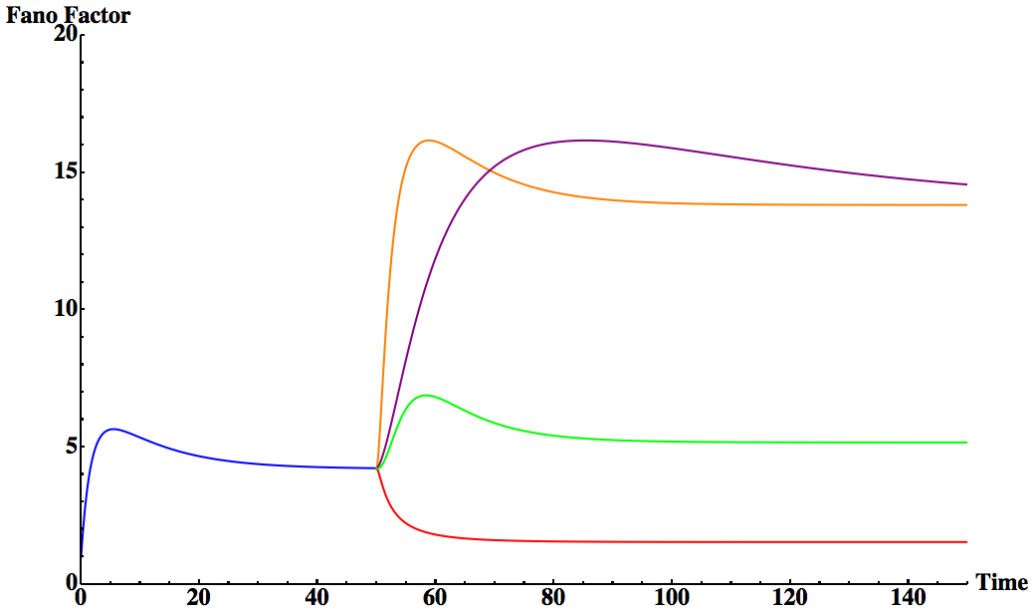


FIGURE 2.7: Time evolution of the mean and Fano factor of mRNA. Red line is the calculation results of the time evolution of the mean and Fano factor of mRNA. Blue dots show computer simulation results of the time evolution of the mean value and Fano factor of mRNA. parameter λ changes from 0.1 to 0.5 at time $t = 600$ second, values of other parameters: $\nu = 0.1$, $\delta = 0.01$ and $\mu = 0.1$.



(a)



(b)

FIGURE 2.8: How different parameters influence the time evolution process as well as the noise properties of transcription: from $t = 0$ to 50 seconds, $\lambda_0 = 0.2$, $\mu_0 = 1.0$, $\nu_0 = 5.0$ and $\delta_0 = 0.1$. This is represented with blue line. At the time $t = 50$, assuming one of the four parameters changes: If λ changes from 0.2 to 2.0 (red line), or μ changes from 1.0 to 0.1 (green line), or ν changes from 5.0 to 20.0 (orange line), or δ changes from 0.1 to 0.025 (purple line), the mean mRNA will converge to the same value; however, the time properties of this convergence as well as the time evolution of the Fano factor will be quite different.

time can measure the time needed for a system to respond under external stimuli to an equilibrium condition from a non-equilibrium condition. Interestingly, if $b < 1$, the response speed is unrelated to the transcription rate and degradation rate.

2.5 Discussion

In summary, we are the first to derive the time-evolution solution for the mean and Fano factor of mRNA during bursty gene expression as well as the full distribution of mRNA. This is adequate for relating model parameters to experiments in which neurons are genetically modified to allow the observation of transcription at a single activity-dependent gene. The model can now be extended to study how temporal variations in different parameters would affect the process of transcription and mRNA statistics. We note that deriving the mRNA and protein distribution simultaneously is not possible analytically. However, in the limit that protein lifetimes are significantly longer than mRNA lifetimes, proteins can be shown to be gamma-distributed according to the number of mRNA.

3

Quantifying Negative Feedback Regulation by micro-RNAs

miRNAs are transcribed from independent miRNA genes or are portions of introns of protein-coding RNA polymerase II transcripts as precursor RNAs that are processed by the enzymes, Dicer and Drosha. The processed miRNA is assembled into a characteristic stem loop structure, cleaved into single strands and loaded onto specialized proteins of the Argonaute (Ago) family, forming an RNA-induced silencing complex (RISC). The RISC complex can then bind to its target mRNA at complementary sequences (7-8 nucleotides long) in the untranslated $3'$ region of the target. This binding leads to suppression of translation in a number of different ways [64, 65]. Perfect or almost perfect complementary leads to the cleavage of miRNA-mRNA duplex [26, 66]. However, this mechanism is relatively rare in animals [26, 67]. Instead, miRNAs tend to destabilize mRNAs by deadenlylation, leading to marked reduction in their abundance, and a consequent decrease in protein levels. The most prevalent mechanism of miRNA action is to repress translation by blocking steps in translation initiation or elongation [26, 65]. The repressed mRNAs accumulate into specialized protein aggregates called P-bodies, where they are either degraded or stored. Impor-

tantly, accumulation of miRNA/mRNA complexes into these P-bodies is correlated with fewer translating ribosomes, leading to lowered translational output for that mRNA. In all cases, mRNA and miRNA can pair with each other in a stoichiometric fashion and move to a translationally incompetent pool.

Extensive work has shown that the complex machinery of gene expression can be described by a coarse-grained model that treats transcription of mRNA and translation of the mRNA message into proteins as discrete events, lumping together many of the component steps into single processes. We assume that the gene encoding the transcription factor (TF) is transcribed at a rate α_m, possibly specified by upstream environmental factors and translated from the mRNA at a constant rate k_p. The TF is degraded at a constant rate γ_p. The transcription factor in turn activates miRNA synthesis. Thus, the rate of miRNA synthesis, α_μ is a function of the number of transcription factors modeled as a Hill function:

$$\alpha_\mu = \sigma p^n/(p^n + k_d^n), \tag{3.1}$$

where σ is the constant transcriptional rate with sufficient TFs, k_d is the dissociation constant of transcription factor complex from the promoter region of miRNA gene, and n is the Hill coefficient. A coefficient of one indicates TFs bind to the gene regulatory region independently of each other and coefficients greater than one indicate positive cooperativity between TFs. As many TFs dimerize and activate the transcription of the miRNA (i.e., fos/fos, fos/jun, creb, etc.), we choose a Hill exponent of 2. Thus, the transcriptional rate of miRNA can be written as $\alpha_\mu = g(p) = \sigma p^2/(p^2 + k_d^2)$. This implicitly assumes that the dimerization of the transcription factors and their binding to the miRNA gene promoter is rapid relative to other timescales in the system. However, we note that relaxing the dimerization assumption does not qualitatively alter our results below. Note that here we have assumed that the promoters of the TF and miRNA genes are always active with no

"bursting" due to remodeling of the chromatin environment [16, 19, 68]. Both mRNA and miRNA are degraded at constant rates γ_m and γ_μ respectively.

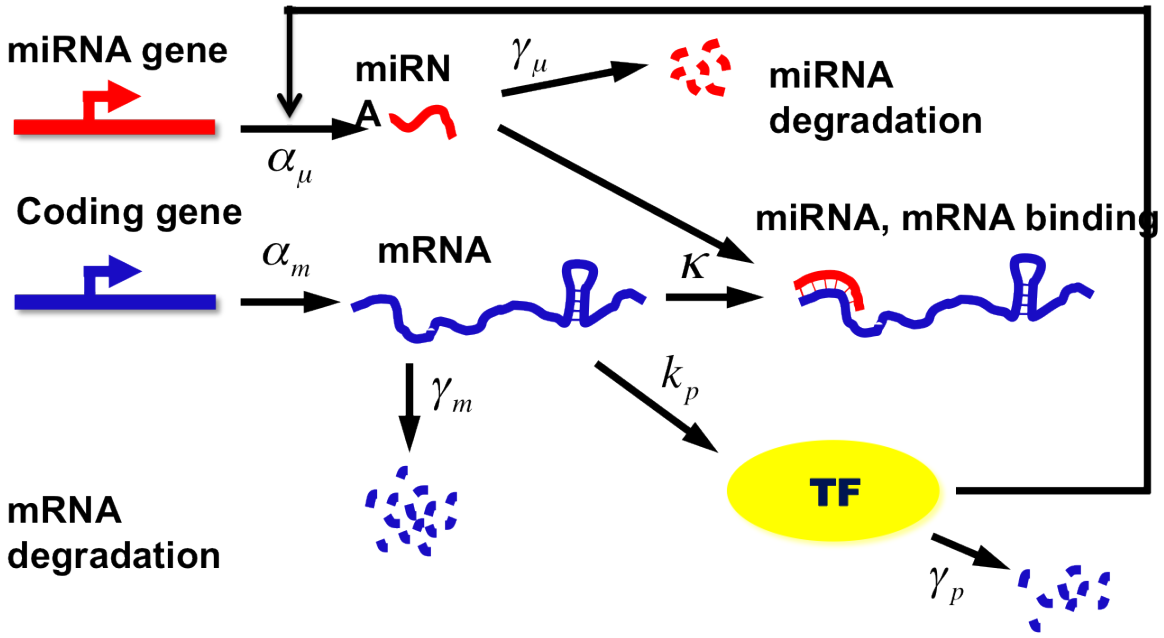


FIGURE 3.1: Schematic illustration of the miRNA-mediated negative feedback loop.

Figure 3.1 shows a schematic diagram of the generic miRNA based feedback network. The component processes of translational regulation by miRNA have been described by Levine et al. and can be generalized to include the negative feedback. A key parameter in this model the probability with which miRNA is co-degraded with the mRNA in the processed state. Considering the limit where miRNA and mRNA interact as an irreversible second-order process that forms a RISC complex at a constant rate κ, yields the "sequestration model ". Importantly, we note that the suppression of translation following miRNA/mRNA interaction is relatively rapid [64], justifying our assumption that the mRNA/RISC complex is effectively incapable of translation. On the other hand, assuming that the miRNA is released and available for reuse leads to a model where miRNAs act catalytically to suppress translation leads to a second model, which we term the "kinetic suppression model". Below, we

analyze the steady-state and noise properties of the miRNA-based feedback network in the two limiting cases and postpone the discussion of the more complete model to a later publication.

3.1 Sequestration Model

A large body of evidence suggests that gene expression is inherently stochastic in nature [7, 8, 15–21], with both intrinsic fluctuations generated by the noisy timing of individual chemical reactions and extrinsic fluctuations due to environmental and other cell-extrinsic factors. Assuming that the intermediate states of the miRNA-mRNA complex are at steady-state, the phase space of the network is characterized by the following three variables, the mRNA number, m, the miRNA number, μ and the protein number, p. The probability of having m mRNA, μ miRNA and p protein molecules at time t thus satisfies the following master equation:

$$\begin{aligned}
\frac{d}{dt}(m,\mu,p) &= \alpha_m P(m-1,\mu,p) + (m+1)\gamma_m P(m+1,\mu,p) \\
&+ g(p)P(m,\mu-1,p) + (\mu+1)\gamma_\mu P(m,\mu+1,p) \\
&+ mk_p P(m,\mu,p-1) + (p+1)\gamma_p P(m,\mu,p+1) \\
&+ (m+1)(\mu+1)\kappa P(m+1,\mu+1,p) \\
&- \{\alpha_m + m\gamma_m + g(p) + \mu\gamma_\mu + mk_p + p\gamma_p + m\mu\kappa\}P(m,\mu,p) \quad (3.2)
\end{aligned}$$

where $P(m,\mu,p)$ is the joint probability for mRNA, miRNA and protein numbers to be m,μ and p respectively. The steady state joint distribution, $P(m,\mu,p)$ cannot be solved analytically. However, experimentally accessible variables are often not the entire distribution, but the mean molecule numbers and their variance. A generating function approach can be generally used to derive these moments, but the presence of the nonlinear term, $\kappa m\mu$, means that the moment equations do not close. However, some progress can be made by multiplying this master equation in turn by m, μ

and p, and summing over all possible m, μ and p, to obtain the familiar mass action equations:

$$\begin{cases} \dfrac{d\langle m\rangle}{dt} = \alpha_m - \gamma_m\langle m\rangle - \kappa\langle m\mu\rangle \\ \dfrac{d\langle \mu\rangle}{dt} = \alpha_\mu - \gamma_\mu\langle \mu\rangle - \kappa\langle m\mu\rangle \\ \dfrac{d\langle p\rangle}{dt} = k_p\langle m\rangle - \gamma_p\langle p\rangle. \end{cases} \tag{3.3}$$

The angle brackets represent the average values of a large ensemble of different realization of these stochastic processes. At steady-state, the average value over large population equals the mean value over the time, showing correspondence between the mass-action and the mean-field models.

3.1.1 miRNA-based Feedback Introduces an Expression Threshold

These mass action equations cannot be solved analytically. In the following, we will explore the general steady-state properties and the nature of intrinsic fluctuations within this negative feedback circuit. For concreteness, we fixed some of the parameters based on experimental observations [69–72]. Specifically γ_m, γ_p have been measured in eukaryotic cells. Across the population, typically $\frac{\gamma_m}{\gamma_p} \sim 10$, i.e., protein life times are significantly longer than their mRNA. Since the rates of degradation of miRNA have not been extensively measured, we assumed $\gamma_m = \gamma_\mu$. In order to make analytic progress, we derive a mean-field model by assuming that the miRNA and mRNA numbers are uncorrelated. Then, the nonlinear term in Eq. B.4 factorizes to yield $\langle m\mu\rangle = \langle m\rangle\langle\mu\rangle$. This allows us to simplify the equations and obtain steady state solutions for the mean-field equations. Under this approximation, mRNA production can be treated as a birth-death process, with a birth rate α and an effective degradation rate $\gamma_m^* = \gamma_m + \kappa\langle\mu\rangle$. Since miRNA-mediated suppression involves an intermediary species, another natural control parameter is the rate of mRNA synthesis

itself. As such, this serves as a proxy for upstream control factors, such as environmental signals, developmental events etc., that engage transcriptional machinery to initiate synthesis of the TF.

As shown in Figure 3.2, the mean mRNA number and the mean TF number exhibit a threshold-linear behavior as a function of the mRNA transcription rate α_m in this negative feedback loop. This has been previously shown for the case when miRNA serves to repress translation in feed-forward fashion [73] and also appears to be qualitatively operant in a similar fashion even when the miRNA acts in a feedback mode. Thus, in either case, miRNAs serve to impose an expression threshold allowing cells to buffer against environmental fluctuations.

A second natural control parameter is the peak transcriptional rate of miRNA, σ. When it is very small relative to mRNA production (α), (as in Figure 2 for σ=0.01), the miRNA production is very small, almost negligible. Thus, the system can be treated as the simplest case with the property of linear relationship between gene product and α_m: $\langle m \rangle = \alpha_m/\gamma_m, \langle p \rangle = \alpha_m * k_p/(\gamma_m\gamma_p)$.. For larger σ, depending on the value of transcription rate of α_m compared to the peak transcriptional rate of miRNA, σ, the system can be classified into three regimes: the repressed regime ($\alpha_m \ll \sigma$), the crossover regime ($\alpha_m \approx \sigma$) and the expressing regime ($\alpha_m \gg \sigma$). While in the repressed regime, mRNA synthesis is strongly repressed by miRNA, keeping overall TF levels very low. This threshold ensures that only strong enough signals, which can drive the α_m value to the expressing regime can trigger the synthesis of gene products.

Increasing the strength of the negative feedback, κ, leads to a sharper crossover between the repressed regime and the expressing regime until it reaches a saturating value as shown in Figure 3.3. Beyond the threshold ($\alpha \approx \sigma$), i.e., in the expressing regime, there is a linear relationship between the number of mRNA (m), TFs (p) and α_m and the slope of the $\langle m \rangle$-α_m curve represents the sensitivity of the system

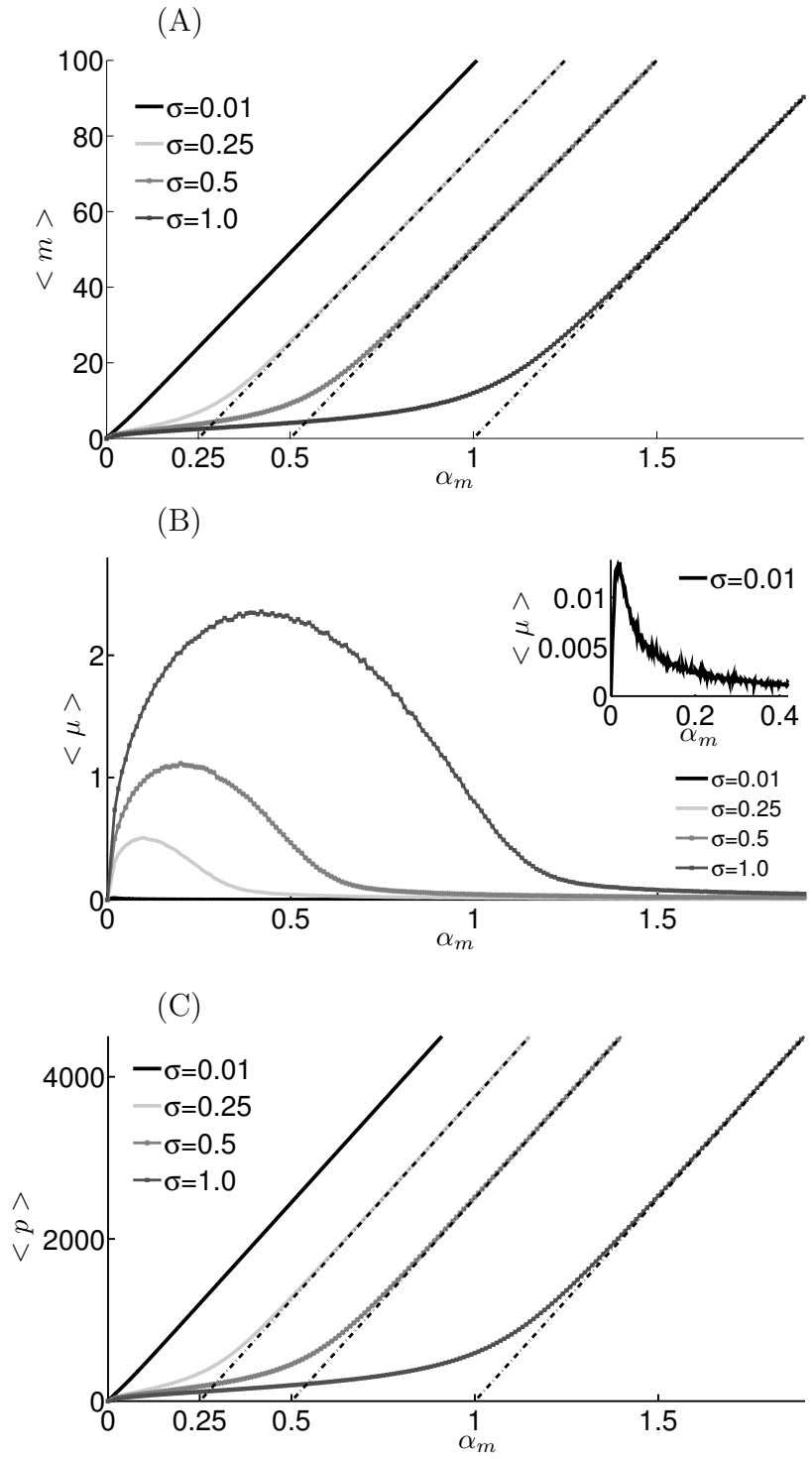


FIGURE 3.2: miRNA-based feedback introduces an expression threshold in sequestration model. The mean value of mRNA (A), miRNA (B) and protein (C) are shown as a function of α_m for four different values of the peak miRNA transcriptional rate (σ). When σ is extremely small, m and p are proportional to α_m because there is almost no miRNA synthesis. For larger σ, the miRNA-based feedback introduces a threshold at $\alpha_m \approx \sigma$. All the asymptotic lines are parallel to each other. $\gamma_m = \gamma_\mu = 0.01$, $\gamma_p = 0.002$, $\kappa = 1.0$, $k_p = 0.1$ and $k_d = 200.0$.

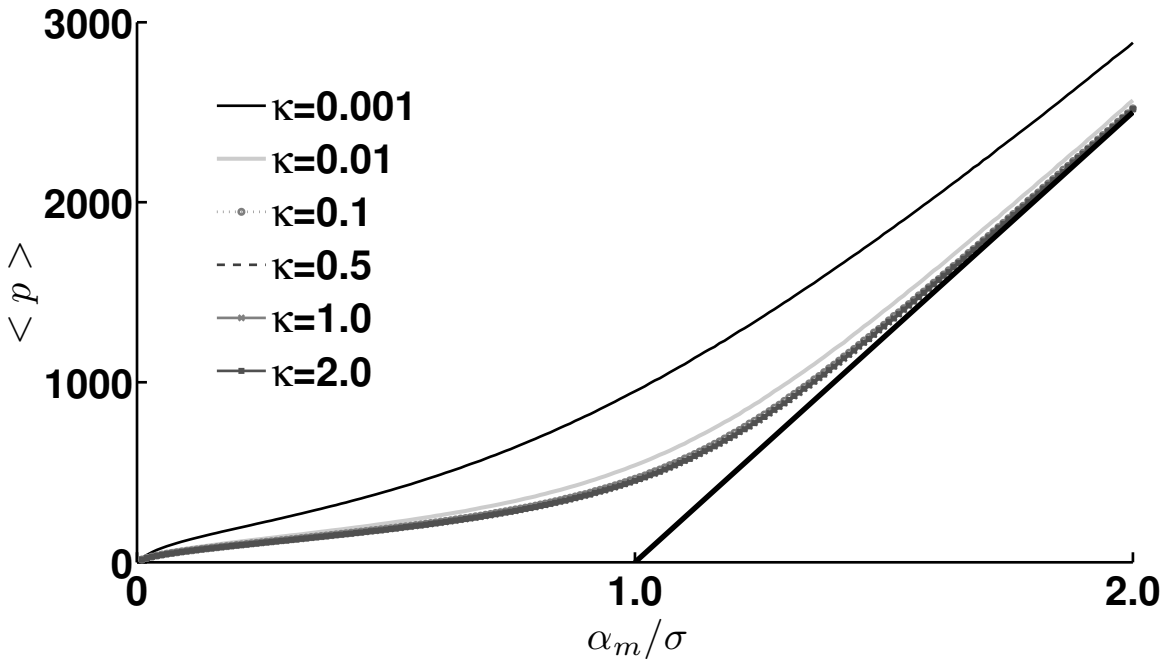


FIGURE 3.3: Increasing miRNA-mediated feedback strength sharpens the expression threshold. Mean protein numbers as a function of upstream transcription rate. The different curves show how threshold expression depends on the strength of miRNA/mRNA association rate. Mean protein expression shows a crossover regime from low expression to linear expression with increasing transcription rate. As the strength of miRNA/mRNA association, κ, increases, the threshold becomes sharper. The solid curve is a perfect threshold-linear behavior. The parameters are set as $\gamma_m = \gamma_\mu = 0.01, \gamma_p = 0.002, \sigma = 0.5, k_p = 0.1, k_d = 200.0$.

in response to external signals. Both m and p are proportional to α_m: $\frac{d\langle m\rangle}{d\alpha_m} = 1/\gamma_m$, $\frac{d\langle p\rangle}{d\alpha_m} = k_p/(\gamma_m\gamma_p)$. The linearized relationship between $< p >$ and α_m is plotted as black line in Figure 3.3, denoted by $< p >= \frac{k_p}{\gamma_m\gamma_p}(\alpha_m - \sigma)$. While the mRNA and protein abundance show a threshold-linear behavior, the miRNA levels within this feedback circuit exhibit a non-monotonic behavior (Figure 3.2B): at low mRNA synthesis rate, there is very little synthesis of the TF and consequently, the synthesis of the miRNA is low. At high synthesis rates, most of the miRNAs stoichiometrically combine with mRNAs and accumulate in the translationally inactive pool. Since the mRNA is in excess, all the miRNAs are consumed, leaving a large number of translationally competent mRNAs that can engage in synthesis of the TF protein.

Moreover, since this pool does not feel the effect of miRNA based repression, the effective degradation rate of this excess pool of mRNAs is the native degradation rate of the mRNA, γ_m. We note that the non-monotonicity of mean miRNA number and the threshold linear behavior of the mean protein levels are observed for a wide variety of parameter combinations.

3.1.2 miRNA based Negative Feedback Amplifies Noise in the Sequestration Model

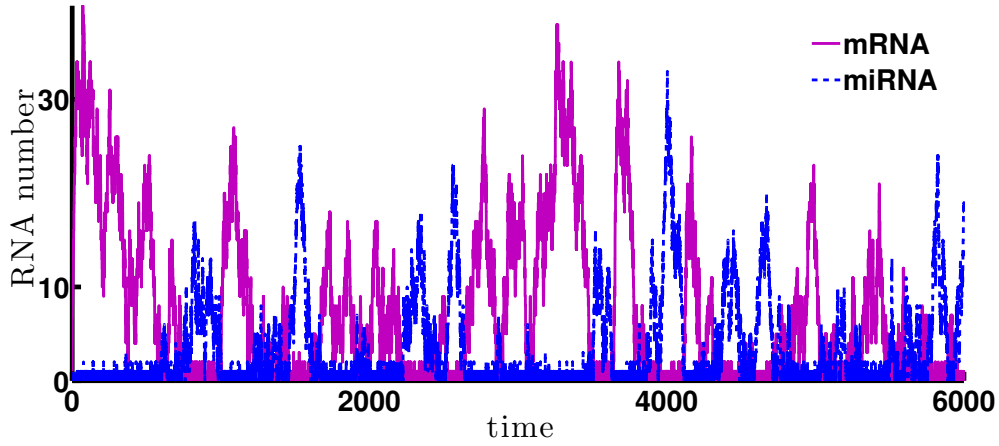


FIGURE 3.4: miRNA and mRNA levels are anti-correlated. Temporal evolution of mRNA (black) and miRNA (gray) in a Monte-Carlo simulation of the sequestration model using the Gillespie algorithm. The parameters are set as $\alpha_m = 1, \gamma_m = \gamma_\mu = 0.01, \gamma_p = 0.001, \kappa = 1.0, \sigma = 1.0, k_d = 200.0, k_p = 0.1$. The anti-correlation suggests that assuming $\langle m\mu \rangle = \langle m \rangle \langle \mu \rangle$ is not valid.

A conventional interpretation of negative feedback motifs in genetic circuits is that they generally serve to decrease expression noise, suppressing fluctuations while maintaining near constant mean levels of the components. However, depending on the timescales of the various component process (RNA polymerase binding, repressor multimerization and binding etc.), noise levels can moderately increase relative to unregulated systems with increasing negative feedback strength [74–78]. These insights have been derived from examining the behavior of genetic circuits that involve

genes that code for repressor proteins that block their own transcription by binding to promoter (or promoter-proximal) regions of their own genes. In order to assess the impact of miRNA based negative feedback, we next examined the intrinsic noise properties of our network using the Fano factor as a measure of the fluctuations. The Fano factor [79] is typically independent of system volume and measures how much the size of internal fluctuations deviates from what is expected from Poisson statistics, for which the Fano factor equals one.

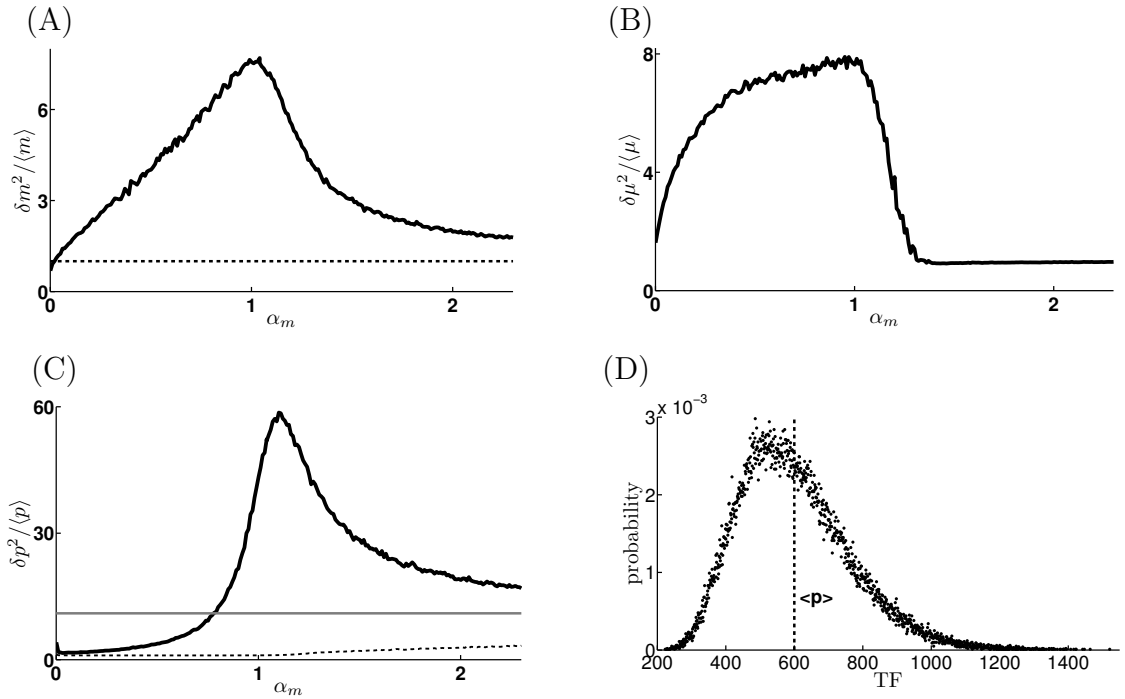


FIGURE 3.5: Negative feedback amplifies expression noise in the sequestration model. (A)-(C). Fano factor of mRNA, miRNA and protein are plotted versus α_m. Solid lines represent the simulation results and dashed lines represent the analytical calculation from the mean field model. The mean field model cannot be used to describe the system around the threshold, where the mRNA and miRNA levels are comparable. The solid straight line in (C) represents the asymptote for protein Fano factor value for large α_m. (D). Histogram protein numbers at steady state for $\alpha_m = 1.0$: the mean values are $\langle p \rangle = 600$, $\frac{\delta p^2}{\langle p \rangle} = 43.2$. The protein distribution shows a long tail, which suggests that while mean values may be kept low, protein numbers can exhibit large values across the population allowing the transcription factor to act at promoters with widely different sensitivities.

We first consider how noise properties depend on environmental control signals

that are encoded in the parameter α_m, the synthesis rate of the target mRNA. The mean field model states that steady state mRNA number should reach a Poisson distribution with Fano factor, (the ratio of the mRNA number variance and mean) $\frac{\delta m^2}{\langle m \rangle} = 1$. The Fano factor of the TF can be readily calculated to be

$$\frac{\delta p^2}{\langle p \rangle} = 1 + \frac{k_p}{\gamma_p + \gamma_m + \kappa \langle \mu \rangle}. \tag{3.4}$$

As expected, the Fano factor for protein numbers is larger than 1, because each mRNA leads to the synthesis of a burst of proteins before degradation, with a "burst" size of $k_p/(\gamma_p(1+\gamma'_m))$ [75]. In the mean-field model, the effect of the miRNA is to increase the degradation rate of mRNA, leading to a smaller burst size and lower variability. In order to validate our assumptions, we used the Gillespie algorithm[1] to perform stochastic simulations of the full model. We find that the mRNA and miRNA levels are strongly anti-correlated (Figure 3.4), with periods of high mRNA levels corresponding to low miRNA levels and vice versa, as has been widely noted in experiment [80]. This can be understood as follows: increases in mRNA levels lead to the synthesis of the TF, which then leads to the transcription of the miRNA. These miRNA molecules can now bind to the mRNA, and move them to the translationally inactive pool, resulting in a net loss of both mRNA and miRNA. However, if the mRNA levels are high to begin with, most of the miRNA molecules are saturated, with only an excess of mRNA levels remaining. On the other hand, a large fluctuation in miRNA levels reduces mRNA numbers stoichiometrically, leaving excess miRNA free. These observations imply that our assumption that the nonlinear term can be factorized is invalid and that the effective degradation rate γ^*_m must include the effects of this correlation between mRNA and miRNA levels. Moreover, since the correlation is negative, we would expect that the effective degradation rate be smaller

[1] Gillespie algorithm is a discrete-event simulation algorithm that produces single realizations of the stochastic process that are in exact statistical agreement with the master equation.

than when there is no miRNA-based translational repression. Consequently, we should see an increase in the effective burst size, and hence a larger Fano factor for protein fluctuations than without miRNA-mediated repression.

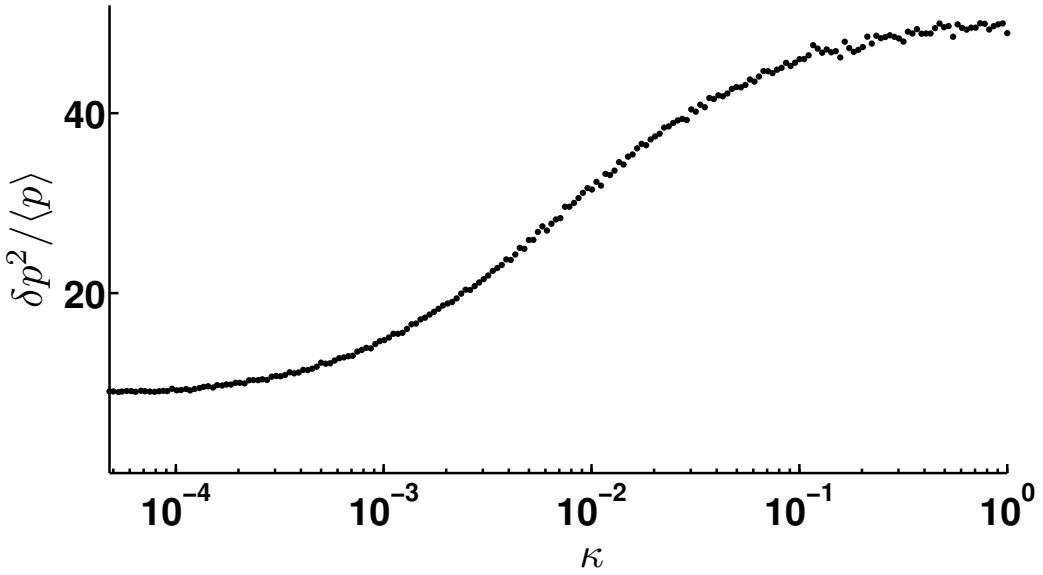


FIGURE 3.6: Expression variability increases with negative feedback strength. Fano factor of TF numbers is plotted versus the strength of bimolecular miRNA-mRNA association rate, κ for sequestration model. The transcription rate, $\alpha_m = 1.2$, at the cross-over point in Figure 3.3. Other parameters were set at $\gamma_m = \gamma_\mu = 0.01$, $\gamma_p = 0.001$, $\sigma = 1.0$, $k_p = 0.1$, $k_d = 100.0$. For κ=0, there is no miRNA-mediated translational repression, and Fano factor is the lowest. The Fano factor increases with κ, until it reaches a saturating value.

To verify our intuitive observations, we conducted large scale simulations to study the noise properties of the network as a function of key control parameters. Surprisingly, the Fano factors of all three components showed non-monotonic behavior as α_m was increased, with peak Fano factors well in excess of what is predicted by the mean field model as well as the case of the unregulated gene (Figure 3.5). Furthermore, the non-monotonicity is obtained over a wide range of values for the negative feedback strength κ, peak miRNA transcription rate, σ and protein synthesis rates k_p (data not shown). Moreover, we find that the mean field model is only applicable in certain limiting regimes. If the mRNA synthesis rate, α_m is much smaller

or much larger than σ, the mean field model can capture the noise properties of the system. In the case where α_m is low and $m \ll \mu$, the mRNA is strongly repressed by miRNA. Because of the excess miRNA, $\langle\mu\rangle$, the denominator is large in Eq. 3.4 is large and thus the Fano factors of mRNA and TF are small. On the other hand, when $m \gg \mu$, the miRNA number is strongly repressed due to the binding of mRNA and miRNA, leaving the excess mRNA translationally active. Thus, the effect of miRNA regulation diminishes and can be neglected. In this case, the Fano factor of both mRNA and protein number asymptotically tends to the value where there is effectively no miRNA-mediated repression, i.e., $\frac{\delta m^2}{\langle m\rangle} = 1, \frac{\delta p^2}{\langle p\rangle} = 1 + \frac{k_p}{\gamma_p+\gamma_m}$, which is plotted as the asymptotic line in Figure 3.5. This is also in accordance with Eq. 3.4 while $\kappa\mu \to 0$. Interestingly, when $\alpha_m \simeq \sigma$, i.e., the synthesis rates of mRNA and miRNA are comparable, the Fano factors of mRNA and TF numbers are much larger than the mean field prediction as can be seen in Figure 3.5. This amplification is due to the anti-correlation between the mRNA and miRNA. Noise in stoichiometrically coupled systems such as miRNA-based gene regulation has been studied earlier [81–84]. These studies suggest the existence of a crossover regime characterized by enhanced stochastic fluctuations. This *near-critical* behavior is reminiscent of the critical fluctuations near phase transitions [85]. Accordingly, we find that the TF number distribution shows a long tail (Figure 3.5D), which suggests that while negative feedback by miRNA keeps the mean TF number low, there can be large temporal or population variation. This variation naturally arises because the effect of miRNA is to reduce the overall number of mRNA. Thus any surviving mRNA has to rapidly engage in a burst of translation before it is consumed either be mRNA decay or binding to a miRNA-loaded RISC complex. We will examine the effect of this variation in a gene cascade in a later section. Moreover, we note that such large fluctuations in the protein levels can be obtained even without transcriptional

bursting, i.e. for promoters that are continuously active.

We next examined the noise properties as a function of the strength of negative feedback, in this case represented by the bimolecular association rate, κ. We focused on the region where the noise properties of the network are amplified relative to no miRNA based feedback. Thus, we fixed $\alpha = 1.2$, a value around which the Fano factor of TF levels peaks (Figure 3.5C), to see how the Fano factor varies with the strength of negative feedback. While $\kappa = 0$, there is no negative feedback, the noise in the system is at a minimum as shown in Figure 3.6, which is close to that of an unregulated system. As κ increases, the interaction between mRNA and miRNA is strengthened and the noise in TF numbers become larger. Then finally, when κ is very large, the Fano factor saturates to an asymptotic value, much larger than that for an unregulated gene. The limiting values of the Fano factor with increasing feedback strength are distinct from the case of negative feedback mediated by a protein repressor, where increasing the strength of negative feedback (i.e. the affinity of the repressor to the gene promoter) for a fixed transcription rate increases the Fano factor over what would be expected for an unrepressed gene while tending to a lower value for weak and intermediate feedback strengths [74, 76–78]

However, we note that the origins of the increased fluctuations are similar. For protein-based repression with a given transcriptional rate, the high affinity of the repressor implies that most of the time, the gene is inactive with few mRNAs being transcribed. Upon brief dissociations of the repressor from the gene, transcription can commence and result in bursts of synthesis of both mRNA and the repressor protein. Thus, the effective timescale of these bursts is determined by the dissociation rate of the repressor. For repressors with weak affinity to the promoter, this additional noise source vanishes as the repressor-gene interaction approaches steady-state and the Fano factor tends to an asymptotic value $\frac{\delta p^2}{\langle p \rangle} = \frac{b}{1+\eta} + 1 = \frac{k_p}{\gamma_p + \gamma_m} + 1$,

the burst size, $b = k_p/(\gamma_m + \gamma_p)$ is the average burst size and $\eta = \gamma_p/\gamma_m$. On the other hand, in miRNA mediated feedback repression, increasing the miRNA/mRNA association rate lowers the overall mRNA levels, leading to rare bursts of synthesis, while reducing it approximates a situation with no feedback regulation. In summary, in the sequestration model, miRNA-based negative feedback in physiologically relevant regimes actually amplifies the noise relative to what would be expected either in case of a protein-repressor mediated feedback [75] or the case where there is no feedback.

3.2 Kinetic Suppression Model

$$\begin{cases} \dfrac{dm}{dt} = & \alpha_m - m\gamma_m + \eta_m \\ \dfrac{d\mu}{dt} = & g(p) - \mu\gamma_\mu + \eta_\mu \\ \dfrac{dp}{dt} = & f(m,\mu) - p\gamma_p + \eta_p, \end{cases} \tag{3.5}$$

where $g(p) = \sigma p^2/(p^2 + k_d^2)$ and $f(m,\mu) = \frac{mk_p}{1+\beta\mu^2}$.

Given the abundance of miRNA and the diversity of targets for a single miRNA, under some conditions, translational regulation by miRNA can be considered to act catalytically, i.e., miRNAs bind to mRNA at the regulating sites and repress translation initiation or elongation with the number of miRNA itself being unchanged. This scenario is valid under conditions of relatively weak miRNA/mRNA binding and large miRNA concentrations. We represent this catalytic mode of action by assuming that the miRNAs act in a Michaelis-Menten fashion to repress translation. The transcription of mRNA and miRNA as well as the degradation of mRNA, miRNA and TF protein have the same form as in the previous (sequestration) model. We model the translational repression by taking translation rates to be decreasing Hill functions of the number of miRNA regulatory molecules. i.e., $k_p^* = f(m,\mu) = \frac{mk_p}{1+\beta\mu^2}$.

In the circuit analyzed here, we denote β as the strength of negative feedback representing the effect of the control of TF synthesis by the miRNA. Given the abundance of the components, we model the post-transcriptional regulation through miRNAs using mass action equations with three molecular species: the number of miRNA molecules μ, the number of target mRNA molecules m, and the number of regulated TF molecules, p [73, 83, 84, 86–88].

The effect of intrinsic noise is included by Langevin terms, η_m, η_μ and η_p denoting the intrinsic fluctuations of the mRNA, miRNA and protein respectively, that describe the statistical fluctuations in the underlying biochemical reactions [89]. The dynamics of these processes can then be described by the following Langevin equations:

The Langevin terms η_i model intrinsic noise by treating the birth and death of the different species as independent Poisson processes, representing the stochastic creation and destruction of mRNA, miRNA and TF. We have dropped the cross-term $\eta_{m,\mu}$ since we assume that miRNAs act catalytically, where these two levels are uncorrelated.

The Langevin terms are characterized within the linear noise approximation [89] by two-point time correlation functions:

$$\begin{cases} \langle \eta_m(t)\eta_m(t')\rangle &= (\alpha_m + m\gamma_m)\delta(t-t') \\ \langle \eta_\mu(t)\eta_\mu(t')\rangle &= (g(p) + \mu\gamma_\mu)\delta(t-t') \\ \langle \eta_p(t)\eta_p(t')\rangle &= (f(m,\mu) + p\gamma_p)\delta(t-t'). \end{cases} \tag{3.6}$$

The linear-noise approximation is a good approximation even for nonlinear systems with small fluctuations. This is confirmed by the simulation results which use the exact Gillespie algorithm (see Figure 3.7). In order to obtain expressions for the noise properties of the different species, we find the steady state solution of the model and then linearize around this to compute the response of the variables m, μ and p to

the Langevin forcing terms η_m, η_μ and η_p. In the linear approximation, the steady state is also the mean value. So $m =< m > +\delta_m$, $\mu =< \mu > +\delta_\mu$ and $p =< p > +\delta_p$. Linearizing the Langevin equations around their steady states, we obtain:

$$\frac{d}{dt}\begin{pmatrix} \delta_m \\ \delta_\mu \\ \delta_p \end{pmatrix} = \begin{pmatrix} -\gamma_m & 0 & 0 \\ 0 & -\gamma_\mu & \frac{\partial g}{\partial p} \\ \frac{\partial f}{\partial m} & \frac{\partial f}{\partial \mu} & -\gamma_p \end{pmatrix}\begin{pmatrix} \delta_m \\ \delta_\mu \\ \delta_p \end{pmatrix} + \begin{pmatrix} \eta_m \\ \eta_\mu \\ \eta_p \end{pmatrix} \tag{3.7}$$

We now transform these linearized equations into Fourier space, with $\hat{\delta}_i(\omega)$ and $\hat{\eta}_i(\omega)$ corresponding to the temporal variables $\delta_i(t)$ and $\eta_i(t)$ where i equals to m,μ or p in the spectral domain. Thus,

$$\begin{pmatrix} \hat{\delta}_m(\omega) \\ \hat{\delta}_\mu(\omega) \\ \hat{\delta}_p(\omega) \end{pmatrix} = M^{-1}\begin{pmatrix} \hat{\eta}_m(\omega) \\ \hat{\eta}_\mu(\omega) \\ \hat{\eta}_p(\omega) \end{pmatrix} \tag{3.8}$$

where

$$M = \begin{pmatrix} \gamma_m + i\omega & 0 & 0 \\ 0 & \gamma_\mu + i\omega & -\frac{\partial g}{\partial p} \\ -\frac{\partial f}{\partial m} & -\frac{\partial f}{\partial \mu} & \gamma_p + i\omega \end{pmatrix} \tag{3.9}$$

Using the Wiener-Khinchin theorem, the spectral density

$$\left|\hat{\eta_i}(\omega)\right|^2 = 2\pi \int_{-\infty}^{\infty} \left\langle \eta_i(t)\eta_i^*(t-\tau)\right\rangle e^{-i\omega t} d\tau, \tag{3.10}$$

we can get:

$$\begin{cases} \left|\hat{\eta_m}(\omega)\right|^2 = 2\pi(\alpha_m + m\gamma_m) \\ \left|\hat{\eta_\mu}(\omega)\right|^2 = 2\pi(g\,(p) + \mu\gamma_\mu) \\ \left|\hat{\eta_p}(\omega)\right|^2 = 2\pi(f\,(m,\mu) + p\gamma_p) \end{cases} \tag{3.11}$$

Based on these expressions, we can obtain $\left|\hat{\delta_m}(\omega)\right|^2$, $\left|\hat{\delta_\mu}(\omega)\right|^2$ and $\left|\hat{\delta_p}(\omega)\right|^2$ using Eq. 3.8 (Appendix Equation B.9). Next, using the relation $\left\langle f\,(t)^2\right\rangle = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\left|\hat{f}(\omega)\right|^2 d\omega$

to inverse transform back to the time domain, we can get the exact solution of the variance $\langle \delta m^2 \rangle$, $\langle \delta \mu^2 \rangle$ and $\langle \delta p^2 \rangle$. These expressions are somewhat lengthy and we have omitted them for brevity.

3.2.1 Negative feedback represses noise in kinetic suppression model

We now analyze the properties of the number fluctuations of different species in the network. Because in the kinetic suppression model, miRNAs act catalytically to repress the translation of TF, mRNAs are always Poisson distributed. We now focus on the fluctuations in the number of TFs which can be quite different depending on the negative feedback and promoter strengths (Figure 3.7). When the negative feedback strength β, equals zero, the Fano factor of TF numbers equals $1 + \frac{k_p}{\gamma_p + \gamma_m}$. With sufficient production of mRNA (α_m not too small), the Fano factor slightly increases from the zero-feedback value for small values of β, i.e. weak feedback strengths. For increasing β, the Fano factor decreases very rapidly to a small value over a long range of β. Note the correspondence between the expressions derived from the linear noise approximation and the full model using stochastic simulations. For very large values of β, denoting large negative feedback, the analytic approximation breaks down since the mean TF levels are small and the relative fluctuations are high. In this case the Fano factor asymptotes to a value above what would be expected for the no repression case.

We next study the effect of varying the strength of negative feedback on the TF number fluctuations as a function of the promoter strength α_m (Figure 3.8). We find that for weak promoters, the effect of negative feedback is to continuously decrease the variability. This is in part due to the fact that weak mRNA production implies an even weaker synthesis of the miRNA itself. However, as the promoter strength is increased, we find the emergence of a peak in the protein number fluctuations at very weak feedback strength, subsequent suppression and then increasing

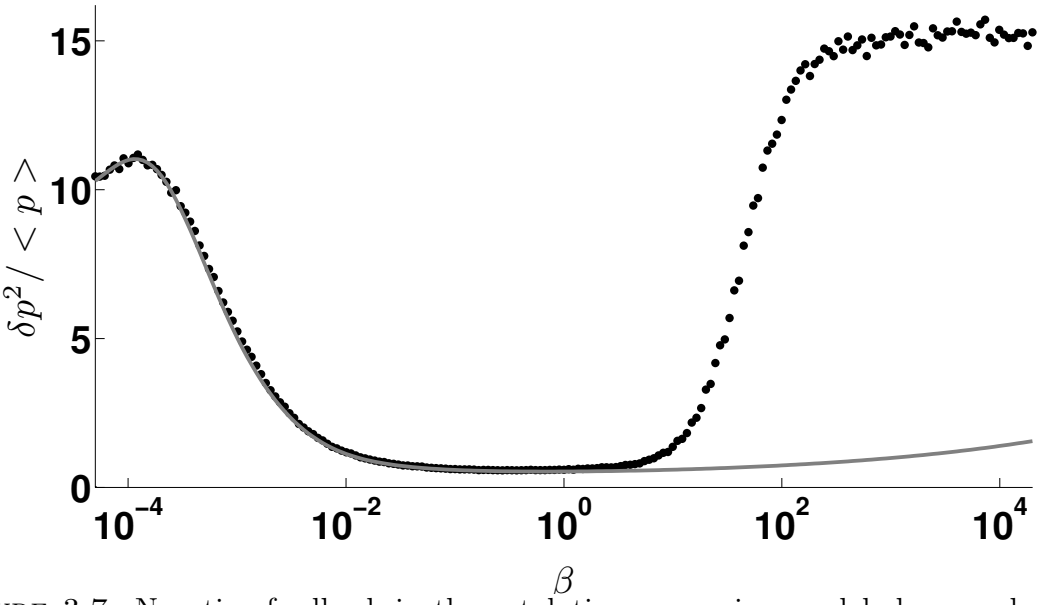


FIGURE 3.7: Negative feedback in the catalytic suppression model shows reduced expression variability over a large range of feedback strength. The Fano factor of TF is plotted versus β for kinetic suppression model. The solid line is the analytical solution from the linear noise approximation method, dots represent results of Gillespie algorithm simulation. Other parameters are set as: $\alpha_m = 1.2, \gamma_m = \gamma_\mu = 0.01, \gamma_p = 0.001, \sigma = 1.0, k_p = 0.1, k_d = 100.0$. Although there is a temporary increase of the noise for low feedback strengths , it decreases very rapidly over a long range. For large values of β, linear noise approximation breaks down and single molecule effects dominate. Note that the variability is lower than in the sequestration model.

fluctuations for very strong feedback. This latter increase is due to the fact that the fluctuations are large and the small noise approximation breaks down. A substantial amount of experimental evidence suggests that miRNA based translational repression usually serves to reduce mean target protein levels in a modest fashion (2-4 fold reduction) [90]. Thus, we anticipate that the large negative feedback regime considered here is more for the sake of completeness and not meant to represent any physiological situation in general.

We next compare the noise properties of the feedback network depending on the mode of action of the miRNA. In sequestration model, where miRNA and mRNA pair stoichiometrically and are rendered translationally incompetent, the relevant parameter that determines the extent of negative feedback is the affinity of the

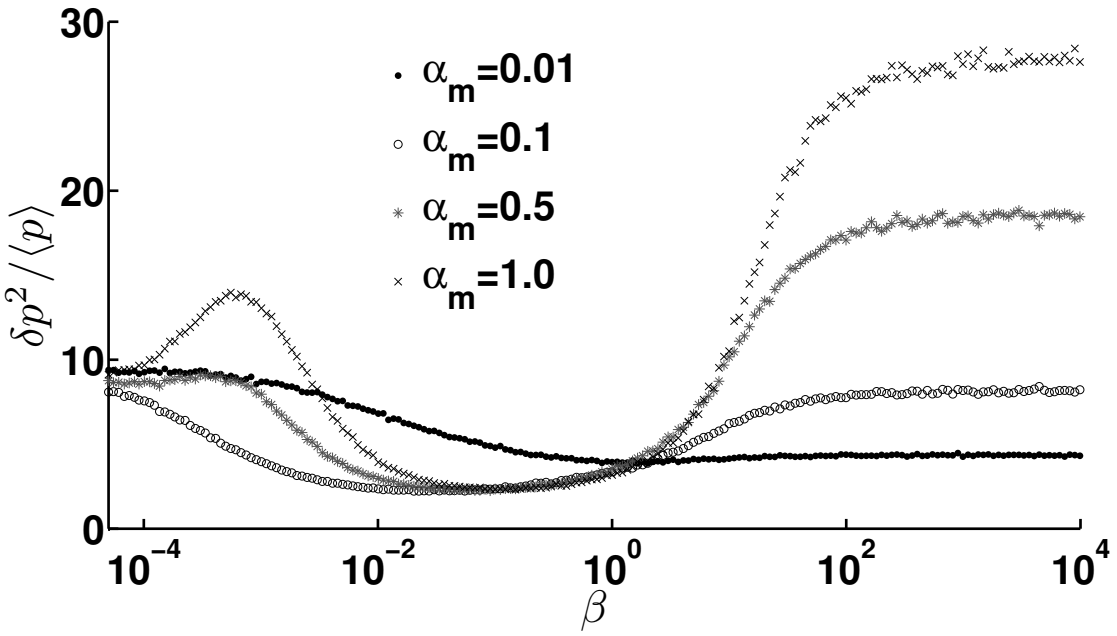


FIGURE 3.8: Negative feedback strength in the catalytic suppression model affects expression variability for different promoter strength. The relationship between the Fano factor of TF and β with four different values of α_m. $\gamma_m = \gamma_\mu = 0.01, \gamma_p = 0.002, \sigma = 0.5, k_p = 0.1, k_d = 200.0$.

bimolecular interaction, κ. On the other hand, in the kinetic suppression model, the strength of mRNA translation is modulated by the parameter, β. In order to compare the noise properties of these two cases, we set all the other parameters the same for these two models and then chose the feedback strength β and κ so that they result in similar mean value of TF but different noise terms (Figure 3.9A). As shown in Figure 3.9, the two genetic circuits produce almost the same mean value of the TF, but the population distributions of the TF in the two cases are quite different. In the kinetic model, TF numbers are tightly restricted near the mean value, but distribution of TF in the sequestration model is broader and is characterized by long tail which indicates that there can be large bursts in the number of TF.

3.3 The Impact of Transcription Factor Fluctuations in Gene Networks

Considering that the miRNA-regulated protein is a TF, the qualitatively different distributions of TF obtained under different modes of feedback regulation by miRNA can affect the transcription of genes that are also regulated by the same TF. Concretely, we envision the genes downstream of the transcription factor *pitx3*, would exhibit different patterns of gene expression depending on population levels of *pitx3*. Here, we examine how qualitatively different modes of miRNA based negative feedback affect the transcription of such downstream genes using a simple regulatory cascade as shown in (Figure 3.10). We assume the transcription rate of downstream genes is a generic Hill function: $\alpha_{m2} = \frac{\epsilon \mathrm{TF}^n}{\mathrm{TF}^n + {K_D}^n}$ as shown in Figure 3.11A. In this study, we choose $n = 10$ because with large Hill coefficient, the promoter essentially acts like a switch. Qualitatively similar results obtain for smaller n. We assume that the downstream protein ($p2$) is synthesized at a rate k_{p2} and degraded at a rate γ_{p2}, its mRNA is degraded at a rate γ_{m2}. We assume that three representative target genes with different dissociation constants, that denote varying promoter strengths, to see how the variability of the TF levels affect the downstream protein distribution (Figure 3.11A).

For a highly-sensitive downstream gene, (small disassociation constant), the mean TF number is in the saturation regime. Thus, TFs that are either stoichiometrically or catalytically suppressed are equally effective in driving the expression of the downstream gene (Figure 3.12A) with similar mean values. However, owing to the switch-like behavior of p2 promoter, the population distribution of p2 is broader when driven by the stoichiometrically suppressed TF as compared to the catalytically suppressed one. In particular, we note the presence of a long tail of low p2 expressors. On the other hand, for a downstream gene with a weaker promoter, i.e. dissociation constant

such that the mean TF levels lies in the linear regime of $\alpha_{m2}(TF)$ (the dashed curve in Figure 3.11A), expression of p2 is reduced. However, the p2 distribution is broad and skewed towards high expressors for the case when the TF is stoichiometrically regulated (Figure 3.12B). For the extreme case where the dissociation constant of the downstream gene promoter is much higher than the mean TF levels, TFs that are stoichiometrically regulated by miRNA are able to drive considerable expression (Figure 3.12C) while TFs that are catalytically regulated cannot.

3.4 Discussion

The post-transcriptional control of protein expression in animal cells by micro-RNAs plays an important role in almost every cellular process and changes in their expression may underlie developmental disorders and diseases such as cancer. miRNAs base-pair with seed sequences in the $3'$ UTRs of their target mRNA and block steps in the initiation of transcription, sequestering mRNAs into sites of repression or by accelerating mRNA decay [64]. As a result, miRNAs reduce mRNA and protein abundance, often modestly and sometimes sharply [90, 91]. Genome-wide studies have shown that miRNAs target many transcription factors, which in some cases regulate their own transcription. Given the importance of miRNAs in cellular function an analysis of the impact of miRNA-mediated regulation on the mean levels and fluctuations of genetic circuits is vital. Here, we have shown that depending on the mode of miRNA action, negative feedback by miRNA can have differential impact on the noise levels of protein expression. In particular, if miRNAs act in a stoichiometric fashion, whereby both the target mRNA and miRNA are removed from the population into an inactive pool (Figure 3.1), then negative feedback regulation by miRNA largely amplifies the intrinsic noise in the system, leading to long-tailed distributions of transcriptional factor numbers. Our simulations show that this enhancement of protein number fluctuations is sensitive to environmental factors as

seen in Figures 3.5 and Figure 3.6. However, if miRNAs act catalytically to repress protein synthesis, the net effect is to reduce variability in protein levels, as would be conventionally expected for a negative feedback circuits.

A number of experimental observations justify the distinction of the modes of miRNA action made in our models. Early studies of miRNA effects seemed to reveal that mRNA degradation was minimal but protein expression was reduced consistent with a catalytic mode of regulation [91]. This could result from imperfect seed sequence complementarity between the miRNA and mRNA, the presence of multiple miRNA targets, weak and reversible association of the target mRNA with the RISC machinery, rapid accumulation of the RISC/mRNA complex into P-bodies or accumulation in stress granules accompanied by the release of the miRNA. More recent studies have shown that mRNA degradation is significant [91]. These could arise from a higher degree of complementarity of miRNA seed sequence in the mRNA $3'$ UTR, multiple pairing locations, post-translational modifications of the RISC machinery that enhance binding of miRNA/mRNA and subsequent translational repression and P-body accumulation. These latter effects are best represented mathematically by a sequestration model where both miRNA and mRNA are stoichiometrically degraded. In order to keep the models relatively simple and to gain intuition, we have abstracted many of the intermediate steps, modeling component processes as first and second-order reactions. Such coarse-grained representations have been quite successful in elucidating many aspects of deterministic and stochastic gene networks [9, 10, 75, 92]. We have neglected additional aspects of miRNA biogenesis and function, such as multiple miRNA seed sequences on the same target mRNA, delays in processing mature miRNA from precursor transcripts etc. However, we expect that consideration of these steps would not qualitatively change our results.

Our studies expand the repertoire of miRNA action in gene circuits that govern

cell fate specification and commitment during development, processes where miRNA function was first highlighted. The commitment of cells to specific lineages derives from the coordinated expression of different patterns of genes within a relatively uniform population of cells. These expression patterns are then crystallized by downstream gene networks to result in stable expression of lineage specific genes that is maintained throughout the individual's life. Cell fate choices are often under the control of restricted subsets of upstream transcription factors. How population diversity is achieved from cells that possess identical genomes is a fundamental question of developmental biology. It is well known that genetic circuits with extensive feedback loops, both negative and positive, play an important role in cell fate choices. In particular, feedback imparts a network with multiple steady states, which can denote the multiple cell fates controlled by the network. Moreover, the steady states of feedback circuits are well separated, preventing spontaneous transitions, imparting robustness to the gene circuits controlling cell fates. The role of miRNAs in animal development has been examined in these contexts. Most studies to date have focused on the impact of miRNA-mediated control of the mean protein levels on developmental and cell fate specification circuits [36, 38–41].

However, miRNAs translational repression also shapes the intrinsic variability within developmental gene networks. Noisy gene expression in developmental circuits has the potential to be harmful leading either to arrested development, aberrant positional expression of tissue specific genes or over-representation of specific cell types. miRNAs are thought to tune the fluctuations of protein expression within developmental networks, buffering them against environmental fluctuations. The imposition of an expression threshold by miRNAs renders the network insensitive to small subthreshold variations, preventing stochastic transitions between steady states. This has been directly demonstrated in Drosophila, where the miRNA, miR-7, is required to maintain normal gene expression and sensory organ fate determination under fluc-

tuating temperature conditions [40] by buffering the levels of its downstream target, the transcriptional repressor, *yan*.

Our modeling studies suggest a new role for miRNA-based feedback regulation, namely, by modulating the levels of TFs at the level of translational repression, miRNAs can drive large fluctuations in TF levels across the population. In turn, these fluctuations can drive the expression of different constellations of genes across the population, thereby allowing the expression of multiple cellular phenotypes in a uniform precursor population. Given the extensive complexity of the component processes, cells may be able to tune the manner of miRNA-based feedback, from stoichiometric repression to catalytic repression to tune the level of protein number fluctuations in gene circuits and consequently drive stochastic cell fate choices. A number of recent studies suggest that such tuning may be operant in cells. RISC protein phosphorylation can control the loading of miRNAs [93]. Alternately, the seed site for miRNA binding on the target mRNA may be made more accessible [94]. Thus, cells may control expression noise in miRNA-based negative feedback circuits to determine cell fates in different contexts. In general, cell fate decisions during development are robust, in order to generate reproducible body plans. However, in certain cases, cell fate decisions are made at random, generating cell fate diversity. Diversified cell fates in a homogeneous progenitor population increases the spectrum of responses to environmental stimuli. One example is the choice of Rhodopsin type during photoreceptor differentiation in the Drosophila eye [95]. Thus, miRNA based translational repression may serve as an important mechanism that controls fluctuations of protein number promoting cell fate diversity.

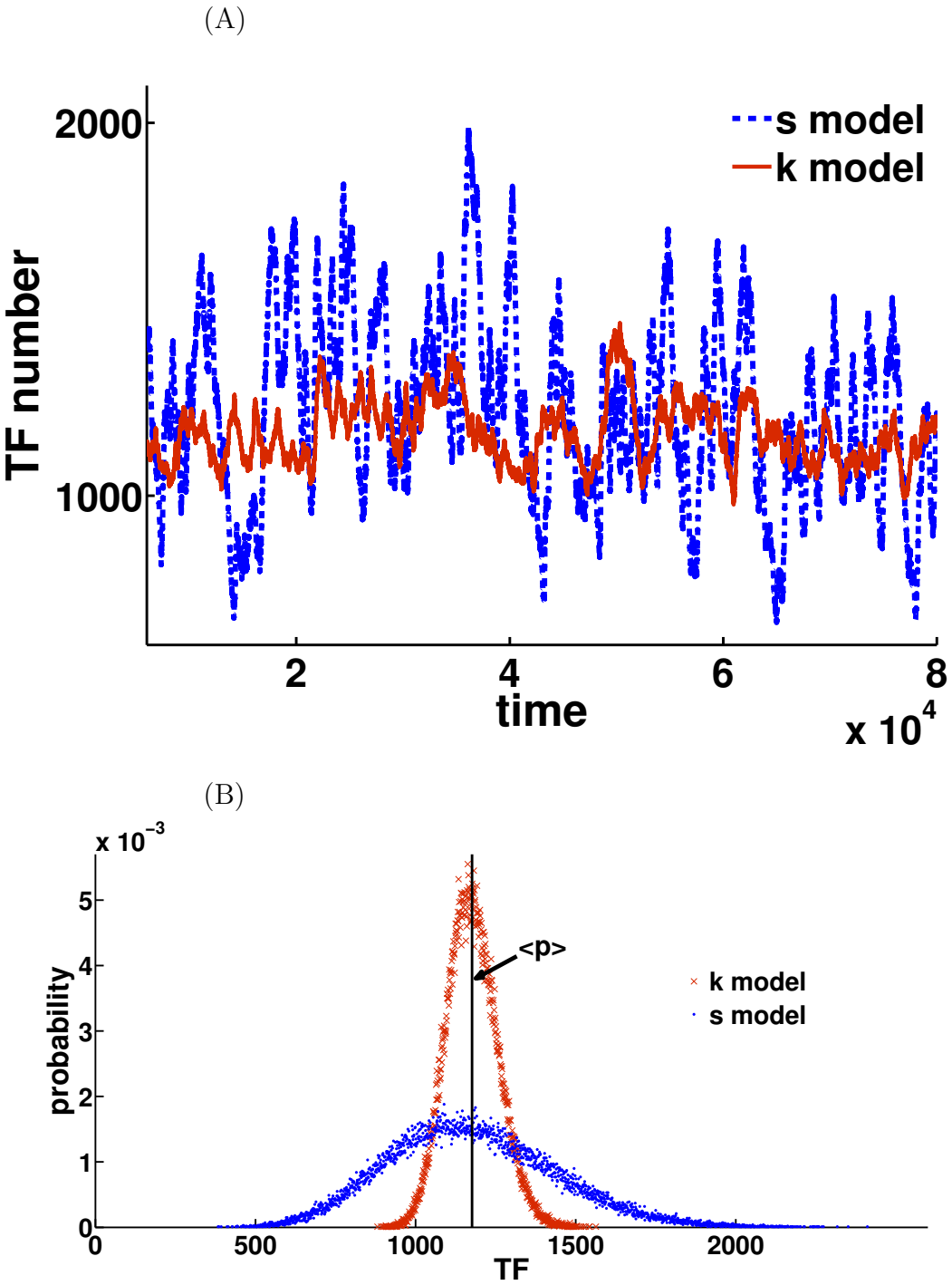


FIGURE 3.9: Expression variability depends on the mode of translational repression by miRNA. Comparison of the different effects on the noise of the two negative feedback schemes. A. Time evolution of the mean TF values for sequestration (black curves) and kinetic suppression model (gray). B. Histogram of the steady state TF distribution for the sequestration model (black) and for the kinetic suppression model (gray). The parameters are $\alpha_m = 1.1, \gamma_m = \gamma_\mu = 0.01, \gamma_p = 0.001, \sigma = 1.0, k_p = 0.1, k_d = 100.0, \kappa = 1.0, \beta = 0.00087$. κ and β are chosen so that their mean TF values are almost the same in both models. In the sequestration model, $< p >= 1174.6, \frac{\delta p^2}{<p>} = 57.34$, However, in the kinetic suppression model, $< p >= 1176.2, \frac{\delta p^2}{<p>} = 5.41$. Thus, in K model, both signal and noise are suppressed; while in S model, signal is suppressed, however, noise is amplified.

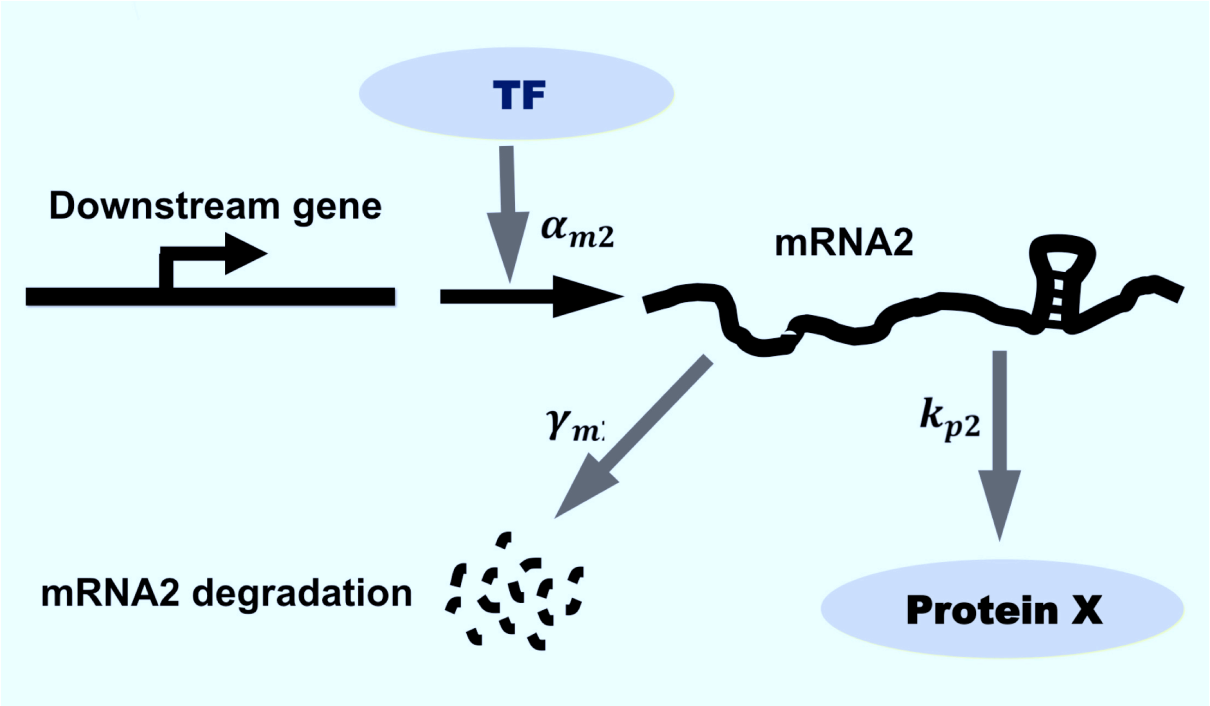


FIGURE 3.10: Scheme of transcriptional cascade involving the feedback-regulated TF and a downstream gene.

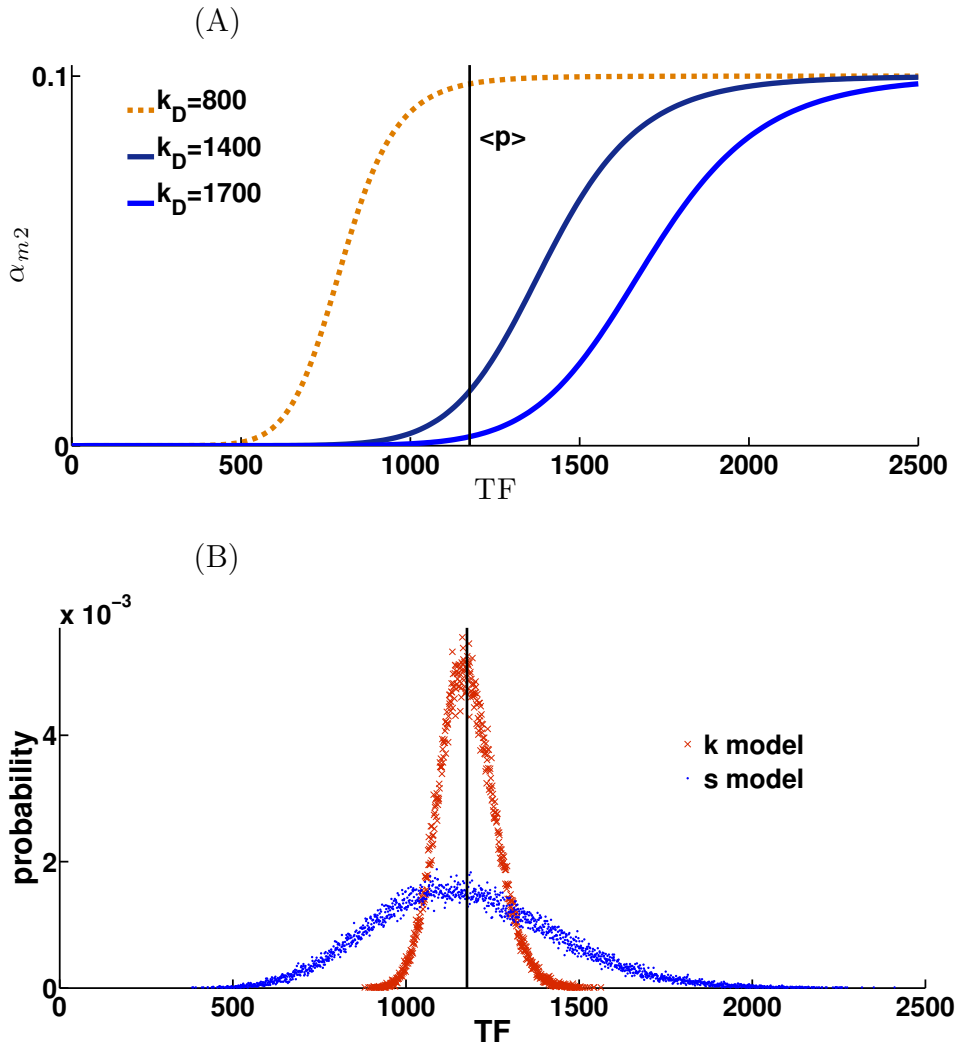


FIGURE 3.11: Scheme of information propagation in downstream gene. (A). Three different model genes with similar promoter strength (Hill coefficient $n = 10$.) but with different sensitivities. (B). The TF number histograms are plotted to display their overlap with the expression regions.

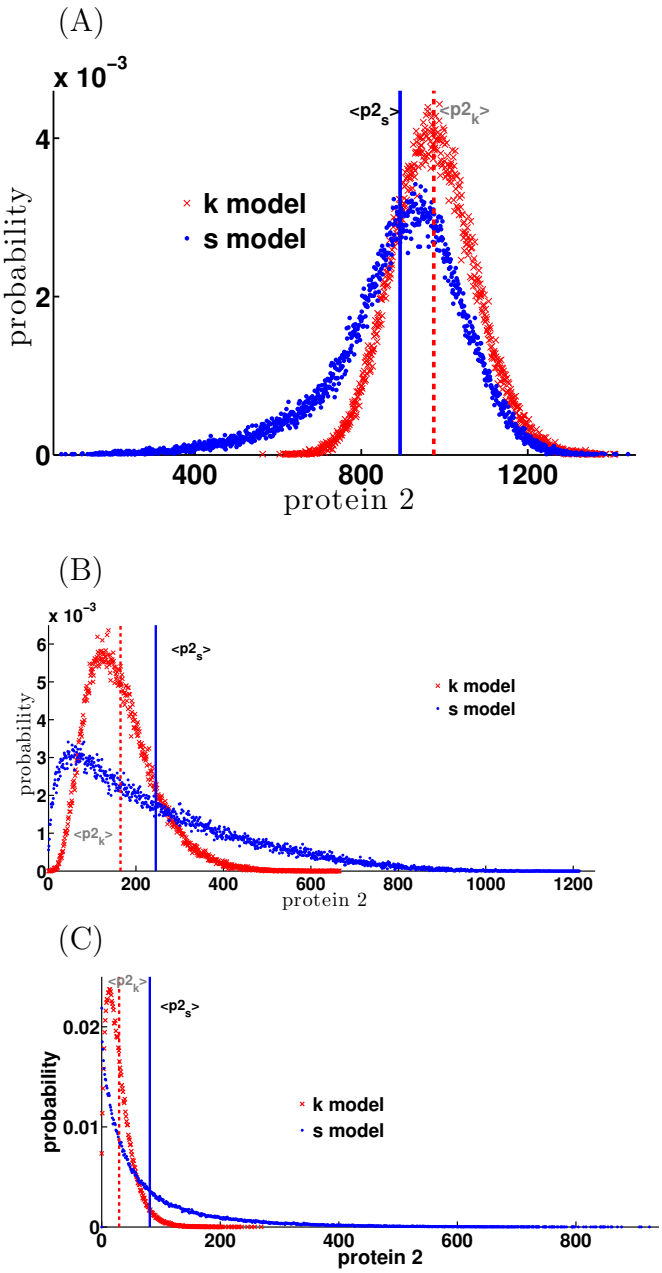


FIGURE 3.12: Mode of miRNA-mediated negative feedback can affect noise transmission to downstream genes in transcriptional cascades. Different noise level effect on downstream gene for different dissociation constants. When K_D is small, the downstream gene of both models are triggered. The mean value of downstream proteins in the K model is larger than in the S model. However, while K_D increases, although both mean values of downstream proteins decreases, The mean value of downstream proteins in the K model will be smaller than in the S model while K_D is over some value because of the long tail noise of S model. (A)$K_D = 800$, (B) $K_D = 1400$, (C)$K_D = 1700$. $\alpha = 1.1, \gamma_m = \gamma_\mu = 0.01, \gamma_p = 0.001, k_d = 100.0, k_p = 0.1, \kappa = 1.0, \beta = 0.00087, \sigma = 1.0, \epsilon = 0.1, \gamma_{m2} = 0.01, k_{p2} = 0.1, \gamma_{p2} = 0.001$.

4

Noise Propagation in Alternative Splicing

Transcription of eukaryotic genes produces a primary transcript (pre-mRNA) which contains both introns (intervening regions) that do not code for proteins and exons, which code for proteins. The process of splicing removes introns, assembling the coding sequences (exons) into a mature mRNA transcript that undergoes further processing before being translated. The process of splicing is accomplished by a set of specialized ribo-nuclear proteins (snRNPs) which assemble de novo around specialized sequences on the pre-mRNA called splice sites. Coincident with the discovery of splicing, it was discovered that the same pre-mRNA can be spliced in different transcripts, in a process called alternative splicing. Alternate splicing is widespread involving nearly 95% of mammalian genes, allowing a much larger diversity of proteins from a limited repertoire of few ten thousand genes in the genome [43, 96].

There are mainly seven different modes of alternative splicing observed: exon skipping, alternative $3'$ splice site selection, alternative $5'$ splice site selection, intron retention, mutually exclusive exons, alternative promoters and alternative polyadeny-

lation [97, 98]. What we are most interested in is exon skipping, which is the most common form of alternative splicing in higher eukaryotes, accounting for nearly 40% of alternative splicing events in higher eukaryotes [98, 99]. In exon skipping, a particular exon may be included in mRNAs under some conditions or in particular tissues, and omitted from the mRNA in others. There are several examples of different cellular functions of alternatively spliced isoforms. These include the process of sex determination in Drosophila [100]. The all-or-none bistable behavior of the two Sex-lethal (sxl) mRNA isoforms determine whether a Drosophila embryo will develop into a male or a female. Another example from the nervous system is the case of the immediate early gene *fosB*. In this case, is an example of intron retention. the ratio of FosB/ΔFosB isoforms have profound influence of the central nervous system functions [101].

The biochemical events associated with the process of gene expression are stochastic, owing to the relatively small numbers of reactants and the effect of thermal fluctuations. The resulting stochastic fluctuations (or noise) in mRNA and protein abundance can serve as an important source of cell-to-cell variability despite identical genetic composition and environmental conditions. Gene expression noise has profound effects on biological function – the variability in protein expression levels enables cells to switch their "fates" as well as generate heterogeneity across a population, which allows the population to respond to adverse conditions. Moreover, many disease states are associated with variability in gene expression [102]. Thus, extensive research has focused on exploring the origins and consequences of gene expression variability [103]. While most studies have focused on the contribution of transcriptional and translational noise to gene expression, that of alternate splicing remains unexplored. Alternative splicing patterns are determined by a combination of parameters, including elongation rate, chromatin and histone modifications, splicing factor abundance and modification, which modulate the recruitment of splicing

factors to the pre-mRNA [104].

While most studies have examined the role of these factors in a deterministic fashion, there have been few studies on the stochastic dynamics of splicing factor diffusion and binding to the pre-mRNA transcripts as well as the probabilistic nature of splicing itself [55, 105]. Here, we incorporate the process of alternate splicing to models of transcription to derive the factors that control the statistics and the relative abundance of alternately spliced isoforms.

4.1 Kinetic Model of Exon Skipping

Each splice site consists of a consensus sequence (a stretch of nucleotides) that can be recognized by spliceosomal components, the small nuclear ribonucleoproteins (snRNPs) U1, U2, U4, U5 and U6 and auxiliary factors, including U2AF65 and U2AF35. A strong splice site means its consensus sequence can be recognized more efficiently by spliceosomal components and in general, strong splice site leads to constitutive splicing, i.e. the associated exon is always present in the final, mature transcript. A weak splice site is a sequence that diverges considerably from the consensus sequence resulting in low-affinity binding of splicing factors. Thus, the competition between adjacent strong and weak splicing site leads to alternative splicing, whereby the exon associated with the weak site is either included or excluded from the mature transcript depending on the context.

Experiments have shown that most of the splicing events occur co-transcriptionally [106–108]. In other words, splicing takes places before the nascent RNA is released from RNA polymerase II. The process of splicing can start as soon as the pre-mRNA comes out from RNA polymerase II. Two mutually exclusive models have been proposed to explain the coupling of alternate splicing and transcription: recruitment coupling model and kinetic coupling model. In recruitment coupling model, the splicing factors are prepositioned on the RNA polymerase II carboxy-terminal do-

main (CTD) and hop on to the splicing site as soon as it emerge from RNA polymerase II [109]. The kinetic model, however, suggests that splicing factor directly assemble into a spliceosome on the splice site due to their high concentration and mobility [53]. Support for the latter model comes from the finding that transcription by slow RNA polymerase II mutants increases alternative exon inclusion[110, 111]. Application of drugs that inhibit elongation process also promote alternative exon inclusion [111, 112]. According to kinetic coupling model, inclusion or exclusion of an exon in the final transcript is influenced by transcription elongation rate as well as splicing factor diffusion. Although lots of experiments have been done to verify how elongation rate affect the ratio of two mRNA isoforms, little theoretical work

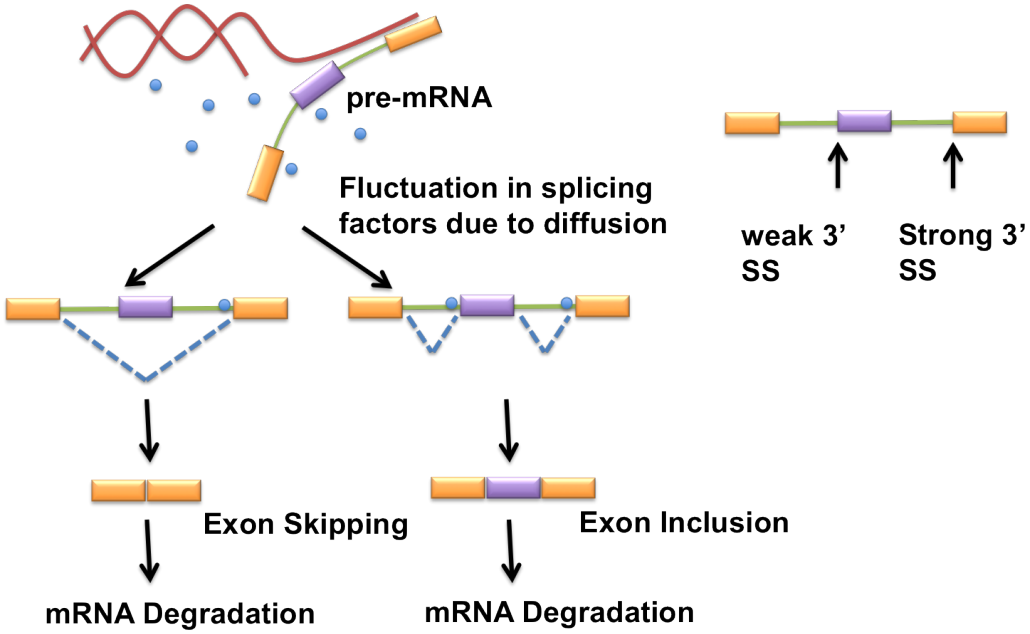


FIGURE 4.1: Kinetic model of exon skipping: There are two splice sites, a weak splice site upstream and a strong splice site downstream of the alternately spliced exon, which compete for the recruitment of splicing factors. If the weaker splice site is unoccupied by the spliceosome, the associated exon is skipped. On the other hand, if splicing factors bind to the weaker splice site, then the alternative exon will be included. Constitutive exons are shown in yellow and alternatively spliced exons are shown in purple. Solid lines represents introns and dashed lines represents alternately spliced transcripts. Blue dots represents splicing factors.

has been done to investigate the functional role of elongation and splicing factor diffusion in alternative splicing.

The kinetic model of exon skipping we propose is shown in Figure 4.1. However, we note that the same model can also be applied for the analysis of intron retention, another prevalent form of alternate splicing. Also in the kinetic model we propose, the function of splicing factor protein is to facilitate splicing, acting as a splicing activator, like U2AF65. It can also represent a splicing inhibitor, such as the protein Polypyrimidine Tract Binding Protein (PTBP), a ribonucleoprotein known to suppress splicing. Thus, a bound splicing inhibitor leads to the retention of the alternately spliced exon (or in some cases the retention of an intron).

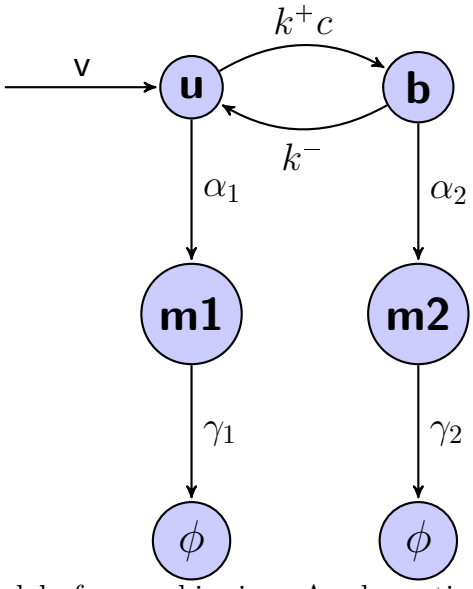


FIGURE 4.2: Simplified model of exon skipping: A schematic diagram of exon skipping showing the relevant states and associated transition rates. v is RNA PolII elongation rate, c is the concentration of the core splicing factor. u is the unbound pre-mRNA (PM1) and b is the bound pre-mRNA (PM2). Two kinds of mRNA isoforms $m1$ and $m2$ are the two alternately spliced mRNA transcripts that can be produced respectively from these two pre-mRNA states. They both degrade at the same rates, γ.

In order to mathematically calculate the noise properties of alternative splicing, we further simplify the above model into Figure 4.2. This is a very generalized

model, in which we assume RNA Polymerase II elongation rate along the gene is v. The concentration of the core splicing factor, which can either be $U2AF65$ or $PTBP$ or any other core splicing factor depending on the scenario, is c. We denote u as the number of pre-mRNA transcripts (PM1) whose weak splicing site is free of splicing factors and b is the number of pre-mRNA (PM2) whose weak splicing site is bound to the splicing factor. These two different states can switch due to the stochastic properties of the binding and unbinding of the splicing factor. Thus, two kinds of mRNA isoforms ($m1$ and $m2$) can be produced, corresponding to these two pre-mRNA states. For convenience, in the remainder of the discussion, we assume that these isoforms degrade with the same rates.

First, by assuming all the parameters as constant, we calculate the mean and variance of each variable as well as correlations between variables using both Master equation approach and linear noise approximation approach. Next, we extend the linear noise approximation to determine the influence of the diffusion of splicing factors. We finally validate our model and theoretical calculation by stochastic simulation.

4.2 Results

4.2.1 Isoform Abundance is Poisson Distributed

For the alternative splicing model described in Figure 4.2, there are four variables: $\vec{x} = \{u, b, m_1, m_2\}$. If all the parameters are constant, then we can analytically solve this system and get the mean and variance of these four variables using two mathematical approaches. Let $P(u, b, m_1, m_2; t)$ represents the probability that at time t, the number of pre-mRNA is u, the number of pre-mRNA with an assembled spliceosome at the weak splice site is b and m_1 and m_2 are the number of the two mature isoforms respectively. Then the chemical master equation for this reaction

system takes the form:

$$\begin{aligned}\frac{\partial P(u,b,m_1,m_2,t)}{\partial t} &= vP(u-1,b,m_1,m_2,t)\\ &+ k^+c(u+1)P(u+1,b-1,m_1,m_2,t) + k^-(b+1)P(u-1,b+1,m_1,m_2,t)\\ &+ \alpha_1(u+1)P(u+1,b,m_1-1,m_2,t) + \alpha_2(b+1)P(u,b+1,m_1,m_2-1,t)\\ &+ \gamma_1(m_1+1)P(u,b,m_1+1,m_2,t) + \gamma_2(m_2+1)P(u,b,m_1,m_2+1,t)\\ &- (v + k^+cu + k^-b + \alpha_1 u + \alpha_2 b + \gamma_1 m_1 + \gamma_2 m_2))P(u,b,m_1,m_2,t). \end{aligned} \tag{4.1}$$

with the appropriate normalization. In order to derive analytical results, we write the moment generating function as:

$$G(z_u, z_b, z_{m_1}, z_{m_2}, t) = \sum_{u,b,m_1,m_2=0}^{\infty} z_u^u z_b^b z_{m_1}^{m_1} z_{m_2}^{m_2} P(u,b,m_1,m_2,t). \tag{4.2}$$

From this, we can obtain the mean levels of the relevant variables (From Appendix Equation C.7):

$$\begin{cases} \langle u \rangle = \dfrac{v}{\alpha_1 + g\alpha_2} \\ \langle b \rangle = \dfrac{vg}{\alpha_1 + g\alpha_2} \\ \langle m_1 \rangle = \dfrac{\alpha_1}{\gamma_1}\dfrac{v}{\alpha_1 + g\alpha_2} \\ \langle m_2 \rangle = \dfrac{\alpha_2}{\gamma_2}\dfrac{vg}{\alpha_1 + g\alpha_2}, \end{cases} \tag{4.3}$$

where $g \equiv \frac{k^+c}{\alpha_2 + k^-}$ is the effective production rate of b from u. It also determines the ratio of two mRNA isoform production $\frac{<m2>}{<m1>} = \frac{\alpha_2/\alpha_1}{\gamma_2/\gamma_1}g$. We can further get the

variance of these four variables (From Appendix Equation C.9):

$$\begin{cases} \sigma_u^2 = \langle u^2 \rangle - \langle u \rangle^2 = \langle u \rangle = \dfrac{v}{\alpha_1 + g\alpha_2} \\ \sigma_b^2 = \langle b^2 \rangle - \langle b \rangle^2 = \langle b \rangle = \dfrac{vg}{\alpha_1 + g\alpha_2} \\ \sigma_{m_1}^2 = \langle {m_1}^2 \rangle - \langle m_1 \rangle^2 = \langle m_1 \rangle = \dfrac{\alpha_1}{\gamma_1} \dfrac{v}{\alpha_1 + g\alpha_2} \\ \sigma_{m_2}^2 = \langle {m_2}^2 \rangle - \langle m_2 \rangle^2 = \langle m_2 \rangle = \dfrac{\alpha_2}{\gamma_2} \dfrac{vg}{\alpha_1 + g\alpha_2}, \end{cases} \tag{4.4}$$

We immediately note that the variance of each variable equals its mean. Therefore, all four variables are Poisson distributed and the Fano factors equal 1 if all the parameters are constant. We can also get the covariance term of m_1 and m_2

$$\mathrm{cov}(m_1, m_2) = \langle (m_1 - \langle m_1 \rangle)(m_2 - \langle m_2 \rangle) \rangle = 0 \tag{4.5}$$

Thus m_1 and m_2 are two independent Poisson distributed random variables and there is no correlation between them if all the parameters are constant. An alternate method to describe the statistics of the reaction system is the more familiar chemical Langevin equations:

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t} = v - (k^+ c + \alpha_1)u + k^- b + (\sqrt{v}\Gamma_1 - \sqrt{k^+ cu}\Gamma_2 + \sqrt{k^- b}\Gamma_3 - \sqrt{\alpha_1 u}\Gamma_4) \\ \dfrac{\mathrm{d}b}{\mathrm{d}t} = k^+ cu - (k^- + \alpha_2)b + (\sqrt{k^+ cu}\Gamma_2 - \sqrt{k^- b}\Gamma_3 - \sqrt{\alpha_2 b}\Gamma_5) \\ \dfrac{\mathrm{d}m_1}{\mathrm{d}t} = \alpha_1 u - \gamma_1 m_1 + (\sqrt{\alpha_1 u}\Gamma_4 - \sqrt{\gamma_1 m_1}\Gamma_6) \\ \dfrac{\mathrm{d}m_2}{\mathrm{d}t} = \alpha_2 b - \gamma_2 m_2 + (\sqrt{\alpha_2 b}\Gamma_5 - \sqrt{\gamma_2 m_2}\Gamma_7). \end{cases} \tag{4.6}$$

where Γ_i $(i = 1 \ldots 7)$ represent seven temporally uncorrelated, statistically independent Gaussian white noise terms introduced by seven chemical reactions as described in Figure 4.2. By using linear noise approximation approach as described

in Appendix C.2.1, we can get the same results as using the Master equation approach. However, this will be more useful as shown below for cases where we consider fluctuations of the level of splicing factor.

We see that in this simplest model, the elongation rate does not affect the ratio of two isoforms. This is because when the elongation rate increases, we simply assume both the isoforms increase in proportion. However, experiments suggest that elongation rate alters the ratio of two mRNA isoforms[104, 110, 111, 113]. In our model, we have assumed that the strong and weak splice sites are simultaneously transcribed, i.e. the transcription rate is fast, while the pre-mRNA processing steps are relatively slow. However, it is possible that for genes in which exon skipping depends on the transcriptional rate, the distance between the constitutive and alternately spliced exons might be relatively large and the weak splice site is upstream (5') from the strong site. Thus, when the elongation rate is slow, the weak splice site is transcribed first and is able to recruit splicing factors before the strong site is transcribed. On the other hand, when the transcript synthesis (PolII elongation) rate is high, both splicing sites become available almost simultaneously, which makes the weak splice site less likely to recruit splicing factors which then results in exon skipping. We are exploring a more general model where we account for explicit coupling between transcription and splicing. In general, the elongation rate affects the concentration of splicing factor available for the weak site binding. If we assume $c = f(v)$ in our model and increasing v will decrease c, then this simplest model still makes sense. Furthermore, as we discuss in the next section for discussing the influence of splicing factor diffusion, splicing factor diffusion will also alter the mRNA isoform ratio when elongation rate changes as observed in the experiments which strongly validate our model.

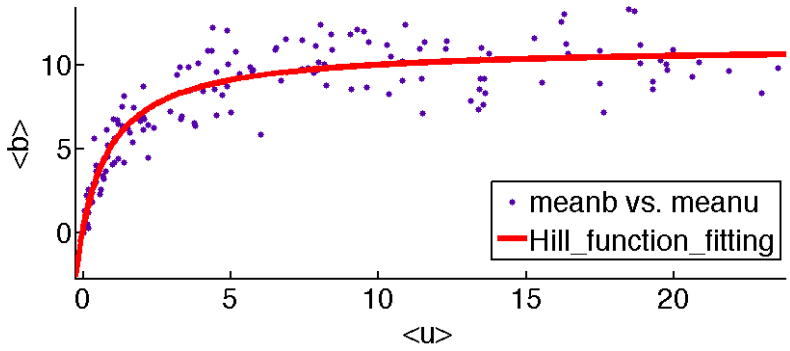
4.2.2 Elongation Rate Alters Isoform Ratio due to Splicing Factor Diffusion

So far, we have assumed that the splicing factor concentration is essentially fixed and sufficiently large such that binding and unbinding to the pre-mRNA transcripts do not alter their availability. However, as shown by Waks *et al.* [55], cells have a limited complement of splicing factors which must be shared between the relatively large number of simultaneously transcribed genes. While the splicing factors bind to the weak splice site to regulate the alternative splicing, the precision of the regulation is limited by the randomness in the arrival of splicing factor molecules at these sites. In order to investigate the influence of splicing factor diffusion, we need to make the following assumption so as to simplify our model: First, we assume splicing factors are highly mobile and diffuse freely in the nucleus. Second, because the diameter of a typical mammalian nucleus is 5-10 microns (1 micron equals 10^3 nanometers), and most alternative splicing occurs co-transcriptionally, or happens very near transcription site, the region splicing occurs is much less than the volume of the mammalian nucleus [114]. Thus, in our models, we can simply assume that the splicing factor that can be absorbed to or released from binding site on the pre-mRNA located at $\vec{x}_0$.

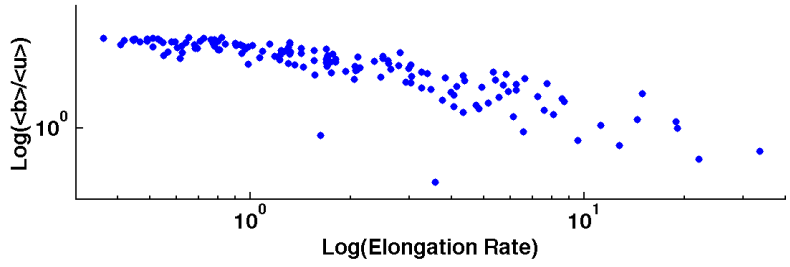
If we assume splicing factor diffuse freely in three dimensions through the surrounding solution to the binding site located at $\vec{x}_0 = 0$. c represents the concentration of splicing factors that can freely diffuse. If splicing factor binds to splicing site, then c decreases. We can write time-derivative of the concentration of splicing factor surround the binding site as a function of 3D diffusion and binding.

$$\frac{\partial c(\vec{x},t)}{\partial t} = D\nabla^2 c(\vec{x},t) - \dot{b}\delta(\vec{x}-\vec{x}_0), \tag{4.7}$$

where D is the diffusion constant and $\delta(\vec{x}-\vec{x}_0)$ is the Dirac-delta function. This equation describes the diffusion of the splicing factor which can be absorbed or



(a)



(b)

FIGURE 4.3: Ratio of mRNA isoforms is altered while elongation rate increases due to splicing factor diffusion. : (a) shows at first, when u increases, b also increases, however, when u becomes large enough, the increase rate of b approaches 0. Due to splicing factor diffusion, b is up-bounded. Dots represents simulation results using the Gillespie algorithm and red line denotes a Hill function fit. (b) The ratio of two mRNA isoforms ($\frac{\langle b\rangle}{\langle u\rangle}$) versus the elongation rate, showing a clear decrease as the elongation rate increases, which is consistent with experimental observations. For simulations, we treat nucleus as a cube and divide it into $100 \times 100\times$ subcubes. At each node, the probability it will be occupied with a splicing factor is set to be 0.001, so there will be around 1000 splicing factors in nucleus. $\alpha_1 = \alpha_2 = 0.5$, $k^+ = 10000/molecule$, $k^- = 0.5$, v ranges from 0.1 to 15.

released from a binding site on pre-mRNA located at $\vec{x}_0$. If we integrate Equation 4.7 over the whole nuclear volume,

$$\begin{aligned}\frac{\mathrm{d}b}{\mathrm{d}t} &= D\iiint_V \nabla^2 c(\vec{x},t)\mathrm{d}V - \iiint_V \frac{\partial c(\vec{x},t)}{\partial t}\mathrm{d}V \\ &= D\oiint_S \nabla c(\vec{x},t)\cdot\vec{n}\mathrm{d}S - \frac{\partial}{\partial t}\Big(\iiint_V c(\vec{x},t)\mathrm{d}V\Big). \end{aligned} \tag{4.8}$$

On the right of Equation 4.8, the first terms represents the net flux through the boundary of nucleus. The second term represents the time derivative of total number of splicing factor within nuclear region. Because we assume a fixed number of splicing factors in nucleus region, some binds to splice sites, some diffuse freely in nucleus. There will be no net flux of c through the boundary of nucleus. Thus the first term on the right of Equation 4.8 will equal 0, which leads to

$$\frac{\mathrm{d}b}{\mathrm{d}t} = -\frac{\partial}{\partial t}\Big(\iiint_V c(\vec{x},t)\mathrm{d}V\Big) \tag{4.9}$$

From Equation 4.9, we can see that the total number of splicing factor available for binding provides a upper limit for the increase of b. When elongation rate increases, more pre-mRNA unbound to splicing factors, u, will be produced. Thus the probability that splicing factors will bind to the splicing site of u and change u to b will increase. Consequently, fewer splicing factors will be available to bind to new pre-mRNA. So $\dot{b}$, the derivative of b will approach 0 when elongation rate reaches a large value. However, u will still increase proportionally with the elongation rate. Thus with the increase of the elongation rate, the ratio of two mRNA isoforms (b/u) will decrease. We also verify this by numerical simulation as shown in Figure 4.3. This also fits the experimental observation [104, 110, 111, 113] . One thing needs to be specified: During computer simulation, k^+ is chosen a very big number 10000 because this probability is "per molecule". If change to "per NM" it will be around $k^+ \approx 0.6/nM$ if assuming the reaction volume to be $V = 100nm^3$

4.2.3 Splicing Factor Diffusion Suppresses Noise

In order to examine the influence of splicing factor diffusion, we assume all the parameters as described in the Scheme 4.2 are constant except the splicing factor concentration c, which is assumed to be determined by diffusion in the nuclear region around the gene. Splicing factors can bind and unbind with u weak splicing site as

described in Equation 4.7). Assume the mean concentration is $\bar{c}$, and assume $c(\vec{x},t)$ will fluctuate around the mean value. Linearizing the concentration around the mean:

$$c(\vec{x},t) = \bar{c} + \delta_c(\vec{x},t) \tag{4.10}$$

We transform the chemical Langevin equations into Fourier space, to obtain the following relationship between the Fourier representations $\delta_c(\vec{x},t)$ is $\hat{\delta}_c(\vec{k},\omega)$:

$$\delta_c(\vec{x},t) = \int \frac{\mathrm{d}\omega}{2\pi} \int \frac{\mathrm{d}^3k}{(2\pi)^3} \mathrm{e}^{i(\omega t + \vec{k}\cdot\vec{x})} \hat{\delta}_c(\vec{k},\omega). \tag{4.11}$$

From Equation 4.7, we can get the following equation in Fourier space:

$$i\omega\hat{\delta}_c(\vec{k},\omega) = -D|k|^2\delta_c(\vec{k},w) - i\omega\hat{\delta}_b(\omega)\mathrm{e}^{-i\vec{k}\cdot\vec{x}_0} \tag{4.12}$$

and

$$\hat{\delta}_c(\vec{k},\omega)\mathrm{e}^{i\vec{k}\cdot\vec{x}_0} = \frac{i\omega\hat{\delta}_b(\omega)}{-i\omega - D|k|^2} \tag{4.13}$$

Thus:

$$\begin{aligned}
\hat{\delta}_c(\vec{x}_0,\omega) &= \int \frac{\mathrm{d}^3k}{(2\pi)^3}\hat{\delta}_c(\vec{k},\omega)\mathrm{e}^{i\vec{k}\cdot\vec{x}_0} \\
&= -i\omega\hat{\delta}_b(\omega)\int \frac{\mathrm{d}^3k}{(2\pi)^3}\frac{1}{i\omega + D|k|^2} \qquad (4.14)\\
&= -i\omega\hat{\delta}_b(\omega)\int \frac{1}{2\pi^2}\frac{|k|^2\mathrm{d}|k|}{i\omega + D|k|^2} \qquad (4.15)
\end{aligned}$$

This integral is ultraviolet divergent, because we have assumed that the binding site is infinitely small and binding only occurs at one point $\vec{x}_0$ as described in the Dirac-delta function. A more realistic treatment is to assume the binding site has a finite linear size a, which is equivalent to cutting off the k integral at a relative large value $\Lambda \sim \pi/a$. Furthermore, in experimental observations, the time interval between observations is always much longer than the correlation time of the noise.

Thus, the observed variable is the average occupancy of the splice site over some characteristic time. Thus, it is justified to take the low frequency limit of the noise spectrum. Hence, $\omega \ll D/a^2$ and

$$\hat{\delta}_c(\vec{x}_0, \omega \ll D/a^2) = -i\omega\hat{\delta}_b(\omega)\int_0^\Lambda \frac{\mathrm{d}|k|}{2\pi^2 D} = -\frac{i\omega\hat{\delta}_b(\omega)}{2\pi Da} \tag{4.16}$$

Combining the Lagevin equation as stated in Equation 4.6 and the splicing factor concentration equation 4.7, and using Fourier transformed variables, we can get the variance of all four variables u,b,m_1 and m_2 (Appendix Equation C.36). However, since the expression of the last two variables are cumbersome, we only list the variance of u and b:

$$\begin{cases} \sigma_u^2 = \dfrac{v}{\alpha_1 + g\alpha_2}\dfrac{k^+c + (\alpha_2 + k^- + \alpha_2 g\theta)(1+\theta) + \alpha_1(1+\theta)^2}{(1+\theta)(\alpha_1 + \alpha_2 + k^- + k^+c + \alpha_1\theta)}, \\ \sigma_b^2 = \dfrac{vg}{\alpha_1 + g\alpha_2}\dfrac{\alpha_2 + k^- + (\alpha_1 + k^+c)(1+\theta)}{(1+\theta)(\alpha_1 + \alpha_2 + k^- + k^+c + \alpha_1\theta)}, \end{cases} \tag{4.17}$$

where $\theta = \frac{k^+\langle u\rangle}{2\pi Da} = \frac{k^+ v}{2\pi Da(\alpha_1 + g\alpha_2)}$

We can see, on the right of the equations, the first term represents the noise without the diffusion influence, it's the same as in Equation 4.4. The second term represents the influence of splicing factor diffusion. We can see for the variance of b, because $\frac{\alpha_2 + k^- + (\alpha_1 + k^+c)(1+\theta)}{(1+\theta)(\alpha_1 + \alpha_2 + k^- + k^+c + \alpha_1\theta)} < 1$ is always true, so the following statement always stands:

$$\sigma_b^2 < \frac{vg}{\alpha_1 + g\alpha_2}, \tag{4.18}$$

This means the variance of b will always be smaller than Poisson and the Fano factor of b will always less than 1. The variance of u sometimes can be bigger than Poisson, sometimes can be smaller than Poisson, depending on the value of the parameters.

4.4a 4.4b are the theoretical calculation results calculated from Equation 4.17, 4.4c,4.4d are the results from numerical simulation. They plot the variance versus

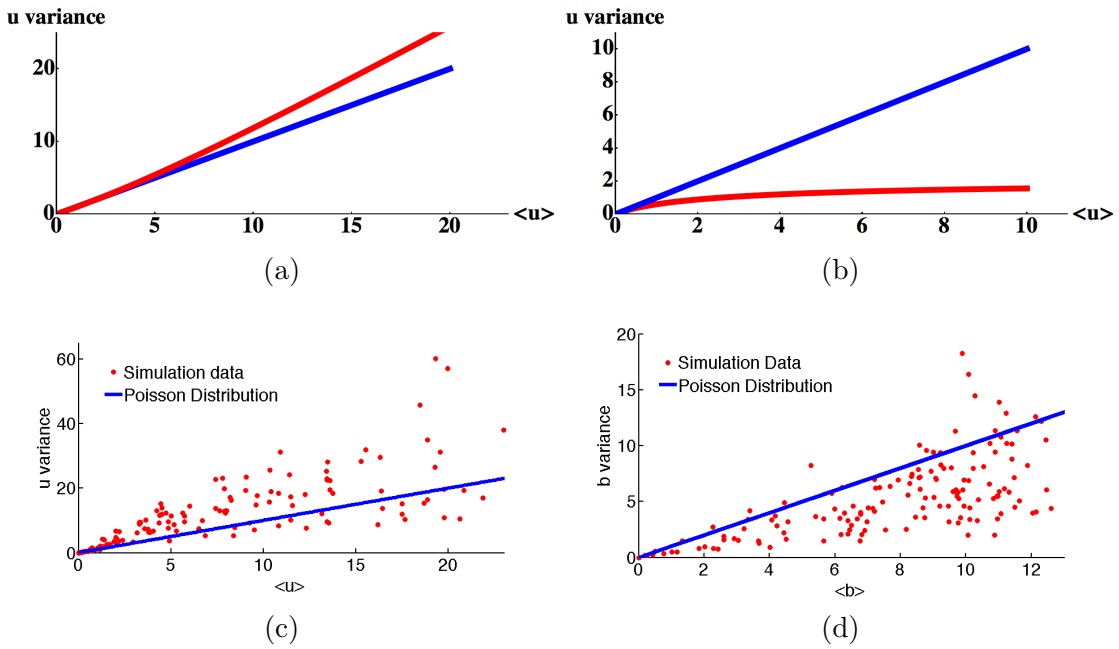


FIGURE 4.4: Noise properties of pre-mRNAs in alternative splicing: 4.4a ,4.4b are the theoretical calculation results calculated from Equation 4.17, 4.4c, 4.4d are the results from numerical simulation. They plot the variance versus mean value of variable u and b. Blue line represents the case that without considering diffusion, all the parameters are constant, i.e., it represents Poisson distribution, in which case the variance is strictly equals the mean value. So we can see the noise of variable u is slightly larger than the Poisson noise and the noise of b is repressed and smaller than the Poisson noise. For simulation, $\alpha_1 = \alpha_2 = 0.5$, $k^+ = 10000/molecule$, $k^- = 0.5$, v ranges from 0.1 to 15. For theoretical calculation, $k^+ = 0.5/nM$, $\bar{c} = 1.0nM$, $\theta = \langle u \rangle$. All other parameter are the same as in simulation.

mean value of variable u and b. Blue line represents the case that without considering diffusion, all the parameters are constant, i.e., it represents Poisson distribution, in which case the variance is strictly equals the mean value. So we can see the noise of variable u is slightly larger than the Poisson noise and the noise of b is repressed and smaller than the Poisson noise. Figure 4.5 shows that due to different noise properties of u and b, while noise propagate from pre-mRNA to mature mRNAs, noise properties of m_1 and m_2 will also be different. The additional noise of u from splicing factor diffusion will propagate and lead to noise amplification of m_1. However, due to repressed noise properties of b, m_2 will be quasi-Poisson distributed.

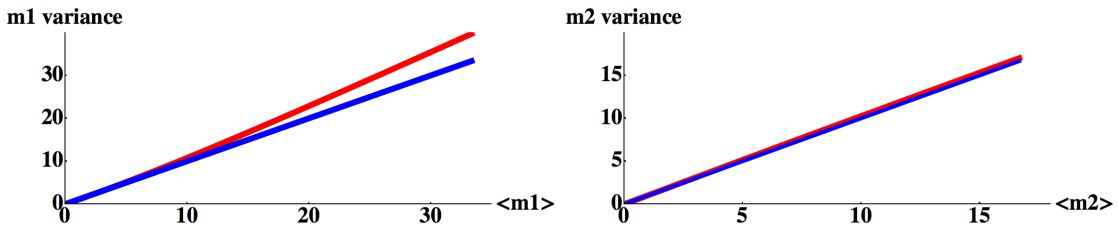


FIGURE 4.5: Noise properties of mature mRNAs in alternative splicing: These two figures are from theoretical calculation. They plot the variance versus mean value of variable m_1 and $m_2$while changing the parameter v. Blue line represents the case that without considering diffusion, all the parameters are constant, i.e., it represents Poisson distribution, in which case the variance is strictly equals the mean value. So we can see the noise of variable m_1 is amplified and larger than the Poisson noise, however m_2 are quasi-Poisson distributed. this is due to the repressed noise properties of b. $\alpha_1 = \alpha_2 = 0.5$, $\gamma_1 = \gamma_2 = 0.3$ $k^+ = 0.5/nM$, $k^- = 0.5$, $\theta = \langle u \rangle$, $\bar{c} = 1.0nM$. v ranges from 0.1 to 15.

4.3 Conclusion and Discussion

Here, we proposed a mathematical model of exon skipping and investigated the noise properties of this system. Our results have important implication for stochastic modeling of gene expression. In stochastic model of gene expression, mRNA synthesis for an active promoter is typically modeled as a Poisson process. For an mRNA transcript that undergoes alternative splicing, this is also accurate in the presents of a stable internal environment (if all variables are constant, both of the mRNA isoforms will still be Poisson). However, fluctuation in elongation rate will result in positive correlation between two mRNA isoforms and fluctuation in splicing factor concentration will result in negative correlation between two mRNA isoforms. Our results also indicate that splicing factor diffusion may have different impact on two mRNA isoforms. Splicing factor diffusion will decrease the variance of pre-mRNA with splicing factor bound to the weak splicing site (b), but increase the variance of pre-mRNA without splicing factor bound to the weak splicing site (u). Mature mRNA produced from the pre-mRNA with the assembled spliceosome on the weak splice site (b) will be quasi-Poisson distributed due to the diffusive fluctuations of

spliceosome components. This could be a way for the biological system to control the noise propagation. We also show that the elongation rate of RNA PolII will alter the ratio of two mRNA isoforms due to splicing factor diffusion. A lot of experiments report that RNA polymerase II elongation can affect the ratio of two mRNA isoforms [104, 115]. Slow elongation rate favors commitment to alternative exon inclusion and fast elongation rate will lead to higher alternative exon skipping. Our theoretical calculation indicates that this may due to splicing factor diffusion. Longer exposure time of upstream weak splicing site may indicate higher chance for splicing factor binding.

There are many cases in which alternative splicing plays an important role in cell functioning. For example, ΔFosB, a truncated splice variant of FosB, has been implicated in the development of drug addiction and control of the reward system in the brain [116, 117]. External stimuli, such as cocaine, can alter the ratio of FosB, ΔFosB for a long time, even after long periods of abstinence [118]. This is an example of intron retention. Another interesting case is sex determination in Drosophila. It is an example of exon skipping. Sex determination in Drosophila is controlled by a cascade of splicing factors that are themselves alternatively spliced, ultimately leading to the sex-specific expression of two different variants of the Doublesex (Dsx) transcription factor. there are a lot more examples indicate how functional significance alternative splicing is. So investigate how external stimuli regulate alternative splicing and result in tissue- and stage-specific protein isoforms with different functions in in germ cells, muscle and the central nervous system is of great importance. As variability in gene expression levels can impact various cellular behavior [52, 103, 119], Little is known the consequences of variability in spliced mRNA isoform ratios. The experimental observations support the idea that co-transcriptional splicing is the default mechanism and checkpoints that prevents the unspliced transcript release are likely to exist in higher eukaryotes [114]. However, some experimental observation that some

alternative splicing may not necessary co-transcriptional [55, 120]. In this situation, splicing is un-coupled from transcription and pre-mRNA may diffuse in nucleoplasm and have a more complex interaction with splicing factors. More research needs to be done to investigate the noise properties of alternative splicing in this situation.

5

Conclusion and Future Outlook

Neurons, the fundamental units of the brain, can change their structure, inter connectivity and biochemical properties in response to electrical activity. Such activity dependent changes form the molecular basis of learning, memory formation and the development of neuronal networks in the brain. These changes are in part due to the induction of specialized genes, termed activity-regulated genes that initiate or terminate protein production in response to activity. While the dynamics of gene induction by activity have been well studied in large populations of neurons, the response of activity-dependent genes in single neurons is not known. In particular, because of the small numbers of molecules involved in the gene production as well as their weak interactions, the variation of the number of these molecules can be crucial and stochastic effects must be considered.

Neurons signal to each other by rapid changes in their membrane electrical potential (the action potential). Action potentials can then cause the release of chemicals at connections between neurons (synapses), which in turn leads to changes in membrane potentials of the recipient cells by means of specialized proteins in the cell membrane. Action potentials not only mediate signaling between neurons, but also control

transcription of genes important for neuronal function, a process known as activity-dependent gene expression. Among the many functions of activity-dependent gene expression is the alteration of the levels of key proteins at synapses, thereby altering the synaptic connection between neurons for days or longer in response to changes of sensory experience [56–60].

Current experimental work assesses activity dependence of transcription in bulk population assays. However, due to the large diversity of the neuronal firing patterns, it remains unclear how activity-driven genes in single neurons respond to external stimuli and what is the limit that external stimuli can alter the gene expression.

Most theoretical studies have focused on steady state responses of genes to stimuli [5, 10, 75, 92, 103], in particular because the experimental systems under study are non-stationary. Bacteria and yeast cells grow and divide within hours, diluting levels of proteins, which introduces a natural time-scale of decay and obscures the effect of temporal changes in transcription in any individual cell. Moreover, technical considerations limit the time scale of environmental variations that can be applied[121]. Therefore, theoretical studies designed to study prokaryotic gene regulation cannot be simply extended to understand the quantitative nature of transcriptional responses to rapidly varying stimuli.

As part of my Ph.D. work (Chapter 2), I have proposed a mathematical model to obtain the temporal evolution solution of transcription. In the future, I shall try to use this solution to try to characterize how activity dependent genes in neurons translate time-varying signals into an appropriate protein output at the level of individual neurons. Such characterization is important in many respects. First, individual neurons tune the expression levels of voltage-dependent channels in response to activity, in order to maintain a metabolically appropriate level of action potential activity, termed homeostasis. Secondly, activity dependent changes in single neurons caused by behaviorally important stimuli is important for memory formation[122–124].

This characterization can be divided into two categories: First, how do time-varying signals affect the stochastic nature of transcription? As discussed previously, calcium influx into neurons (by the activation of voltage-dependent calcium channels) leads to the activation of transcription factors which can bind to promoter regions of activity-dependent genes to initiate the transcriptional response[125, 126]. In addition to initiation, neuronal activity (acting through calcium sensitive enzymes) can regulate several facets of gene regulation [127]. Also, calcium signals differing in amplitude, spatial and temporal properties can trigger different transcriptional responses[128, 129]. In a reduced model, as presented in dissertation studies, this would correspond to activity dependence of gene activation (or inactivation), transcriptional initiation and elongation. In this way, we can encode the Ca^{2+} or transcription factor concentration variables into the transcriptional rate changes to obtain the temporal characteristics of the mRNA distribution. This can tell us how mRNA distribution changes over time and how long it will take to reach the steady state due to different external environment which can help to properly interpret experimental mRNA counting from single cells.

A second, fundamental question is: how do expression levels of activity-dependent genes change corresponding to changes in concentration of transcription factors? Information theory is a powerful way of characterizing the quality of information transfer. It give us the opportunity to analysis the transcriptional machinery in terms of its input, output and transfer characteristics without any knowledge required of its internal workings[130]. In general, this regulation machinery can be thought of as an input/output device in which the input is the concentration of transcription factors (c) and the output is the concentration of the gene product (g) as described in Scheme 5.1. To explore these issues, we need to quantify the number of distinguishable settings of the transcription machinery and its temporal dynamics. One measure that is ideally suited for this task is Shannon's information theory, which

FIGURE 5.1: Scheme of the information channel of transcription.

has been widely used to quantify information flow in biochemical and gene networks. The core concept in information theory is the entropy $H(g)$ and the mutual information $I(c;g)$ given the input signal c and the output signal g. Entropy $H(g)$ measures the uncertainty of the output signal, conditional entropy $H(g|c)$ is the entropy of the output g given the knowledge of the input c. Mutual information $I(c;g)$ is defined as the reduction in the uncertainty of g due to the knowledge of c. Stated differently, it is a measure of the mutual dependence of the input and output variables. Thus $I(c;g) = H(g) - H(g|c)$[131].

If we measure the information in bits, then

$$I(c;g) = \sum_{c,g} p(c,g) log_2 \frac{p(c,g)}{p(c)p(g)}. \tag{5.1}$$

where p(c,g) is the joint probability mass function of variable c and g.

Using information theoretic approach, one can investigate the possibility of reliable communication over unreliable channels. In order to frame this, we need to know the temporal characteristics of the input signals as well as that of the output signals, which I have done in my dissertation study. However, calculating the mutual information is in general formidable. Until recently, it can only be obtained analytically by assuming that the input signals obey Gaussian statistics[130] which provides a great challenge for the future studies.

The post-transcriptional control of protein expression in animal cells by microRNAs plays an important role in almost every cellular process and changes in their expression may underlie developmental disorders and diseases such as cancer. miRNAs base-pair with seed sequences in the $3'$ UTRs of their target mRNA and block steps in the initiation of transcription, sequestering mRNAs into sites of repression or

by accelerating mRNA decay [64]. As a result, miRNAs reduce mRNA and protein abundance, often modestly and sometimes sharply [90, 91]. Genome-wide studies have shown that miRNAs target many transcription factors, which in some cases regulate their own transcription. Given the importance of miRNAs in cellular function an analysis of the impact of miRNA-mediated regulation on the mean levels and fluctuations of genetic circuits is vital. Here, we have shown that depending on the mode of miRNA action, negative feedback by miRNA can have differential impact on the noise levels of protein expression. In particular, if miRNAs act in a stoichiometric fashion, whereby both the target mRNA and miRNA are removed from the population into an inactive pool (Figure 3.1), then negative feedback regulation by miRNA largely amplifies the intrinsic noise in the system, leading to long-tailed distributions of transcriptional factor numbers. Our simulations show that this enhancement of protein number fluctuations is sensitive to environmental factors as seen in Figures 3.5 and Figure 3.6. However, if miRNAs act catalytically to repress protein synthesis, the net effect is to reduce variability in protein levels, as would be conventionally expected for a negative feedback circuits.

A number of experimental observations justify the distinction of the modes of miRNA action made in our models. Early studies of miRNA effects seemed to reveal that mRNA degradation was minimal but protein expression was reduced consistent with a catalytic mode of regulation [91]. This could result from imperfect seed sequence complementarity between the miRNA and mRNA, the presence of multiple miRNA targets, weak and reversible association of the target mRNA with the RISC machinery, rapid accumulation of the RISC/mRNA complex into P-bodies or accumulation in stress granules accompanied by the release of the miRNA. More recent studies have shown that mRNA degradation is significant [91]. These could arise from a higher degree of complementarity of miRNA seed sequence in the mRNA $3'$ UTR, multiple pairing locations, post-translational modifications of the

RISC machinery that enhance binding of miRNA/mRNA and subsequent translational repression and P-body accumulation. These latter effects are best represented mathematically by a sequestration model where both miRNA and mRNA are stoichiometrically degraded. In order to keep the models relatively simple and to gain intuition, we have abstracted many of the intermediate steps, modeling component processes as first and second-order reactions. Such coarse-grained representations have been quite successful in elucidating many aspects of deterministic and stochastic gene networks [9, 10, 75, 92]. We have neglected additional aspects of miRNA biogenesis and function, such as multiple miRNA seed sequences on the same target mRNA, delays in processing mature miRNA from precursor transcripts etc. However, we expect that consideration of these steps would not qualitatively change our results.

Our studies in Chapter 3 expand the repertoire of miRNA action in gene circuits that govern cell fate specification and commitment during development, processes where miRNA function was first highlighted. Our modeling studies suggest a new role for miRNA-based feedback regulation, namely, by modulating the levels of TFs at the level of translational repression, miRNAs can drive large fluctuations in TF levels across the population. In turn, these fluctuations can drive the expression of different constellations of genes across the population, thereby allowing the expression of multiple cellular phenotypes in a uniform precursor population. Given the extensive complexity of the component processes, cells may be able to tune the manner of miRNA-based feedback, from stoichiometric repression to catalytic repression to tune the level of protein number fluctuations in gene circuits and consequently drive stochastic cell fate choices. A number of recent studies suggest that such tuning may be operant in cells. RISC protein phosphorylation can control the loading of miRNAs [93]. Alternately, the seed site for miRNA binding on the target mRNA may be made more accessible [94]. Thus, cells may control expression noise in

miRNA-based negative feedback circuits to determine cell fates in different contexts. In general, cell fate decisions during development are robust, in order to generate reproducible body plans. However, in certain cases, cell fate decisions are made at random, generating cell fate diversity. Diversified cell fates in a homogeneous progenitor population increases the spectrum of responses to environmental stimuli. One example is the choice of Rhodopsin type during photoreceptor differentiation in the Drosophila eye [95]. Thus, miRNA based translational repression may serve as an important mechanism that controls fluctuations of protein number promoting cell fate diversity.

Alternative Splicing is a key process that contributes to the creation of phenotypic complexity among higher eukaryotes. My Ph.D. work presented in Chapter 4 suggests an important role of splicing factor diffusion on regulates the noise propagation in transcription as well as control mRNA isoforms ratio. The diffusive influence to the total noise in alternative splicing has only recently been recognized as significant. Our pioneer work indicated a theoretical way to analyze the role of diffusion. In our mathematical model, we brutally treat the splicing factor binding site as a single point. In the future, it can be expand to be an interaction volume. Also we assume splicing factor diffuse freely in nucleus, however, the diffusion constant of splicing factor in nucleoplasm and in speckle are different. Splicing factor can also being produced and degrade. All these point to the basic fundamental problems: how biological processes, regulate and propagate the noise so as to perform precise biological function with cell to cell variability presents.

Appendix A

Supplementary Information for Chapter 2

A.1 Bursty Transcription Model

A.1.1 Model Description

Transcription in cells is actually discontinuous. Bursty transcription means gene can be turned "on" or "off" with the regulation of gene regulatory protein, also known as transcription factor, which act as switches. Gene can be activated when gene regulatory protein recognize and bind to a specific regulatory region near the gene to be transcribed. Mathematical model of transcription for this situation can be described as follows:

$$\begin{aligned} D_I &\xrightarrow{\lambda} D_A && \text{activated gene} \\ D_A &\xrightarrow{\mu} D_I && \text{repressed gene} \\ D_A &\xrightarrow{\nu} D_A + \text{mRNA} && \text{transcription} \\ \text{mRNA} &\xrightarrow{\delta} \phi && \text{mRNA degradation} \end{aligned} \tag{A.1}$$

D_I represents a repressed gene, D_A stands for an activated gene. The reaction

rates are λ, μ, ν and δ respectively. For simplicity, we assume all the rates are constant. The phase space of this system can be written as

$$S = \{(i, m) \quad i \in \{0, 1\}, m \in \mathcal{N}\}. \tag{A.2}$$

In any state, $i = 0$ denotes that the gene is inactive or cannot transcribe mRNA while $i = 1$ implies that the gene is active and can transcribe mRNA. m denotes the number of mRNA which are produced at a rate ν and degraded at a rate $\delta \times m$. The gene remains active during a random time distributed exponentially with parameter μ on average. So from $(0, m)$ the system can jump to

$$\begin{aligned} (1, m) \quad &\text{with rate} \quad \lambda \\ (0, m-1) \quad &\text{with rate} \quad \delta m \end{aligned}$$

From state $(1, m)$ the system can jump to

$$\begin{aligned} (0, m) \quad &\text{with rate} \quad \mu \\ (1, m+1) \quad &\text{with rate} \quad \nu \\ (1, m-1) \quad &\text{with rate} \quad \delta m \end{aligned}$$

This set of transitions defines a Markov process on S as before. Let $p_{0,m}$ be the probability that the gene is in the inactive state and the number of mRNA is m. Also, $p_{1,m}$ be the analogous probability when the gene is in the active state. So we have the following system of equations that govern these probabilities:

$$\begin{aligned} \frac{dp_{0,m}}{dt} &= -(\lambda + m\delta)p_{0,m}(t) + \delta(m+1)p_{0,m+1}(t) + \mu p_{1,m}(t) \\ \frac{dp_{1,m}}{dt} &= -(\mu + \nu + m\delta)p_{1,m}(t) + \delta(m+1)p_{1,m+1}(t) + \nu p_{1,m-1}(t) + \lambda p_{0,m}(t) \\ \frac{dp_{1,0}}{dt} &= -(\mu + \nu)p_{1,0}(t) + \delta p_{1,1}(t) + \lambda p_{0,0}. \end{aligned} \tag{A.3}$$

The Generating function for this process:

$$G_0(z,t) = \sum_{m=0}^{\infty} z^m p_{0,m}(t) \tag{A.4}$$

$$G_1(z,t) = \sum_{m=0}^{\infty} z^m p_{1,m}(t). \tag{A.5}$$

The sum $G(z,t) = G_0(z,t) + G_1(z,t)$ is the generating function for the total mRNA probability distribution at time t, so $G(1,t) = 1$.

Multiply the equations A.3 by z^m and sum m from zero to infinity to obtain:

$$\frac{\partial G_0(z,t)}{\partial t} = -\lambda G_0(z,t) + \mu G_1(z,t) + \delta(1-z)\frac{\partial G_0(z,t)}{\partial z} \tag{A.6a}$$

$$\frac{\partial G_1(z,t)}{\partial t} = \lambda G_0(z,t) - \mu G_1(z,t) + \delta(1-z)\frac{\partial G_1(z,t)}{\partial z} - \nu(1-z)G_1(z,t). \tag{A.6b}$$

We can solve for the case $z = 1$ from the above equations as

$$G_0(1,t) = \sum_{m=0}^{\infty} p_{0,m}(t) = \frac{\mu}{\lambda+\mu} + \left(\gamma_0 - \frac{\mu}{\lambda+\mu}\right) \mathrm{e}^{-(\lambda+\mu)t} \tag{A.7a}$$

$$G_1(1,t) = \sum_{m=0}^{\infty} p_{1,m}(t) = \frac{\lambda}{\lambda+\mu} + \left(\gamma_1 - \frac{\lambda}{\lambda+\mu}\right) \mathrm{e}^{-(\lambda+\mu)t}. \tag{A.7b}$$

This gives us the probability distribution of the gene activity at time t. γ_0 and γ_1 are the probabilities that initially the system starts from inactive state $i = 0$ and active state $i = 1$ respectively, i.e., $\gamma_0 \equiv \sum_{m=0}^{\infty} p_{0,m}(t=0)$, $\gamma_1 \equiv \sum_{m=0}^{\infty} p_{1,m}(t=0)$. If we set the initial state to be inactive, then $\gamma_0 = 1$ and $\gamma_1 = 0$.

A.1.2 Steady State Solution

In order to solve for the steady state distribution, we can set the time derivatives to 0 in Eqn. A.6. Let $g_0(z) \equiv \lim_{t\to\infty} G_0(z,t)$, $g_1(z) \equiv \lim_{t\to\infty} G_1(z,t)$ and $g(z) \equiv$

$\lim_{t\to\infty} G(z,t) = g_0(z) + g_1(z)$.

$$\delta(z-1)\frac{\partial g_0(z)}{\partial z} = -\lambda g_0(z) + \mu g_1(z) \tag{A.8a}$$

$$\delta(z-1)\frac{\partial g_1(z)}{\partial z} = \lambda g_0(z) - \mu g_1(z) + \nu(z-1)g_1(z). \tag{A.8b}$$

Note that this pair of equations are singular at $z = 1$ and we cannot directly solve it.

If $z = 1$, we can get $g_0(1) = \frac{\mu}{\lambda+\mu}$ and $g_1(1) = \frac{\lambda}{\lambda+\mu}$, which is the same results as in Eqn. A.7 with $t \to \infty$.

If $z \neq 1$, add the two equations together, we can get:

$$\delta\frac{\partial g(z)}{\partial z} = \nu g_1(z). \tag{A.9}$$

Differentiate the above equation and substitute Eqn. A.8 to get the following second order equation

$$\delta^2(z-1)g''(z) + \delta(\lambda+\mu-\nu(z-1))g'(z) - \lambda\nu g(z) = 0. \tag{A.10}$$

Let $a \equiv \frac{\lambda}{\delta}$, $b \equiv \frac{1}{\delta}(\lambda+\mu)$ and $y \equiv \frac{\nu}{\delta}(z-1)$, we can get:

$$yg''(y) + (b-y)g'(y) - ag(y) = 0. \tag{A.11}$$

This is a Kummer's equation (Confluent Hypergeometric Function). There are two linear independent solution for this equation: ${}_1F_1(a,b;y)$ and $U(a,b;y)$. The boundary condition is $g(0) = 1$.Because the second diverse when $y = 0$. So we can discard it and then get the exact solution

$$g(y) = {}_1F_1(a,b;y). \tag{A.12}$$

i.e.

$$g(z) = {}_1F_1(\frac{\lambda}{\delta}, \frac{\lambda+\mu}{\delta}; \frac{\nu}{\delta}(z-1)). \tag{A.13}$$

It is also possible to solve this using a power series expansion of $g(y) = \sum_{n=0}^{\infty} C_n y^n$.

From the boundary condition $g(y=0) = 1$, we can get $C_0 = 1$. So

$$g'(y) = \sum_{n=1}^{\infty} C_n n y^{n-1} \tag{A.14}$$

$$g''(y) = \sum_{n=2}^{\infty} C_n n(n-1) y^{n-2}. \tag{A.15}$$

Put them into Equation A.11, we can get

$$\sum_{n=0}^{\infty} \left\{ C_{n+1}(n+1)(n+b) - C_n(n+a) \right\} y^n = 0. \tag{A.16}$$

The coefficient of every order of y should be vanished, so we can get the recurrence relation as below:

$$C_0 = 1 \tag{A.17}$$

$$\frac{C_{n+1}}{C_n} = \frac{n+a}{(n+1)(n+b)}. \tag{A.18}$$

Thus we can get:

$$C_n = \frac{(a)_n}{(b)_n n!}, \quad g(y) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n n!} y^n \tag{A.19}$$

where $(a)_0 \equiv 1$,$(a)_n \equiv a(a+1)(a+2)\ldots(a+n-1)$.

Because the confluent hypergeometric series have the form:

$$ {}_1F_1(a,b;y) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n n!} y^n, \tag{A.20}$$

we can get $g(y) = {}_1F_1(a,b;y)$, which is just the same as Eqn. A.12.

In sum, the exact solution of the steady state generating function is:

$$g(z) = {}_1F_1(a,b;\frac{\nu}{\delta}(z-1)) = \sum_{n=0}^{\infty} \frac{(a)_n}{(b)_n n!} (\frac{\nu}{\delta})^n (z-1)^n. \tag{A.21}$$

From this solution, we can get all the moment of the mRNA because

$$\frac{\partial^n g(z)}{\partial z^n} \mid_{z=1} = \langle m(m-1)\dots(m-n+1)\rangle. \tag{A.22}$$

The first three term are listed below:

$$\langle m\rangle = \frac{\nu\lambda}{\delta(\mu+\lambda)} \tag{A.23}$$

$$\langle m(m-1)\rangle = \frac{\nu^2\lambda(\lambda+\delta)}{\delta^2(\mu+\lambda)(\mu+\lambda+\delta)} \tag{A.24}$$

$$\langle m(m-1)(m-2)\rangle = \frac{\nu^3\lambda(\lambda+\delta)(\lambda+2\delta)}{\delta^3(\mu+\lambda)(\mu+\lambda+\delta)(\mu+\lambda+2\delta)}. \tag{A.25}$$

So the mean mRNA level for steady state is

$$\langle m\rangle = \frac{\nu\lambda}{\delta(\mu+\lambda)}, \tag{A.26}$$

and the Fano factor is

$$F = \frac{\delta m^2}{\langle m\rangle} = \frac{\langle m^2\rangle - \langle m\rangle^2}{\langle m\rangle} = 1 + \frac{\nu\mu}{(\mu+\lambda)(\mu+\lambda+\delta)}. \tag{A.27}$$

As shown in Figure 2.2, the mean for bursty gene is always lower than the mean for the gene which is always active (u is always equal to 0). The Fano factor is always larger than the active gene whose Fano factor is always one. When the activation rate and inactivation rate are the same (i.e., $\mu = \lambda$), the mean for bursty gene is half the value of that for activated gene. While the inactivation rate increasing, the level of the mean mRNA number per cell decreases, however, Fano factor first increase, then decrease. Especially when $\mu = \sqrt{\lambda(\lambda+\delta)}$, Fano factor achieve maximum: $F = 1 + \frac{v}{(\sqrt{\lambda}+\sqrt{\lambda+\delta})^2}$.

A.1.3 Complex Analysis to get Steady State Probability Distribution

Below we will use the complex analysis to calculate the probability distribution for any mRNA number m at steady state. First, we need to generalize the variable z to be a complex variable.

Cauchy-Goursat theorem tells us that if $f(z)$ is analytic inside and on a simple closed curve C, then

$$\oint_C f(z)dz = 0. \tag{A.28}$$

As important as the above theorem is, Cauchy integral formula is at least equally important in complex analysis about line integrals for holomorphic functions in the complex plane. It says that if a function $f(z)$ is analytic on a simple closed curve C and in the region enclosed by C, if a is any point inside C, then

$$\oint_C \frac{f(z)}{z-a}dz = 2\pi i f(a). \tag{A.29}$$

Cauchy integral formula can also be generalized to get the nth derivative of $f(a)$.

$$f^{(n)}(a) = \frac{n!}{2\pi i}\oint_C \frac{f(z)}{(z-a)^{n+1}}dz, \quad n = 0, 1, 2, ... \tag{A.30}$$

Based on the above formula, if $f(z)$ can be represented by a Laurent series of the form:

$$f(z) = \sum_{n=0}^{\infty} a_n(z-a)^n + \sum_{n=0}^{\infty} \frac{b_n}{(z-a)^n}. \tag{A.31}$$

then the coefficients are:

$$a_n = \frac{1}{2\pi i}\oint_C \frac{f(\zeta)}{(\zeta-a)^{n+1}}d\zeta \tag{A.32}$$

$$b_n = \frac{1}{2\pi i}\oint_C \frac{f(\zeta)}{(\zeta-a)^{-n+1}}d\zeta. \tag{A.33}$$

Due to residue theorem, the coefficient of $\frac{1}{z-a}$ in the Laurent series of $f(z)$ (i.e., b_1) is called the residue of $f(z)$ at $z = a$. Thus

$$a_n = \text{Residue}[\frac{f(z)}{(z-a)^{n+1}}]. \tag{A.34}$$

Thus

$$p_m(t \to \infty) = \text{Residue}[\frac{g(z)}{z^{m+1}}]. \tag{A.35}$$

The full analytical expression of $p_m(t \to \infty)$ can be calculated as following:

$$p_m(t \to \infty) = \frac{1}{m!}\frac{(a)_m}{(b)_m}(\frac{v}{\delta})^m {}_1F_1(m+a, m+b; -\frac{v}{\delta}). \tag{A.36}$$

where $(a)_0 \equiv 1$ and $(a)_m \equiv a(a+1)(a+2)\dots(a+m-1)$, as defined previously in confluent hypergeometric series.

Using the complex analysis, we successfully get the full list of steady state probability distribution for mRNA.

A.1.4 Time Dependent Solution

$$\frac{\partial G_0(z,t)}{\partial t} = -\lambda G_0(z,t) + \mu G_1(z,t) + \delta(1-z)\frac{\partial G_0(z,t)}{\partial z} \tag{A.37a}$$

$$\frac{\partial G_1(z,t)}{\partial t} = \lambda G_0(z,t) - \mu G_1(z,t) + \delta(1-z)\frac{\partial G_1(z,t)}{\partial z} - \nu(1-z)G_1(z,t). \tag{A.37b}$$

Next we try to get the time dependent solution. How can we deduct time-dependent solution from the above equations of generating function?

First, we can do the following change of variables

$$x = \frac{\nu}{\delta}(z-1) \tag{A.38}$$

$$w = (1-z)e^{-\delta t}. \tag{A.39}$$

So $G(z,t) \rightarrow \overline{G}(x,w)$. Therefore, derivatives in the transformed coordinates become

$$\frac{\partial}{\partial t} = -\delta w \frac{\partial}{\partial w} \tag{A.40}$$

$$\frac{\partial}{\partial z} = \frac{\nu}{\delta}\left(\frac{\partial}{\partial x} + \frac{w}{x}\frac{\partial}{\partial w}\right). \tag{A.41}$$

The equations now become

$$\delta x \frac{\partial \overline{G}_0}{\partial x} = -\lambda \overline{G}_0 + \mu \overline{G}_1 \tag{A.42}$$

$$\delta x \frac{\partial \overline{G}_1}{\partial x} = \lambda \overline{G}_0 - \mu \overline{G}_1 + \delta x \overline{G}_1. \tag{A.43}$$

Note that these equations are independent of w as a consequence of the transformation. The dependence on w will be found in the boundary conditions. Adding these two equations, we have

$$\frac{\partial \overline{G}}{\partial x} = \overline{G}_1. \tag{A.44}$$

Now, we can differentiate the Eqn. A.44 w.r.t. x and using Eqn. A.43 and Eqn. A.44 to obtain

$$x\frac{\partial^2 \overline{G}}{\partial x^2} + (b-x)\frac{\partial \overline{G}}{\partial x} - a\overline{G} = 0. \tag{A.45}$$

where $b \equiv \frac{\lambda+\mu}{\delta}$ and $a \equiv \frac{\lambda}{\delta}$. This is a Confluent hypergeometric equation. The solutions of this equation can be written in the form

$$\overline{G} = A(w)_1F_1(a,b;x) + B(w)x^{(1-b)}\ {}_1F_1(1+a-b, 2-b; x). \tag{A.46}$$

If we assume $b < 1$, then the second term is valid when $x = 0$, however, we will find out b doesn't necessary to be less than 1 after using the boundary conditions to fix the coefficient $B(w)$. The Boundary conditions for $G(z,t)$ are:

$$\begin{cases} G(z,0) = 1 \\ G(1,t) = 1 \end{cases} \tag{A.47}$$

Transform the coordinate into $\overline{G}(x,w)$, the boundary conditions become

$$\begin{cases} \overline{G}(-\frac{v}{\delta}w, w) = 1 \\ \overline{G}(0,0) = 1 \end{cases} \tag{A.48}$$

During the transformation $G(1,t) \to \overline{G}(0,0)$, the boundary condition degenerate from a line to a dot. So information is lost during the transformation. Also the Jacobian determinant (Eqn. A.49) is zero at boundary $(1,t)$.

$$\frac{\partial(x,w)}{\partial(z,t)} = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial t} \\ \frac{\partial w}{\partial z} & \frac{\partial w}{\partial t} \end{vmatrix} = \begin{vmatrix} \frac{\nu}{\delta} & 0 \\ -e^{-\delta t} & \delta(z-1)e^{-\delta t} \end{vmatrix} = \nu(z-1)e^{-\delta t}. \tag{A.49}$$

In order to solve the problem, we need to find another boundary condition. Fortunately, we find from Eqn. A.44, while $t = 0$, $\frac{\partial \overline{G}}{\partial x} = 0$. The boundary conditions we use can be written as follows:

$$\begin{cases} \overline{G}(-\frac{v}{\delta}w, w) = 1 \\ \frac{\partial \overline{G}(x,w)}{\partial x}|_{x=-\frac{v}{\delta}w} = 0 \end{cases} \tag{A.50}$$

Introduce $U(a,b;x) \equiv x^{(1-b)}\ {}_1F_1(1+a-b, 2-b; x)$ and $U'(a,b;x) \equiv \frac{\partial U(a,b;x)}{\partial x}$.

Due to the properties of the confluent hypergeometric function:

$$\frac{\mathrm{d}^n}{\mathrm{d}z^n}{}_1F_1(a,b;z) = \frac{(a)_n}{(b)_n}{}_1F_1(a+n, b+n; z), \tag{A.51}$$

$$\frac{\mathrm{d}^n}{\mathrm{d}z^n}\left(z^{b-1}{}_1F_1(a,b;z)\right) = (b-n)_n z^{b-n-1}{}_1F_1(a, b-n; z), \tag{A.52}$$

we can get the following relations:

$$\frac{\partial_1 F_1(a,b;x)}{\partial x} = \frac{a}{b}{}_1F_1(a+1,b+1,x) \tag{A.53}$$

$$\frac{\partial U(a,b;x)}{\partial x} = (1-b)U(a+1,b+1;x) \tag{A.54}$$

Apply the boundary condition (Equation A.50) to the general solution of generating function (Equation A.46), we can get the following equations for solving the coefficient $A(w)$ and $B(w)$:

$$\begin{cases} A(w){}_1F_1(a,b;-\dfrac{\nu w}{\delta}) + B(w)U(a,b;-\dfrac{\nu w}{\delta}) = 1 \\ A(w)\dfrac{a}{b}{}_1F_1(a+1,b+1;-\dfrac{\nu w}{\delta}) + B(w)(1-b)U(a+1,b+1;-\dfrac{\nu w}{\delta}) = 0 \end{cases} \tag{A.55}$$

Define $\theta(w) = (1 - \frac{\nu w}{\delta}\frac{a}{b(b-1)}\frac{{}_1F_1(a+1,b+1;-\frac{\nu w}{\delta}){}_1F_1(1+a-b,2-b;-\frac{\nu w}{\delta})}{{}_1F_1(a,b;-\frac{\nu w}{\delta}){}_1F_1(1+a-b,1-b;-\frac{\nu w}{\delta})})^{-1}$, then we can get:

$$\begin{cases} A(w) = \dfrac{\theta(w)}{{}_1F_1(a,b;-\frac{\nu w}{\delta})} \\ B(w) = \dfrac{1-\theta(w)}{U(a,b;-\frac{\nu w}{\delta})} \end{cases} \tag{A.56}$$

So the generating function have the following form:

$$\overline{G}(x,w) = \frac{{}_1F_1(a,b;x)}{{}_1F_1(a,b;-\frac{\nu w}{\delta})}\theta(w) + (-\frac{\delta x}{\nu w})^{1-b}\frac{{}_1F_1(1+a-b,2-b;x)}{{}_1F_1(1+a-b,2-b;-\frac{\nu w}{\delta})}(1-\theta(w)). \tag{A.57}$$

From Eqn. A.57, we can find the value of b is not confined in region $(0,1)$, b can be any positive number because x/w is always positive.

Next, transform $\overline{G}(x,w)$ back to $G(z,t)$ using equation A.39:

$$G(z,t) = \frac{{}_1F_1(a,b;\frac{\nu}{\delta}(z-1))}{{}_1F_1(a,b;\frac{\nu}{\delta}(z-1)e^{-\delta t})}\theta(z,t)$$
$$+ \frac{{}_1F_1(1+a-b,2-b;\frac{\nu}{\delta}(z-1))}{{}_1F_1(1+a-b,2-b;\frac{\nu}{\delta}(z-1)e^{-\delta t})}(1-\theta(z,t))e^{(1-b)\delta t}. \quad \text{(A.58)}$$

where

$$\theta(z,t) = \big(1 + \frac{aX}{b(b-1)}\frac{{}_1F_1(a+1,b+1;X){}_1F_1(1+a-b,2-b;X)}{{}_1F_1(a,b;X){}_1F_1(1+a-b,1-b;X)}\big)^{-1}. \quad \text{(A.59)}$$

$X \equiv X(z,t) = \frac{\nu}{\delta}(z-1)e^{-\delta t}$. It's defined here to shorten the expression.

Eqn. A.58 satisfies the boundary condition (A.47) and b can be any positive value here. When $t \to \infty$, $\theta(z, t \to \infty) \to 1$, thus $G(z, t \to \infty) \to {}_1F_1(a,b;\frac{\nu}{\delta}(z-1))$, which is the same as Eqn. A.21.

From the time-dependent generating function, we can get all the time-dependent moment with the relation:

$$\frac{\partial^n G(z,t)}{\partial z^n}\ |_{z=1} = \langle m(t)(m(t)-1)\dots(m(t)-n+1)\rangle. \quad \text{(A.60)}$$

and

$$\frac{\partial^n {}_1F_1(a,b;z)}{\partial z^n} = \frac{(a)_n}{(b)_n}{}_1F_1(a+n,b+n,z). \quad \text{(A.61)}$$

The first two terms are list below:

$$\langle m(t)\rangle = \frac{\nu a(b-1-be^{-\delta t}+e^{-b\delta t})}{\delta b(b-1)} \quad \text{(A.62)}$$

$$\langle m(t)(m(t)-1)\rangle = \frac{\nu^2 a}{\delta^2 b}\left\{\frac{2(a-b)}{(b-1)(b+1)}e^{-(b+1)\delta t} - \frac{2(a-b+1)}{(b-2)(b-1)}e^{-b\delta t}\right.$$
$$\left. + \frac{b(a-1)}{(b-2)(b-1)}e^{-2\delta t} + \frac{2a}{b-1}e^{-\delta t} + \frac{a+1}{b+1}\right\}. \quad \text{(A.63)}$$

The mean value is verified to be the same as we calculated from the mean field theory.

Thus the Fano factor $F = \frac{\delta m^2}{\langle m \rangle} = \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle}$ can be written as:

$$\begin{aligned} F =& 1 - \langle m(t) \rangle \\ &+ \frac{\nu}{\delta} \frac{\{ \frac{2(a-b)}{b+1} e^{-(b+1)\delta t} + \frac{2(b-a-1)}{b-2} e^{-b\delta t} + \frac{(a-1)b}{b-2} e^{-2\delta t} - 2ae^{-\delta t} + \frac{(a+1)(b-1)}{b+1} \}}{b(1-e^{-\delta t}) - (1-e^{-b\delta t})}. \end{aligned} \quad \text{(A.64)}$$

Where $a = \frac{\lambda}{\delta}$, $b = \frac{\lambda+\mu}{\delta}$. It's exactly the same as Eqn. A.23 when $t \to \infty$.

A.2 Basic Idea of Computer Simulation

First, use GSL random number library to generate two random numbers r_1,r_2 uniformly distributed from $(0, 1]$. the first random number is for the time interval between two sequential events. $\Delta t = -\ln r_1$, so δt can range from zero to infinity. As mentioned in the introduction, the four events, mRNA and protein synthesize and degradation have different rate. We can arrange them as shown in diagram A.1, then use the random number r_2 to choose which one to occur after the time interval δt.

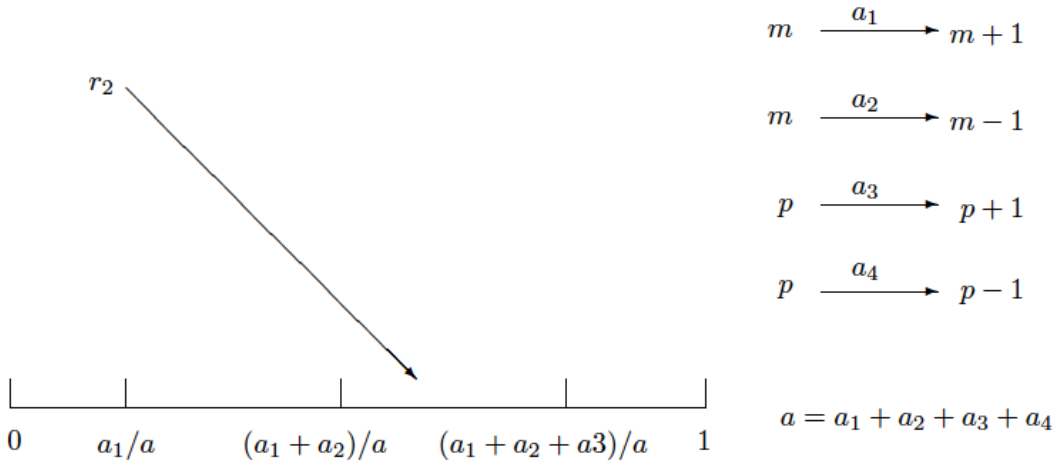


FIGURE A.1: scheme of the principal of computer simulation

Appendix B

Supplementary Information for Chapter 3

B.1 Sequestration Model and Kinetic Suppression Model are Limiting Cases of a Full Model

Here we show that our two models (sequestration and kinetic suppression) can be obtained as the limiting cases of a generalized model of miRNA-based feedback circuit that incorporates the salient aspects of the biology of miRNA action. As described by Levine et al. [132], we consider 3 species of mRNA: free mRNA (m), the silenced miRNA-mRNA complex (m^*), representing the accumulation of Ago proteins that prevent translation initiation of mRNA and processed mRNA (m^{**}), which accumulate into large ribonuclear particles. Assuming that the binding of miRNA to free mRNA and the subsequent assembly of Ago proteins ($m \rightarrow m^*$) and the processing of miRNA-mRNA complex ($m^* \rightarrow m^{**}$), is irreversible, we obtain the

following mass action equations:

$$\begin{cases} \dfrac{d\langle m\rangle}{dt} = \alpha_m - \gamma_m\langle m\rangle - \kappa\langle m\mu\rangle \\ \dfrac{d\langle m^*\rangle}{dt} = \kappa\langle m\mu\rangle - \eta\langle m^*\rangle - \gamma_m^*\langle m^*\rangle \\ \dfrac{d\langle m^{**}\rangle}{dt} = \eta\langle m^*\rangle - \gamma_m^{**}\langle m^{**}\rangle \\ \dfrac{d\langle \mu\rangle}{dt} = \alpha_\mu - \gamma_\mu\langle\mu\rangle - \kappa\langle m\mu\rangle + (1-q)\gamma_m^{**}m^{**} + \gamma_m^*\langle m^*\rangle \\ \dfrac{d\langle p\rangle}{dt} = k_p\langle m\rangle - \gamma_p\langle p\rangle. \end{cases} \tag{B.1}$$

While α-terms accounts for synthesis of RNA species, the γ-terms account for the degradation of RNA species and protein respectively. Here κ denotes the effective rate constant of miRNA-mRNA binding and Ago-complex assembly on the duplex. η denotes the rate of processing of this complex. k_p is the rate of translation of proteins. q denotes the probability for a miRNA to be co-degraded with the mRNA in the processed state. The limit $q = 1$ is the case that all the miRNA co-degrades with the mRNA. In this case, if we assume that m^* state is transient and as soon as mRNA gets into that state, it is processed rapidly into the m^{**} state, we can ignore the degradation of m^*. The mass action equations for m, μ and p can be written as:

$$\begin{cases} \dfrac{d\langle m\rangle}{dt} = \alpha_m - \gamma_m\langle m\rangle - \kappa\langle m\mu\rangle \\ \dfrac{d\langle \mu\rangle}{dt} = \alpha_\mu - \gamma_\mu\langle\mu\rangle - \kappa\langle m\mu\rangle \\ \dfrac{d\langle p\rangle}{dt} = k_p\langle m\rangle - \gamma_p\langle p\rangle. \end{cases} \tag{B.2}$$

which are exactly the same as the mass action equations in our sequestration model. Note that the processed or translationally silent states play no role in the feedback,

which is entirely determined by the fact that the ultimate fate of the miRNA-mRNA complex is accumulation into a translationally incompetent pool.

For $q = 0$, the case when miRNAs act catalytically to suppress translation from free mRNA, we assume (as in [132]), that the m^* and m^{**} species are near steady state. Further assuming that $\gamma_m \sim \gamma_m^* \sim \gamma_m^{**}$, which means the degradation rate of all mRNA species are the same, we can introduce a new variable: $m_{eff} = m + m^* + m^{**}$. This yields

$$\frac{d\langle m_{eff}\rangle}{dt} = \alpha_m - \gamma_m \langle m_{eff}\rangle \tag{B.3}$$

Again assuming that m is not correlated with the number of miRNA μ, $\langle m\mu\rangle = \langle m\rangle\langle\mu\rangle$ and the steady-state assumption for the intermediate species, we obtain:

$$\begin{cases} \dfrac{d\langle m_{eff}\rangle}{dt} = \alpha_m - \gamma_m \langle m_{eff}\rangle \\ \dfrac{d\langle \mu\rangle}{dt} = \alpha_\mu - \gamma_\mu \langle \mu\rangle \\ \dfrac{d\langle p\rangle}{dt} = k_p \langle m_{eff}\rangle \left(\dfrac{1}{1 + \frac{\kappa}{\gamma_m}\langle\mu\rangle}\right) - \gamma_p \langle p\rangle. \end{cases} \tag{B.4}$$

This is the same form as Eqn. 3.5 with $f(m, u) = \frac{k_p m}{1+\beta\mu}$, which functions as repression of translation.

B.2 Verifying the Linear Noise Approximation Method

Below, we verify that the linear noise approximation is applicable for the kinetic suppression model. Translational repression by miRNA in this model, $f(m, u)$ has the familiar Michaelis-Menten form $\frac{mk_p}{1+\beta\mu^n}$. For $n = 0$, translation is not repressed by the number of miRNA. In this case, the Fourier space variances can be calculated as

below:

$$\begin{cases} \left|\hat{\delta_m}(\omega)\right|^2 = \frac{4\pi\alpha_m}{{\gamma_m}^2+\omega^2} \\ \left|\hat{\delta_\mu}(\omega)\right|^2 = \frac{4\pi\{({\gamma_m}^2+\omega^2)[(\frac{\partial g(p)}{\partial p})^2 p\gamma_p+\mu\gamma_\mu({\gamma_p}^2+\omega^2)]+(\frac{\partial g(p)}{\partial p})^2{k_p}^2\alpha_m\}}{({\gamma_m}^2+\omega^2)({\gamma_\mu}^2+\omega^2)({\gamma_p}^2+\omega^2)} \\ \left|\hat{\delta_p}(\omega)\right|^2 = \frac{4\pi[p\gamma_p({\gamma_m}^2+\omega^2)+\alpha_m{k_p}^2]}{({\gamma_m}^2+\omega^2)({\gamma_p}^2+\omega^2)} \end{cases} \tag{B.5}$$

Because $\langle\delta x^2\rangle = \frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}\left|\hat{\delta_x}(\omega)\right|^2 d\omega$, we can get the exact solution as:

$$\begin{cases} \langle\delta m^2\rangle &= \dfrac{\alpha_m}{\gamma_m} \\ \langle\delta\mu^2\rangle &= \langle\mu\rangle + \dfrac{(\frac{\partial g(p)}{\partial p})^2\langle p\rangle}{\gamma_\mu(\gamma_\mu+\gamma_p)} + \dfrac{(\frac{\partial g(p)}{\partial p})^2{k_p}^2(\gamma_m+\gamma_\mu+\gamma_p)\alpha_m}{\gamma_m\gamma_\mu\gamma_p(\gamma_m+\gamma_\mu)(\gamma_m+\gamma_p)(\gamma_\mu+\gamma_p)} \\ \langle\delta p^2\rangle &= \dfrac{\alpha_m k_p}{\gamma_m\gamma_p}(1+\dfrac{k_p}{\gamma_m+\gamma_p}). \end{cases} \tag{B.6}$$

where $\langle m\rangle$, $\langle\mu\rangle$ and $\langle p\rangle$ is the steady state solution of the mass action equations. $\langle m\rangle = \frac{\alpha_m}{\gamma_m}$, $\langle\mu\rangle = g(\langle p\rangle)/\gamma_\mu$ and $\langle p\rangle = \frac{alpha_m k_p}{\gamma_m\gamma_p}$. The variance of m and p is consistent with the solution of gene expression without miRNA suppression [75].

For $n = 1$, we proceed by linearizing the Michaelis functions that link miRNA and protein abundance to the synthesis and repression of the protein and miRNA respectively. This is justified since number distributions of the relevant variables have finite widths and sample only a small region of the domain of the Michaelis functions, $f(m,\mu) = mF(\mu) = k_{rp} + k_1\mu + k_2 m$, where $k_1 = \frac{\partial F}{\partial\mu}|_{\langle\mu\rangle}\langle m\rangle$, $k_2 = F|_{\langle\mu\rangle}$ and $k_{rp} = -k_1\langle\mu\rangle$. Fourier transforming the linearized Langevin equations, we can write down the variance as:

$$\begin{cases} \left|\hat{\delta_m}(\omega)\right|^2 = \frac{4\pi\alpha_m}{{\gamma_m}^2+\omega^2} \\ \left|\hat{\delta_\mu}(\omega)\right|^2 = \frac{4\pi\alpha_\mu}{{\gamma_\mu}^2+\omega^2} \\ \left|\hat{\delta_p}(\omega)\right|^2 = 4\pi\{\frac{\alpha_m{k_2}^2}{(\gamma_m^2+\omega^2)(\gamma_m^2+\omega^2)} + \frac{\alpha_\mu{k_1}^2}{(\gamma_\mu^2+\omega^2)(\gamma_p^2+\omega^2)} + \frac{f(m,\mu)}{\gamma_p^2+\omega^2}\} \end{cases} \tag{B.7}$$

Rearranging and transforming back to the time domain yields:

$$\begin{cases} \langle \delta m^2 \rangle & = \dfrac{\alpha_m}{\gamma_m} \\ \langle \delta \mu^2 \rangle & = \dfrac{\alpha_\mu}{\gamma_m} \\ \langle \delta p^2 \rangle & = \dfrac{f(m,\mu)}{\gamma_p} + \dfrac{\alpha_\mu \gamma_m {k_1}^2 (\gamma_m + \gamma_p) + \alpha_m \gamma_\mu {k_2}^2 (\gamma_\mu + \gamma_p)}{\gamma_m \gamma_\mu \gamma_p (\gamma_m + \gamma_p)(\gamma_\mu + \gamma_p)}. \end{cases} \tag{B.8}$$

Further assuming that $\gamma_m = \gamma_\mu = \gamma$, we can get $\frac{\langle \delta p^2 \rangle}{\langle p \rangle} = 1 + \frac{k_1^2 \alpha_\mu + k_2^2 \alpha_m}{\alpha_m k_2 (\gamma + \gamma_p)}$, which is exactly the same as that derived by Komorowski et al. [133], using other methods for the deduction. In all, we can show that the linear noise approximation is a useful method for investigating the fluctuations in gene expression systems while expression noise is small. Naturally, this approximation breaks down for large values of negative feedback, where the mean levels fall and the variance is an appreciable fraction of the mean. Finally, we give the Fourier space variance for the $n = 2$ case that we have considered as:

$$\begin{cases} \left|\hat{\delta_m}(\omega)\right|^2 = \frac{4\pi\alpha_m}{{\gamma_m}^2 + \omega^2} \\ \left|\hat{\delta_\mu}(\omega)\right|^2 = \frac{4\pi\{(\gamma_m^2+\omega^2)[g_1^2 p\gamma_p + \mu\gamma_\mu(\gamma_p^2+\omega^2)] + f_1^2 g_1^2 \alpha_m\}}{(\gamma_m^2+\omega^2)[\omega^4 + (2f_2 g_1 + \gamma_\mu^2 + \gamma_p^2)\omega^2 + (f_2 g_1 - \gamma_\mu\gamma_p)^2]} \\ \left|\hat{\delta_p}(\omega)\right|^2 = \frac{4\pi\{(\gamma_m^2+\omega^2)[f_2^2 \mu\gamma_\mu + p\gamma_p(\gamma_\mu^2+\omega^2)] + f_1^2(\gamma_\mu+\omega^2)\alpha_m\}}{(\gamma_m^2+\omega^2)[\omega^4 + (2f_2 g_1 + \gamma_\mu^2 + \gamma_p^2)\omega^2 + (f_2 g_1 - \gamma_\mu\gamma_p)^2]}, \end{cases} \tag{B.9}$$

where $f_1 = \frac{\partial f(m,\mu)}{\partial m}$, $f_2 = \frac{\partial f(m,\mu)}{\partial \mu}$ and $g_1 = \frac{\partial g(p)}{\partial p}$. Since the time-domain expressions are lengthy and uninformative, we omit them.

Appendix C

Supplementary Information for Chapter 4

If all the parameters in scheme 4.2 are constant, we can use both the Master equation approach and the linear noise approximation approach to analyze the system. We can also calculate how splicing factor diffusion influent the noise properties of alternative splicing using linear noise approximation method and computer simulation.

C.1 The Master Equation Approach

Master equation is used to describe the time evolution of a system that can be modeled as being in exactly one of countable number of states at any given time, and where switching between states is treated probabilistically. It states that the change of probability of being in a given state depend on probabilities of transition to and from any other states in the system. If it can be directly solved, It provides the full probability distribution. However, this is not often the case because of the complexity of the system. Still we can use moment generating function to get the mean and variance of the distribution.

There are four variables in the alternative splicing model we proposed:

$\vec{x} = \{u, b, m_1, m_2\}$. The Master equation for these variables is:

$$\begin{aligned}\frac{\partial P(u,b,m_1,m_2,t)}{\partial t} &= vP(u-1,b,m_1,m_2,t) \\ &+ k^+c(u+1)P(u+1,b-1,m_1,m_2,t) + k^-(b+1)P(u-1,b+1,m_1,m_2,t) \\ &+ \alpha_1(u+1)P(u+1,b,m_1-1,m_2,t) + \alpha_2(b+1)P(u,b+1,m_1,m_2-1,t) \\ &+ \gamma_1(m_1+1)P(u,b,m_1+1,m_2,t) + \gamma_2(m_2+1)P(u,b,m_1,m_2+1,t) \\ &- (v + k^+cu + k^-b + \alpha_1 u + \alpha_2 b + \gamma_1 m_1 + \gamma_2 m_2))P(u,b,m_1,m_2,t). \end{aligned} \tag{C.1}$$

The moment generating function for this system is given by:

$$G(z_u, z_b, z_{m_1}, z_{m_2}, t) = \sum_{u,b,m_1,m_2=0}^{\infty} z_u^u z_b^b z_{m_1}^{m_1} z_{m_2}^{m_2} P(u,b,m_1,m_2,t). \tag{C.2}$$

Generating function is very useful for calculating moments because of the following properties:

$$\begin{aligned} G|_1 &= 1 & \frac{\partial G}{\partial x_i}|_1 &= \langle x_i \rangle \\ \frac{\partial^2 G}{\partial x_i^2}|_1 &= \langle x_i(x_i-1) \rangle & \frac{\partial^2 G}{\partial x_i \partial x_j}|_1 &= \langle x_i x_j \rangle, \end{aligned} \tag{C.3}$$

where $x_i \in \{u, b, m_1, m_2\}$ and $|_1$ means that the function is evaluated at $x_i = 1$ for all i.

Multiplying the Master equation above (Equation C.1) by $z_u^u z_b^b z_{m_1}^{m_1} z_{m_2}^{m_2}$ on both sides and sum over all possible value of these four variables, we can get:

$$\begin{aligned}\frac{\partial G}{\partial t} =& v(z_u - 1)G + k^+c(z_b - z_u)\frac{\partial G}{\partial z_u} + k^-(z_u - z_b)\frac{\partial G}{\partial z_b} + \alpha_1(z_{m1} - z_u)\frac{\partial G}{\partial z_u} \\ &+ \alpha_2(z_{m_2} - z_b)\frac{\partial G}{\partial z_b} + \gamma_1(1 - z_{m_1})\frac{\partial G}{\partial z_{m_1}} + \gamma_2(1 - z_{m_2})\frac{\partial G}{\partial z_{m_2}}. \end{aligned} \tag{C.4}$$

In steady state, $\frac{\partial G}{\partial t} = 0$. Resort Equation C.4 to the following form:

$$\begin{aligned} 0 = &(z_u - 1)(vG - k^+ c\frac{\partial G}{\partial z_u} + k^- \frac{\partial G}{\partial z_b} - \alpha_1 \frac{\partial G}{\partial z_u}) \\ &+ (z_b - 1)(k^+ c\frac{\partial G}{\partial z_u} - k^- \frac{\partial G}{\partial z_b} - \alpha_2 \frac{\partial G}{\partial z_b}) \\ &+ (z_{m_1} - 1)(\alpha_1 \frac{\partial G}{\partial z_u} - \gamma_1 \frac{\partial G}{\partial z_{m_1}}) \\ &+ (z_{m_2} - 1)(\alpha_2 \frac{\partial G}{\partial z_b} - \gamma_2 \frac{\partial G}{\partial z_{m_2}}). \end{aligned} \tag{C.5}$$

By taking the derivative with respect to x_i ($x_i \in \{u, b, m_1, m_2\}$) respectively and then let $z_u = z_b = z_{m_1} = z_{m_2} = 1$, we can get:

$$\begin{cases} v - (k^+ c + \alpha_1)\langle u\rangle + k^- \langle b\rangle = 0 \\ k^+ c\langle u\rangle - (k^- + \alpha_2)\langle b\rangle = 0 \\ \alpha_1\langle u\rangle - \gamma_1\langle m_1\rangle = 0 \\ \alpha_2\langle b\rangle - \gamma_2\langle m_1\rangle = 0. \end{cases} \tag{C.6}$$

So we got the steady state mean values as below:

$$\begin{cases} \langle u\rangle = \dfrac{v}{\alpha_1 + g\alpha_2} \\ \langle b\rangle = \dfrac{vg}{\alpha_1 + g\alpha_2} \\ \langle m_1\rangle = \dfrac{\alpha_1}{\gamma_1}\dfrac{v}{\alpha_1 + g\alpha_2} \\ \langle m_2\rangle = \dfrac{\alpha_2}{\gamma_2}\dfrac{vg}{\alpha_1 + g\alpha_2}, \end{cases} \tag{C.7}$$

where $g \equiv \frac{k^+ c}{\alpha_2 + k^-}$. It's the effective production rate of b from u.

In order to get the variance of the variables, we need to take a second-order derivative of Equation C.5 and then let $z_u = z_b = z_{m_1} = z_{m_2} = 1$. We also need to use the properties of generating function as described in Equation C.3.

$$
\begin{cases}
v\langle u\rangle - (k^+c+\alpha_1)(\langle u^2\rangle - \langle u\rangle) + k^-\langle ub\rangle = 0 \\
k^+c\langle ub\rangle - (k^- + \alpha_2)(\langle b^2\rangle - \langle b\rangle) = 0 \\
\alpha_1\langle um_1\rangle - \gamma_1(\langle m_1^2\rangle - \langle m_1\rangle) = 0 \\
\alpha_2\langle bm_2\rangle - \gamma_2(\langle m_2^2\rangle - \langle m_2\rangle) = 0 \\
v\langle b\rangle - (k^+c + k^- + \alpha_1 + \alpha_2)\langle ub\rangle + k^-(\langle b^2\rangle - \langle b\rangle) + k^+c(\langle u^2\rangle - \langle u\rangle) = 0 \\
v\langle m_1\rangle - (k^+c + \alpha_1 + \gamma_1)\langle um_1\rangle + k^-\langle bm_1\rangle + \alpha_1(\langle u^2\rangle - \langle u\rangle) = 0 \\
v\langle m_2\rangle - (k^+c + \alpha_1 + \gamma_2)\langle um_2\rangle + k^-\langle bm_2\rangle + \alpha_2\langle ub\rangle = 0 \\
k^+c\langle um_1\rangle - (k^- + \alpha_2 + \gamma_1)\langle bm_1\rangle + \alpha_1\langle ub\rangle = 0 \\
k^+c\langle um_2\rangle - (k^- + \alpha_2 + \gamma_2)\langle bm_2\rangle + \alpha_2(\langle b^2\rangle - \langle b\rangle) = 0 \\
\alpha_1\langle um_2\rangle + \alpha_2\langle bm_1\rangle - (\gamma_1+\gamma_2)\langle m_1m_2\rangle = 0.
\end{cases}
\tag{C.8}
$$

Although the above equation looks scaring, it's not hard to solve. There are 10 variables and 10 equations. We can easily get the variance of each variable as well as the covariance between variables.

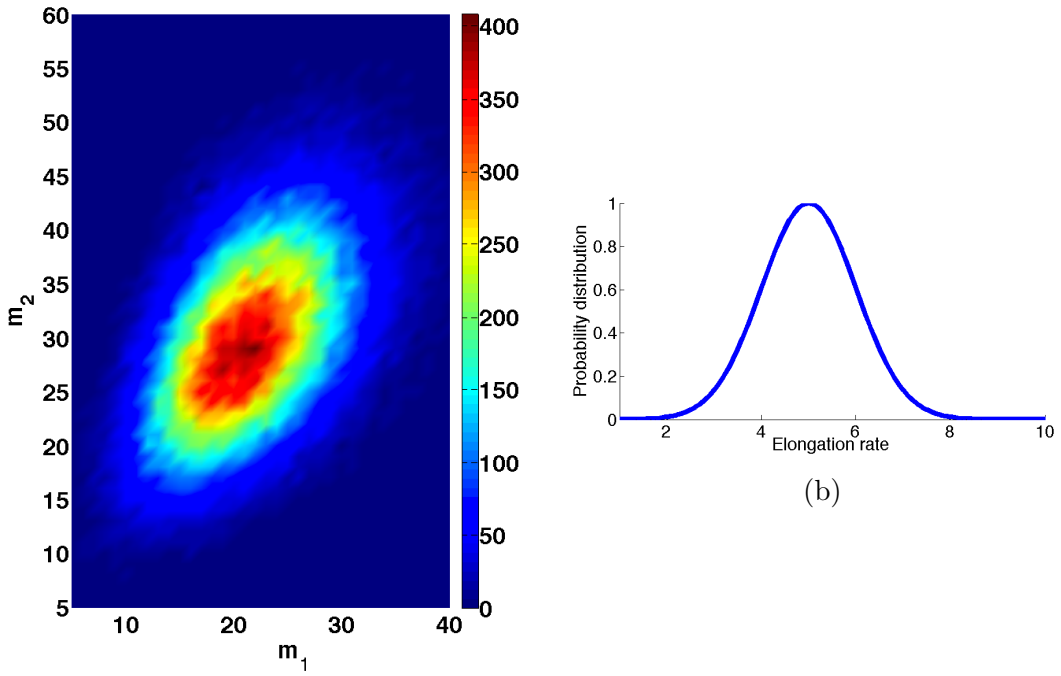
So the variances are:

$$
\begin{cases}
\sigma_u^2 = \langle u^2\rangle - \langle u\rangle^2 = \dfrac{v}{\alpha_1 + g\alpha_2} \\
\sigma_b^2 = \langle b^2\rangle - \langle b\rangle^2 = \dfrac{vg}{\alpha_1 + g\alpha_2} \\
\sigma_{m_1}^2 = \langle {m_1}^2\rangle - \langle m_1\rangle^2 = \dfrac{\alpha_1}{\gamma_1}\dfrac{v}{\alpha_1 + g\alpha_2} \\
\sigma_{m_2}^2 = \langle {m_2}^2\rangle - \langle m_2\rangle^2 = \dfrac{\alpha_2}{\gamma_2}\dfrac{vg}{\alpha_1 + g\alpha_2},
\end{cases}
\tag{C.9}
$$

The variance of each variable equals the mean of that variable. So all four variables are Poisson distributed.

We can also get the covariance of m_1 and m_2

$$
\mathrm{cov}(m_1, m_2) = \langle (m_1 - \langle m_1\rangle)(m_2 - \langle m_2\rangle)\rangle = 0 \tag{C.10}
$$



(a)

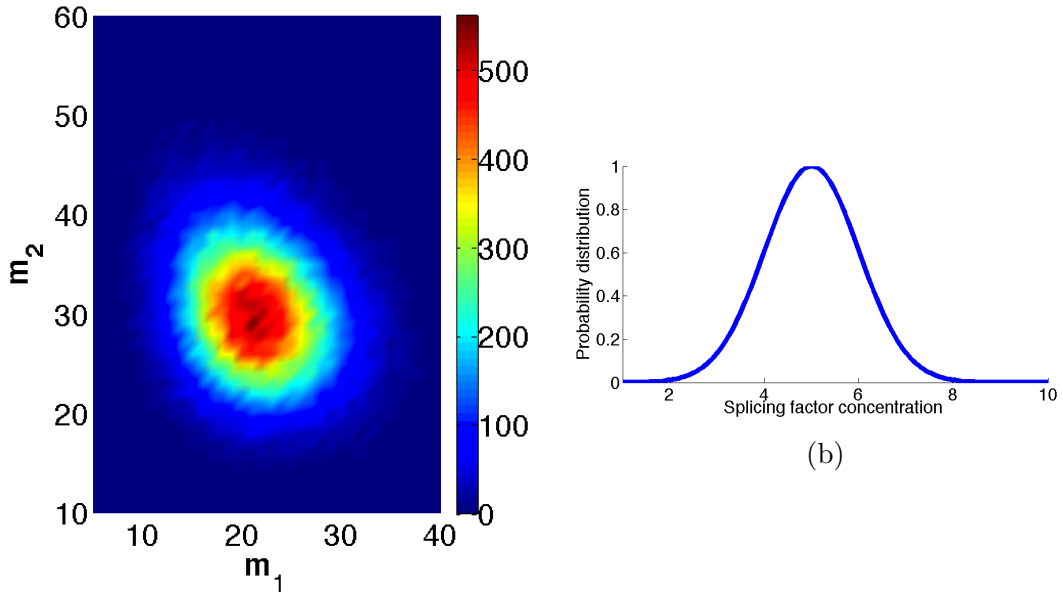
(b)

FIGURE C.1: Influence of elongation rate fluctuation on mRNA isoforms: (a) shows the correlation relation between m_1 and m_2 with Gaussian distributed elongation rate v. So two mRNA isoforms are positive correlated if there are fluctuation in elongation rate. (b) plots the distribution of the elongation rate. It's Gaussian distributed with mean=5, variance=1.

Thus m_1 and m_2 are two independent Poisson distributed random variables. Generally, while all the parameters are constant, u,b,m_1 and m_2 are four independent Poisson distributed random variables.

Figure C.1 is a simulation result showing that if the elongation rate is noisy(We assume it's Gaussian distributed), two mRNA isoforms change almost proportionally. This indicates a positive correlation relation between two mRNA isoforms.

If splicing factor concentration is Gaussian distributed (Figure C.2b), two mRNA isoforms will be aiti-correlated (Figure C.2a).



(a) (b)

FIGURE C.2: Influence of splicing factor concentration fluctuation on mRNA isoforms: (a) shows the correlation relation between m_1 and m_2 with Gaussian distributed splicing factor concentration c. (b) is the distribution of the splicing factor concentration. It's Gaussian distributed with mean=5, variance=1.

C.2 The Linear Noise Approximation Approach

C.2.1 When Splicing Factor c is a Constant

We can also use linear noise approximation method to solve this problem. The basic concept is assuming the system has already reached steady state. Due to the nature of stochasticity, there will be small fluctuation of each molecular species around their mean. This is an elegant way to calculation variation. The basic idea is to first write Langevin equations for the system, then linearize the Langevin equations around their steady states to get the equations for the displacements away from the mean values. Afterwards, use Fourier Transform to calculate power spectrum, then inverse Fourier Transform to get the variance of the variables. More specific calculation is provided as below.

For the alternative splicing model described in Figure 4.2, there are 4 molecular

species $\{u, b, m_1, m_2\}$ interact inside some fixed volume Ω and at constant temperature through 7 reaction channels $\{R_1, R_2, \ldots R_7\}$. These channels thus introduce 7 independent white noises into the system.

Some important vectors and matrices as listed:

- **Molecular vector**: $\vec{x} \equiv (u, b, m_1, m_2)^T$;
- **Reaction rate vector**: $\vec{a} \equiv (v, k^+cu, k^-b, \alpha_1 u, \alpha_2 b, \gamma_1 m_1, \gamma_2 m_2)^T$,

  $a_j(\vec{x})dt$ is the probability given $\vec{X}(t) = \vec{x}$, that reaction R_j will occur in the next infinitesimal time interval $[t, t + dt)$, (j=1,2, ... 7);
- **White noise vector**: $\vec{\Gamma} \equiv (\Gamma_1, \Gamma_2, \Gamma_3, \Gamma_4, \Gamma_5, \Gamma_6, \Gamma_7)^T$,

  $\Gamma_j(t)$ are temporally uncorrelated, statistically independent Gaussian white noise introduced by reaction R_j;
- **State change matrix**: **S**,

  S_{ij} represents the change in the number of x_i produced by R_j reaction (i=1,2,3,4; j=1,2, ... 7) ;
- **Noise amplitude matrix $\mathbf{\Psi}$**,

  Ψ_{ij} is the changes in the number of x_i molecules produced by the Gaussian white noise Γ_j from Reaction R_j.

$$\Psi_{ij} = S_{ij}\sqrt{a_j(\vec{x}(t))}. \tag{C.11}$$

From the above definitions, we can get:

$$\mathbf{S} = \begin{pmatrix} 1 & -1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} \tag{C.12}$$

$$\mathbf{\Psi} = \begin{pmatrix} \sqrt{v} & -\sqrt{k^+cu} & \sqrt{k^-b} & -\sqrt{\alpha_1 u} & 0 & 0 & 0 \\ 0 & \sqrt{k^+cu} & -\sqrt{k^-b} & 0 & -\sqrt{\alpha_2 b} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\alpha_1 u} & 0 & -\sqrt{\gamma_1 m_1} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\alpha_2 b} & 0 & -\sqrt{\gamma_2 m_2} \end{pmatrix} \tag{C.13}$$

So the Langevin equation can be written as:

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = \sum_{j=1}^{7} S_{ij} a_j(\vec{x}(t)) + \sum_{j=1}^{7} \Psi_{ij} \Gamma_j(t) \qquad (i = 1, 2, 3, 4) \tag{C.14}$$

This is equivalent to the following equations if we write it term by term:

$$\begin{cases} \dfrac{\mathrm{d}u}{\mathrm{d}t} = v - (k^+c + \alpha_1)u + k^-b + (\sqrt{v}\Gamma_1 - \sqrt{k^+cu}\Gamma_2 + \sqrt{k^-b}\Gamma_3 - \sqrt{\alpha_1 u}\Gamma_4) \\ \dfrac{\mathrm{d}b}{\mathrm{d}t} = k^+cu - (k^- + \alpha_2)b + (\sqrt{k^+cu}\Gamma_2 - \sqrt{k^-b}\Gamma_3 - \sqrt{\alpha_2 b}\Gamma_5) \\ \dfrac{\mathrm{d}m_1}{\mathrm{d}t} = \alpha_1 u - \gamma_1 m_1 + (\sqrt{\alpha_1 u}\Gamma_4 - \sqrt{\gamma_1 m_1}\Gamma_6) \\ \dfrac{\mathrm{d}m_2}{\mathrm{d}t} = \alpha_2 b - \gamma_2 m_2 + (\sqrt{\alpha_2 b}\Gamma_5 - \sqrt{\gamma_2 m_2}\Gamma_7). \end{cases} \tag{C.15}$$

Because $\Gamma_j(t)$ ($j = 1, 2, \ldots 7$) are temporally uncorrelated, statistically independent Gaussian white noises. They have the following time correlation property:

$$\langle \Gamma_i(t)\Gamma_j(t') \rangle = \delta_{ij}\delta(t - t'), \tag{C.16}$$

where on the right of the equation, the first is Kronecker's delta function and the second is Dirac delta function.

The steady state mean value is the same as the mean value calculated above using Master equation approach (Equation C.7).

Substitute variables in Langevin equations to the following:

$$\begin{cases} u(t) = \langle u \rangle + \delta_u(t) \\ b(t) = \langle b \rangle + \delta_b(t) \\ m_1(t) = \langle m_1 \rangle + \delta_{m_1}(t) \\ m_2(t) = \langle m_2 \rangle + \delta_{m_2}(t). \end{cases} \tag{C.17}$$

Linearizing the Langevin equations to obtain:

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \delta_u \\ \delta_b \\ \delta_{m_1} \\ \delta_{m_2} \end{pmatrix} = \mathrm{A} \begin{pmatrix} \delta_u \\ \delta_b \\ \delta_{m_1} \\ \delta_{m_2} \end{pmatrix} + \Psi_1 \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \\ \Gamma_3 \\ \Gamma_4 \\ \Gamma_5 \\ \Gamma_6 \\ \Gamma_7 \end{pmatrix} \tag{C.18}$$

where

$$\mathbf{A} = \begin{pmatrix} -(k^+c + \alpha_1) & k^- & 0 & 0 \\ k^+c & -(k^- + \alpha_2) & 0 & 0 \\ \alpha_1 & 0 & -\gamma_1 & 0 \\ 0 & \alpha_2 & 0 & -\gamma_2, \end{pmatrix} \tag{C.19}$$

and

$$\Psi_1 = \begin{pmatrix} \sqrt{v} & -\sqrt{k^+c\langle u \rangle} & \sqrt{k^-\langle b \rangle} & -\sqrt{\alpha_1\langle u \rangle} & 0 & 0 & 0 \\ 0 & \sqrt{k^+c\langle u \rangle} & -\sqrt{k^-\langle b \rangle} & 0 & -\sqrt{\alpha_2\langle b \rangle} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\alpha_1\langle u \rangle} & 0 & -\sqrt{\gamma_1\langle m_1 \rangle} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\alpha_2\langle b \rangle} & 0 & -\sqrt{\gamma_2\langle m_2 \rangle} \end{pmatrix}. \tag{C.20}$$

We now transform these linearized equations into Fourier space, with $\hat{\delta}_i(\omega)$ corresponding to the temporal variables $\delta_i(t)$, where i equals to u, b, m_1 or m_2 in the spectral domain. and $\hat{\Gamma}_j(\omega)$ corresponding to the temporal variable $\Gamma_j(t)$ where

$j \in (1, 2, \ldots 7)$ in the spectral domain.

$$\delta_i(t) = \int \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{i\omega t} \hat{\delta}_i(\omega); \tag{C.21}$$

$$\Gamma_j(t) = \int \frac{\mathrm{d}\omega}{2\pi} \mathrm{e}^{i\omega t} \hat{\Gamma}_j(\omega). \tag{C.22}$$

Thus,

$$\begin{pmatrix} \hat{\delta}_u(\omega) \\ \hat{\delta}_b(\omega) \\ \hat{\delta}_{m_1}(\omega) \\ \hat{\delta}_{m_2}(\omega) \end{pmatrix} = \mathbf{M}^{-1} \mathbf{\Psi}_1 \begin{pmatrix} \hat{\Gamma}_1(\omega) \\ \hat{\Gamma}_2(\omega) \\ \hat{\Gamma}_3(\omega) \\ \hat{\Gamma}_4(\omega) \\ \hat{\Gamma}_5(\omega) \\ \hat{\Gamma}_6(\omega) \\ \hat{\Gamma}_7(\omega) \end{pmatrix} \tag{C.23}$$

where Ψ_1 is defined in Equation C.20 and

$$\mathbf{M} = \begin{pmatrix} k^+c + \alpha_1 + i\omega & -k^- & 0 & 0 \\ -k^+c & k^- + \alpha_2 + i\omega & 0 & 0 \\ -\alpha_1 & 0 & \gamma_1 + i\omega & 0 \\ 0 & -\alpha_2 & 0 & \gamma_2 + i\omega \end{pmatrix}. \tag{C.24}$$

Using the Wiener-Khinchin theorem $\left|\hat{\Gamma}_i(\omega)\right|^2 = \int_{-\infty}^{\infty} \langle \Gamma_i^*(t) \Gamma_i(t+\tau) \rangle \, e^{-i\omega\tau} d\tau$, we can get:

$$\left|\hat{\Gamma}_i(\omega)\right|^2 = 1 \qquad i = 1, 2, \ldots 7 \tag{C.25}$$

Based on these expressions, we can obtain $\left|\hat{\delta}_u(\omega)\right|^2$, $\left|\hat{\delta}_b(\omega)\right|^2$, $\left|\hat{\delta}_{m_1}(\omega)\right|^2$ and $\left|\hat{\delta}_{m_2}(\omega)\right|^2$.

Next, using the relation $\left\langle f\,(t)^2 \right\rangle = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left|\hat{f}(\omega)\right|^2 d\omega$ to inverse transform back to

the time domain to obtain $\langle \delta u^2 \rangle$, $\langle \delta b^2 \rangle$, $\langle \delta m_1^2 \rangle$ and $\langle \delta m_2^2 \rangle$.

$$\begin{cases} \sigma_u^2 = \langle \delta u^2 \rangle = \dfrac{v(\alpha_2 + k^-)}{\alpha_1\alpha_2 + \alpha_1 k^- + \alpha_2 k^+ c} \\ \sigma_b^2 = \langle \delta b^2 \rangle = \dfrac{v k^+ c}{\alpha_1\alpha_2 + \alpha_1 k^- + \alpha_2 k^+ c} \\ \sigma_{m_1}^2 = \langle \delta m_1^2 \rangle = \dfrac{\alpha_1}{\gamma_1} \dfrac{v(\alpha_2 + k^-)}{\alpha_1\alpha_2 + \alpha_1 k^- + \alpha_2 k^+ c} \\ \sigma_{m_2}^2 = \langle \delta m_2^2 \rangle = \dfrac{\alpha_2}{\gamma_2} \dfrac{v k^+ c}{\alpha_1\alpha_2 + \alpha_1 k^- + \alpha_2 k^+ c}. \end{cases} \tag{C.26}$$

These are exactly the same as obtained in Equations C.9 while we use the Master equation approach.

C.2.2 Considering 3D Diffusion of Splicing Factors

For the alternative splicing model described in Figure 4.2, there are 4 molecular species interact inside some fixed volume Ω at constant temperature. If including the diffusion of the splicing factors, we can get the following equations:

$$\begin{cases} \dfrac{\partial c}{\partial t} = D\nabla^2 c(\vec{x}, t) - \dot{b}\delta(\vec{x} - \vec{x_0}) + \mathcal{G} - \mathcal{D} \\ \dfrac{\mathrm{d}u}{\mathrm{d}t} = v - (k^+ c + \alpha_1)u + k^- b + (\sqrt{v}\Gamma_1 - \sqrt{k^+ c u}\Gamma_2 + \sqrt{k^- b}\Gamma_3 - \sqrt{\alpha_1 u}\Gamma_4) \\ \dfrac{\mathrm{d}b}{\mathrm{d}t} = k^+ c u - (k^- + \alpha_2)b + (\sqrt{k^+ c u}\Gamma_2 - \sqrt{k^- b}\Gamma_3 - \sqrt{\alpha_2 b}\Gamma_5) \\ \dfrac{\mathrm{d}m_1}{\mathrm{d}t} = \alpha_1 u - \gamma_1 m_1 + (\sqrt{\alpha_1 u}\Gamma_4 - \sqrt{\gamma_1 m_1}\Gamma_6) \\ \dfrac{\mathrm{d}m_2}{\mathrm{d}t} = \alpha_2 b - \gamma_2 m_2 + (\sqrt{\alpha_2 b}\Gamma_5 - \sqrt{\gamma_2 m_2}\Gamma_7). \end{cases} \tag{C.27}$$

$\delta(\vec{x} - \vec{x_0})$ denotes the Dirac-delta function. These splicing factors can be produced at sources $\mathcal{G}$ and degradate at sink $\mathcal{D}$. For simplicity we will assume that splicing

factors are present at a fixed total number in the cell and they do not decay, i.e., $\mathcal{G} = \mathcal{D} = 0$. c represents splicing factors that can freely diffuse inside the volume. If a splicing factor binds to pre-mRNA splicing site, c will be reduce by 1.

The Langevin noise terms $\vec{\Gamma}$ are characterized within the linear noise approximation by the following time correlation functions::

$$\langle \Gamma_i(t)\Gamma_j(t')\rangle = \delta_{ij}\delta(t-t'), \tag{C.28}$$

where on the right of the equation, the first is Kronecker's delta function and the second is Dirac delta function.

We assume the mean concentration of splicing factor is $\langle c\rangle$. Substitute variables in Equations C.27 to the following form:

$$\begin{cases} c(\vec{x},t) = \langle c\rangle + \delta_c(\vec{x},t) \\ u = \langle u\rangle + \delta_u(t) \\ b = \langle b\rangle + \delta_b(t) \\ m_1 = \langle m_1\rangle + \delta_{m_1}(t) \\ m_2 = \langle m_2\rangle + \delta_{m_2}(t). \end{cases} \tag{C.29}$$

Linearize Equations C.27 and transform into Fourier space, then use the Fourier representative of splicing factor concentration as shown in Equation 4.16, we will get the following equation matrix:

$$i\omega \begin{pmatrix} \hat{\delta}_u(\omega) \\ \hat{\delta}_b(\omega) \\ \hat{\delta}_{m_1}(\omega) \\ \hat{\delta}_{m_2}(\omega) \end{pmatrix} = \mathbf{Ad} \begin{pmatrix} \hat{\delta}_u(\omega) \\ \hat{\delta}_b(\omega) \\ \hat{\delta}_{m_1}(\omega) \\ \hat{\delta}_{m_2}(\omega) \end{pmatrix} + \mathbf{\Psi}_2 \begin{pmatrix} \hat{\Gamma}_1(\omega) \\ \hat{\Gamma}_2(\omega) \\ \hat{\Gamma}_3(\omega) \\ \hat{\Gamma}_4(\omega) \\ \hat{\Gamma}_5(\omega) \\ \hat{\Gamma}_6(\omega) \\ \hat{\Gamma}_7(\omega) \end{pmatrix}. \tag{C.30}$$

where

$$\mathbf{A}_d = \begin{pmatrix} -(k^+\bar{c}+\alpha_1) & k^- + i\omega\theta & 0 & 0 \\ k^+\bar{c} & -(k^-+\alpha_2) - i\omega\theta & 0 & 0 \\ \alpha_1 & 0 & -\gamma_1 & 0 \\ 0 & \alpha_2 & 0 & -\gamma_2 \end{pmatrix}, \tag{C.31}$$

$$\Psi_2 = \begin{pmatrix} \sqrt{v} & -\sqrt{k^+\langle c\rangle\langle u\rangle} & \sqrt{k^-\langle b\rangle} & -\sqrt{\alpha_1\langle u\rangle} & 0 & 0 & 0 \\ 0 & \sqrt{k^+\langle c\rangle\langle u\rangle} & -\sqrt{k^-\langle b\rangle} & 0 & -\sqrt{\alpha_2\langle b\rangle} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\alpha_1\langle u\rangle} & 0 & -\sqrt{\gamma_1\langle m_1\rangle} & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\alpha_2\langle b\rangle} & 0 & -\sqrt{\gamma_2\langle m_2\rangle} \end{pmatrix}, \tag{C.32}$$

and

$$\theta = \frac{k^+\langle u\rangle}{2\pi Da} = \frac{k^+ v(\alpha_2 + k^-)}{2\pi \mathrm{Da}(\alpha_1\alpha_2 + \alpha_1 k^- + \alpha_2 k^+ c)}. \tag{C.33}$$

Thus we can get the similar equation as in Equation C.23.

$$\begin{pmatrix} \hat{\delta}_u(\omega) \\ \hat{\delta}_b(\omega) \\ \hat{\delta}_{m_1}(\omega) \\ \hat{\delta}_{m_2}(\omega) \end{pmatrix} = \tilde{\mathbf{M}}^{-1}\mathbf{\Psi}_2 \begin{pmatrix} \hat{\Gamma}_1(\omega) \\ \hat{\Gamma}_2(\omega) \\ \hat{\Gamma}_3(\omega) \\ \hat{\Gamma}_4(\omega) \\ \hat{\Gamma}_5(\omega) \\ \hat{\Gamma}_6(\omega) \\ \hat{\Gamma}_7(\omega) \end{pmatrix}. \tag{C.34}$$

where

$$\tilde{\mathbf{M}} = \begin{pmatrix} k^+\bar{c} + \alpha_1 + i\omega & -k^- - i\omega\theta & 0 & 0 \\ -k^+\bar{c} & k^- + \alpha_2 + i\omega(1+\theta) & 0 & 0 \\ -\alpha_1 & 0 & \gamma_1 + i\omega & 0 \\ 0 & -\alpha_2 & 0 & \gamma_2 + i\omega \end{pmatrix} \tag{C.35}$$

The Fourier space analysis is the same as in the previous section. using the Fourier space analysis, we can get the variance of all four variables u,b,m_1 and m_2. However, the expression of the last two variables are too lengthy, we only list the

variance of u and b:

$$\begin{cases} \sigma_u^2 = \dfrac{v}{\alpha_1 + g\alpha_2} \dfrac{k^+c + (\alpha_2 + k^- + \alpha_2 g\theta)(1+\theta) + \alpha_1(1+\theta)^2}{(1+\theta)(\alpha_1 + \alpha_2 + k^- + k^+c + \alpha_1\theta)}, \\ \sigma_b^2 = \dfrac{vg}{\alpha_1 + g\alpha_2} \dfrac{\alpha_2 + k^- + (\alpha_1 + k^+c)(1+\theta)}{(1+\theta)(\alpha_1 + \alpha_2 + k^- + k^+c + \alpha_1\theta)}. \end{cases} \tag{C.36}$$

C.3 Random-Walk Model of Diffusion in Three Dimensions

In order to prove our theoretical calculation, we also build a simulation model. In our model, we simplify the nucleus as a cube and divide it into $100 \times 100 \times 100$ small cubes. Splicing factors were initially assigned randomly on lattice. Then for every time step Δt, splicing factors can move between lattice adjacent points, i.e., each splicing factor can move to one of the six adjacent points with equal probability per step. Because we consider only the isotropic diffusion: $D_x = D_y = D_z = D_0$, we can set the distance between lattice points in x,y,z direction to be the same:

$$\Delta l = \sqrt{6 D_0 \Delta t}. \tag{C.37}$$

The diameter of a typical mammalian nucleus is $5-10$ microns (10^4 nanometers) [134]. If treat nucleus as a cube and divide into $100 \times 100 \times 100$ small cubes. Then the length of each of the small cubes is roughly $\Delta l = 0.1\mu m$. The diffusion coefficient of splicing factors in the nucleus is about $0.50 - 20.00 um^2/sec$ in the nucleoplasm, transcript in the nucleus is $0.03-0.1\mu m^2/sec$ (Table C.1). Because splicing factor can bind to pre-mRNA as soon as the binding site exposes, pre-mRNA may still hang on the DNA (alternative splicing occurs co-transcriptionally), which makes pre-mRNA more immobile. So we can assume pre-mRNA are at the center of the cube and while splicing factor diffuse to the center, it has a probability to bind to the binding site. With splicing factor binding or unbinding to the binding site, different mRNA isoforms are produced. Due to Equation C.37, the time step Δt between jumps is

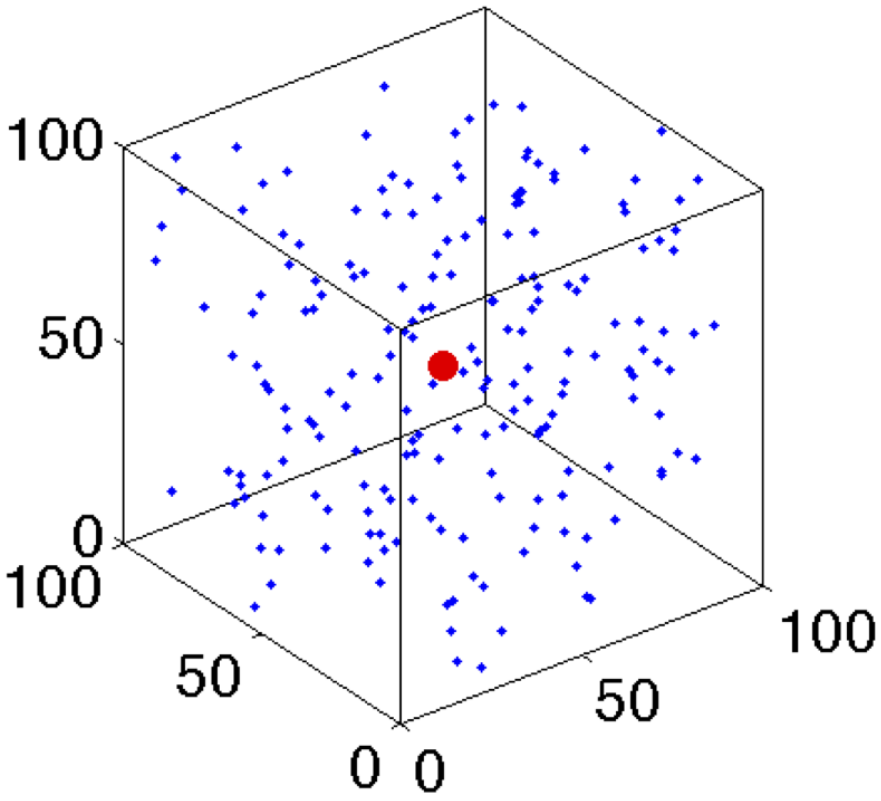


FIGURE C.3: Simulation model of splicing factor diffusion and splicing site binding

roughly $0.0003 - 0.003$ sec. The concentration of splicing factor is roughly $10nM$ to $1\mu m$, so in our model, we assume there are around 1000 splicing factor particles diffuse in nucleus.

If a particle currently at location $\vec{R}(t)$ was directed to move outside of the lattice with one step displacement $\Delta\vec{R}$, it was instantaneous returned back using specular reflection conditions.

$$\vec{R}(t + \Delta t) = \vec{R} - \Delta\vec{R}. \tag{C.38}$$

By using this model, we can numerically calculate the mean and variance as well as the covariance of u and b (Figure 4.4(b),(d)).

Table C.1: Key Numbers in Biology

Cell Size [135]	Bacteria (E. coli)	$0.7 - 1.4\mu m$ diameter, $0.5 - 5\mu m^2$ in volume
	Yeast (S. cerevisiae)	$3 - 6\mu m$ diameter, $20 - 160\mu m^2$ in volume
	Mammalian cell	$100 - 10000\mu m^2$ in volume; (HeLa cell: $500 - 5000\mu m^2$)
Concentration [135]	signaling protein	$\approx 10nM - 1\mu m$
	Concentration of $1nM$	in E. coli ≈ 1 molecule/cell; in HeLa cell ≈ 1000 molecules/cell
Diffusion Coefficient	protein in nucleus	$D = 0.5 - 20.0\mu m^2/s$ [136]
	polyA mRNA in nucleus	$0.03 - 0.1\mu m^2/s$ [137]

Bibliography

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

[2] U. S. Bhalla and R. Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, Jan 1999.

[3] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–5, 2002.

[4] Michael J Holland. Transcript abundance in yeast varies over six orders of magnitude. *J Biol Chem*, 277(17):14363–14366, Apr 2002.

[5] Mukund Thattai and Alexander van Oudenaarden. Stochastic gene expression in fluctuating environments. *Genetics*, 167(1):523–530, May 2004.

[6] D. R. Rigney and W. C. Schieve. Stochastic model of linear, continuous protein synthesis in bacterial populations. *J Theor Biol*, 69(4):761–6, 1977.

[7] E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nat Genet*, 31(1):69–73, 2002.

[8] 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, 2002.

[9] Arjun Raj and Alexander van Oudenaarden. Single-molecule approaches to stochastic gene expression. *Annu Rev Biophys*, 38:255–270, 2009.

[10] Narendra Maheshri and Erin K O'Shea. Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu Rev Biophys Biomol Struct*, 36:413–434, 2007.

[11] Michal Rabani, Joshua Z. Levin, Lin Fan, Xian Adiconis, Raktima Raychowdhury, Manuel Garber, Andreas Gnirke, Chad Nusbaum, Nir Hacohen, Nir Friedman, Ido Amit, and Aviv Regev. Metabolic labeling of rna uncovers principles of rna production and degradation dynamics in mammalian cells. *Nat Biotechnol*, 29(5):436–442, May 2011.

[12] Daniel R Larson, Daniel Zenklusen, Bin Wu, Jeffrey A Chao, and Robert H Singer. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science*, 332(6028):475–478, Apr 2011.

[13] Claire V. Harper, Brbel Finkenstdt, Dan J. Woodcock, Snke Friedrichsen, Sabrina Semprini, Louise Ashall, David G. Spiller, John J. Mullins, David A. Rand, Julian R E. Davis, and Michael R H. White. Dynamic analysis of stochastic transcription cycles. *PLoS Biol*, 9(4):e1000607, Apr 2011.

[14] Lok-Hang So, Anandamohan Ghosh, Chenghang Zong, Leonardo A. Seplveda, Ronen Segev, and Ido Golding. General properties of transcriptional time series in escherichia coli. *Nat Genet*, 43(6):554–560, Jun 2011.

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

[16] J. M. Raser and E. K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–1814, 2004.

[17] Nitzan Rosenfeld, Jonathan W Young, Uri Alon, Peter S Swain, and Michael B Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–1965, Mar 2005.

[18] Ido Golding, Johan Paulsson, Scott M Zawilski, and Edward C Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, Dec 2005.

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

[20] J. R. Chubb, T. Trcek, S. M. Shenoy, and R. H. Singer. Transcriptional pulsing of a developmental gene. *Current Biology*, 16(10):1018–1025, 2006.

[21] Thomas Gregor, David W Tank, Eric F Wieschaus, and William Bialek. Probing the limits to positional information. *Cell*, 130(1):153–164, Jul 2007.

[22] G. Tkacik, T. Gregor, and W. Bialek. The role of input noise in transcriptional regulation. *PLoS ONE*, 3(7):e2774, 2008.

[23] Etay Ziv, Ilya Nemenman, and Chris H Wiggins. Optimal signal processing in small stochastic biochemical networks. *PLoS One*, 2(10):e1077, 2007.

[24] Attila Becskei, Benjamin B Kaufmann, and Alexander van Oudenaarden. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat Genet*, 37(9):937–944, Sep 2005.

[25] 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, 2006.

[26] David P Bartel. Microrna: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, Jan 2004.

[27] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, Jan 2005.

[28] John Tsang, Jun Zhu, and Alexander van Oudenaarden. Microrna-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol Cell*, 26(5):753–767, Jun 2007.

[29] Jongpil Kim, Keiichi Inoue, Jennifer Ishii, William B Vanti, Sergey V Voronov, Elizabeth Murchison, Gregory Hannon, and Asa Abeliovich. A microrna feedback circuit in midbrain dopamine neurons. *Science*, 317(5842):1220–1224, Aug 2007.

[30] Stephen M Cohen, Julius Brennecke, and Alexander Stark. Denoising feedback loops by thresholding–a new role for microrna. *Genes Dev*, 20(20):2769–2772, Oct 2006.

[31] Chunnian Zhao, GuoQiang Sun, Shengxiu Li, and Yanhong Shi. A feedback regulatory loop involving microrna-9 and nuclear receptor tlx in neural stem cell fate determination. *Nat Struct Mol Biol*, 16(4):365–371, Apr 2009.

[32] James C Carrington and Victor Ambros. Role of microrna in plant and animal development. *Science*, 301(5631):336–338, Jul 2003.

[33] Julius Brennecke, David R Hipfner, Alexander Stark, Robert B Russell, and Stephen M Cohen. bantam encodes a developmentally regulated microrna that controls cell proliferation and regulates the proapoptotic gene hid in drosophila. *Cell*, 113(1):25–36, Apr 2003.

[34] Peizhang Xu, Stephanie Y Vernooy, Ming Guo, and Bruce A Hay. The drosophila microrna mir-14 suppresses cell death and is required for normal fat metabolism. *Curr Biol*, 13(9):790–795, Apr 2003.

[35] Robert J Johnston, Sarah Chang, John F Etchberger, Christopher O Ortiz, and Oliver Hobert. Microrna acting in a double-negative feedback loop to control a neuronal cell fate decision. *Proc Natl Acad Sci U S A*, 102(35):12449–12454, Aug 2005.

[36] Yan Li, Fay Wang, Jin-A. Lee, and Fen-Biao Gao. Microrna-9a ensures the precise specification of sensory organ precursors in drosophila. *Genes Dev*, 20(20):2793–2805, Oct 2006.

[37] Fen-Biao Gao. Posttranscriptional control of neuronal development by microrna networks. *Trends Neurosci*, 31(1):20–26, Jan 2008.

[38] Jun Lu, Shangqin Guo, Benjamin L Ebert, Hao Zhang, Xiao Peng, Jocelyn Bosco, Jennifer Pretz, Rita Schlanger, Judy Y Wang, Raymond H Mak, David M Dombkowski, Frederic I Preffer, David T Scadden, and Todd R Golub. Microrna-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. *Dev Cell*, 14(6):843–853, Jun 2008.

[39] Ji Li and Iva Greenwald. Lin-14 inhibition of lin-12 contributes to precision and timing of c. elegans vulval fate patterning. *Curr Biol*, 20(20):1875–1879, Oct 2010.

[40] Xin Li, Justin J Cassidy, Catherine A Reinke, Stephen Fischboeck, and Richard W Carthew. A microrna imparts robustness against environmental fluctuation during development. *Cell*, 137(2):273–282, Apr 2009.

[41] Hctor Herranz and Stephen M Cohen. Micrornas and gene regulatory networks: managing the impact of noise in biological systems. *Genes Dev*, 24(13):1339–1344, Jul 2010.

[42] Deepak Srivastava. Making or breaking the heart: from lineage determination to morphogenesis. *Cell*, 126(6):1037–1048, Sep 2006.

[43] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415, Dec 2008.

[44] Luca Cartegni, Shern L. Chew, and Adrian R. Krainer. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, 3(4):285–298, Apr 2002.

[45] Javier F. Cceres and Alberto R. Kornblihtt. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet*, 18(4):186–193, Apr 2002.

[46] Julian P. Venables. Unbalanced alternative splicing and its significance in cancer. *Bioessays*, 28(4):378–386, Apr 2006.

[47] Prabhakar Rajan, David J. Elliott, Craig N. Robson, and Hing Y. Leung. Alternative splicing and biological heterogeneity in prostate cancer. *Nat Rev Urol*, 6(8):454–460, Aug 2009.

[48] J. Paulsson. Summing up the noise in gene networks. *nature*, 427(6973):415–418, 2004.

[49] J. M. Raser and E. K. O'Shea. Noise in gene expression: Origins, consequences, and control. *Science*, 309(5743):2010–2013, 2005.

[50] Gasper Tkacik, Thomas Gregor, and William Bialek. The role of input noise in transcriptional regulation. *PLoS One*, 3(7):e2774, 2008.

[51] Gasper Tkacik and William Bialek. Diffusion, dimensionality, and noise in transcriptional regulation. *Phys Rev E Stat Nonlin Soft Matter Phys*, 79(5 Pt 1):051901, May 2009.

[52] Shangying Wang and Sridhar Raghavachari. Quantifying negative feedback regulation by micro-rnas. *Phys Biol*, 8(5):055002, Oct 2011.

[53] Fernando Carrillo Oesterreich, Nicole Bieberstein, and Karla M. Neugebauer. Pause locally, splice globally. *Trends Cell Biol*, 21(6):328–335, Jun 2011.

[54] Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010.

[55] Zeev Waks, Allon M. Klein, and Pamela A. Silver. Cell-to-cell variability of alternative rna splicing. *Mol Syst Biol*, 7:506, 2011.

[56] M. A. Lynch. Long-term potentiation and memory. *Physiol Rev*, 84(1):87–136, 2004.

[57] M. E. Greenberg, E. B. Ziff, and L. A. Greene. Stimulation of neuronal acetylcholine-receptors induces rapid gene-transcription. *Science*, 234(4772):80–83, 1986.

[58] M. Sheng, M. A. Thompson, and M. E. Greenberg. Creb: a ca(2+)-regulated transcription factor phosphorylated by calmodulin-dependent kinases. *Science*, 252(5011):1427–30, 1991.

[59] D. Gau, T. Lemberger, C. von Gall, O. Kretz, N. Le Minh, P. Gass, W. Schmid, U. Schibler, H. W. Korf, and G. Schutz. Phosphorylation of creb ser142 regulates light-induced phase shifts of the circadian clock. *Neuron*, 34(2):245–53, 2002.

[60] G. Y. Wu, K. Deisseroth, and R. W. Tsien. Activity-dependent creb phosphorylation: convergence of a fast, sensitive calmodulin kinase pathway and a slow, less sensitive mitogen-activated protein kinase pathway. *Proc Natl Acad Sci U S A*, 98(5):2808–13, 2001.

[61] J. Peccoud and B. Ycart. Markovian modelling of gene products synthesis. *Theoretical Population Biology*, 48:222–234, 1995.

[62] Yuichi Taniguchi, Paul J. Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, Jul 2010.

[63] Edward Yang, Erik van Nimwegen, Mihaela Zavolan, Nikolaus Rajewsky, Mark Schroeder, Marcelo Magnasco, and James E Darnell, Jr. Decay rates of human mrnas: correlation with functional characteristics and sequence attributes. *Genome Res*, 13(8):1863–1872, Aug 2003.

[64] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? *Nat Rev Genet*, 9(2):102–114, Feb 2008.

[65] Ramesh S Pillai. Microrna function: multiple mechanisms for a tiny rna? *RNA*, 11(12):1753–1761, Dec 2005.

[66] Gyrgy Hutvgner and Phillip D Zamore. A microrna in a multiple-turnover rnai enzyme complex. *Science*, 297(5589):2056–2060, Sep 2002.

[67] Yukihide Tomari and Phillip D Zamore. Perspective: machines for rnai. *Genes Dev*, 19(5):517–529, Mar 2005.

[68] S. M. Janicki, T. Tsukamoto, S. E. Salghetti, W. P. Tansey, R. Sachidanandam, K. V. Prasanth, T. Ried, Y. Shav-Tal, E. Bertrand, R. H. Singer, and D. L. Spector. From silencing to gene expression: Real-time analysis in single cells. *Cell*, 116(5):683–698, 2004.

[69] Eran Eden, Naama Geva-Zatorsky, Irina Issaeva, Ariel Cohen, Erez Dekel, Tamar Danon, Lydia Cohen, Avi Mayo, and Uri Alon. Proteome half-life dynamics in living human cells. *Science*, 331(6018):764–768, Feb 2011.

[70] Jrg Grigull, Sanie Mnaimneh, Jeffrey Pootoolal, Mark D Robinson, and Timothy R Hughes. Genome-wide analysis of mrna stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol Cell Biol*, 24(12):5534–5547, Jun 2004.

[71] Marlena Siwiak and Piotr Zielenkiewicz. A comprehensive, quantitative, and genome-wide model of translation. *PLoS Comput Biol*, 6(7):e1000865, 2010.

[72] Christian Miller, Bjrn Schwalb, Kerstin Maier, Daniel Schulz, Sebastian Dmcke, Benedikt Zacher, Andreas Mayer, Jasmin Sydow, Lisa Marcinowski, Lars

Dlken, Dietmar E Martin, Achim Tresch, and Patrick Cramer. Dynamic transcriptome analysis measures rates of mrna synthesis and decay in yeast. *Mol Syst Biol*, 7:458, Jan 2011.

[73] E. Levine, Z. Zhang, T. Kuhlman, and T. Hwa. Quantitative characteristics of gene regulation by small rna. *Plos Biology*, 5(9):1998–2010, 2007.

[74] A. Becskei and L. Serrano. Engineering stability in gene networks by autoregulation. *nature*, 405(6786):590–593, 2000.

[75] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci U S A*, 98(15):8614–8619, Jul 2001.

[76] Gil Hornung and Naama Barkai. Noise propagation and signaling sensitivity in biological networks: a role for positive feedback. *PLoS Comput Biol*, 4(1):e8, Jan 2008.

[77] Yann Dublanche, Konstantinos Michalodimitrakis, Nico Kmmerer, Mathilde Foglierini, and Luis Serrano. Noise in transcription negative feedback loops: simulation and experimental analysis. *Mol Syst Biol*, 2:41, 2006.

[78] Tatiana T Marquez-Lago and Jrg Stelling. Counter-intuitive stochastic behavior of simple gene circuits with negative feedback. *Biophys J*, 98(9):1742–1750, May 2010.

[79] U. Fano. Note on the theory of radiation-induced lethals in drosophila. *Science*, 106(2743):87–88, Jul 1947.

[80] Alexander Stark, Julius Brennecke, Natascha Bushati, Robert B Russell, and Stephen M Cohen. Animal microrna confer robustness to gene expression and have a significant impact on 3'utr evolution. *Cell*, 123(6):1133–1146, Dec 2005.

[81] J. Paulsson and M. Ehrenberg. Noise in a minimal regulatory network: plasmid copy number control. *Q Rev Biophys*, 34(1):1–59, Feb 2001.

[82] J. Elf and M. Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Research*, 13(11):2475–2484, 2003.

[83] Johan Elf, Johan Paulsson, Otto G Berg, and Mns Ehrenberg. Near-critical phenomena in intracellular metabolite pools. *Biophys J*, 84(1):154–170, Jan 2003.

[84] Pankaj Mehta, Sidhartha Goyal, and Ned S Wingreen. A quantitative comparison of srna-based and protein-based gene regulation. *Mol Syst Biol*, 4:221, 2008.

[85] N. Goldenfeld. *Lectures on phase transitions and the renormalization group*. Frontiers in physics. Addison-Wesley, Advanced Book Program, 1992.

[86] Derrick H Lenz, Kenny C Mok, Brendan N Lilley, Rahul V Kulkarni, Ned S Wingreen, and Bonnie L Bassler. The small rna chaperone hfq and multiple small rnas control quorum sensing in vibrio harveyi and vibrio cholerae. *Cell*, 118(1):69–82, Jul 2004.

[87] Namiko Mitarai, Anna M C Andersson, Sandeep Krishna, Szabolcs Semsey, and Kim Sneppen. Efficient degradation and expression prioritization with small rnas. *Phys Biol*, 4(3):164–171, Sep 2007.

[88] Yishai Shimoni, Gilgi Friedlander, Guy Hetzroni, Gali Niv, Shoshy Altuvia, Ofer Biham, and Hanah Margalit. Regulation of gene expression by small non-coding rnas: a quantitative view. *Mol Syst Biol*, 3:138, 2007.

[89] N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. North-holland publishing company, 1981.

[90] Daehyun Baek, Judit Villn, Chanseok Shin, Fernando D Camargo, Steven P Gygi, and David P Bartel. The impact of micrornas on protein output. *Nature*, 455(7209):64–71, Sep 2008.

[91] Huili Guo, Nicholas T Ingolia, Jonathan S Weissman, and David P Bartel. Mammalian micrornas predominantly act to decrease target mrna levels. *Nature*, 466(7308):835–840, Aug 2010.

[92] T. B. Kepler and T. C. Elston. Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophysical Journal*, 81(6):3116–3136, 2001.

[93] Sabine Rdel, Yanli Wang, Ren Lenobel, Roman Krner, He-Hsuan Hsiao, Henning Urlaub, Dinshaw Patel, and Gunter Meister. Phosphorylation of human argonaute proteins affects small rna binding. *Nucleic Acids Res*, Nov 2010.

[94] Martijn Kedde, Marieke van Kouwenhove, Wilbert Zwart, Joachim A F Oude Vrielink, Ran Elkon, and Reuven Agami. A pumilio-induced RNA structure switch in p27-3' UTR controls mir-221 and mir-222 accessibility. *Nat Cell Biol*, 12(10):1014–1020, Oct 2010.

[95] R. Losick and C. Desplan. Stochasticity and cell fate. *Science*, 320(5872):65–68, 2008.

[96] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov 2008.

[97] Arianne J. Matlin, Francis Clark, and Christopher W J. Smith. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*, 6(5):386–398, May 2005.

[98] Hadas Keren, Galit Lev-Maor, and Gil Ast. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*, 11(5):345–355, May 2010.

[99] Alexander V. Alekseyenko, Namshin Kim, and Christopher J. Lee. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. *RNA*, 13(5):661–670, May 2007.

[100] L. R. Bell, E. M. Maine, P. Schedl, and T. W. Cline. Sex-lethal, a drosophila sex determination switch gene, exhibits sex-specific rna splicing and sequence similarity to rna binding proteins. *Cell*, 55(6):1037–1046, Dec 1988.

[101] Victor Marinescu, Patricia A. Loomis, Svetlana Ehmann, Mitchell Beales, and Judith A. Potashkin. Regulation of retention of fosb intron 4 by ptb. *PLoS One*, 2(9):e828, 2007.

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

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

[104] Alberto R. Kornblihtt, Ignacio E. Schor, Mariano All, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J. Muoz. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat Rev Mol Cell Biol*, 14(3):153–165, Mar 2013.

[105] Rajamanickam Murugan and Gabriel Kreiman. Theory on the coupled stochastic dynamics of transcription and splice-site recognition. *PLoS Comput Biol*, 8(11):e1002747, 2012.

[106] Imke Listerman, Aparna K. Sapra, and Karla M. Neugebauer. Cotranscriptional coupling of splicing factor recruitment and precursor messenger rna splicing in mammalian cells. *Nat Struct Mol Biol*, 13(9):815–822, Sep 2006.

[107] Amy Pandya-Jones and Douglas L. Black. Co-transcriptional splicing of constitutive and alternative exons. *RNA*, 15(10):1896–1908, Oct 2009.

[108] Jarnail Singh and Richard A. Padgett. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol*, 16(11):1128–1133, Nov 2009.

[109] Rita Das, Jiong Yu, Zuo Zhang, Melanie P. Gygi, Adrian R. Krainer, Steven P. Gygi, and Robin Reed. Sr proteins function in coupling rnap ii transcription to pre-mrna splicing. *Mol Cell*, 26(6):867–881, Jun 2007.

[110] Manuel de la Mata, Claudio R. Alonso, Sebastin Kadener, Juan P. Fededa, Matas Blaustein, Federico Pelisch, Paula Cramer, David Bentley, and Alberto R. Kornblihtt. A slow rna polymerase ii affects alternative splicing in vivo. *Mol Cell*, 12(2):525–532, Aug 2003.

[111] Alberto R. Kornblihtt. Chromatin, transcript elongation and alternative splicing. *Nat Struct Mol Biol*, 13(1):5–7, Jan 2006.

[112] Joanna Y. Ip, Dominic Schmidt, Qun Pan, Arun K. Ramani, Andrew G. Fraser, Duncan T. Odom, and Benjamin J. Blencowe. Global impact of rna polymerase ii elongation inhibition on alternative splicing regulation. *Genome Res*, 21(3):390–401, Mar 2011.

[113] Manuel J. Muoz, M Soledad Prez Santangelo, Maria P. Paronetto, Manuel de la Mata, Federico Pelisch, Stphanie Boireau, Kira Glover-Cutter, Claudia Ben-Dov, Matas Blaustein, Juan J. Lozano, Gregory Bird, David Bentley, Edouard Bertrand, and Alberto R. Kornblihtt. Dna damage regulates alternative splicing through inhibition of rna polymerase ii elongation. *Cell*, 137(4):708–720, May 2009.

[114] Yehuda Brody, Noa Neufeld, Nicole Bieberstein, Sebastien Z. Causse, Eva-Maria Bhnlein, Karla M. Neugebauer, Xavier Darzacq, and Yaron Shav-Tal. The in vivo kinetics of rna polymerase ii elongation during co-transcriptional splicing. *PLoS Biol*, 9(1):e1000573, 2011.

[115] Reini F. Luco, Mariano Allo, Ignacio E. Schor, Alberto R. Kornblihtt, and Tom Misteli. Epigenetics in alternative pre-mrna splicing. *Cell*, 144(1):16–26, Jan 2011.

[116] Alfred J. Robison and Eric J. Nestler. Transcriptional and epigenetic mechanisms of addiction. *Nat Rev Neurosci*, 12(11):623–637, Nov 2011.

[117] J Thomas Cunningham, W David Knight, Steven W. Mifflin, and Eric J. Nestler. An essential role for deltafosb in the median preoptic nucleus in the sustained hypertensive effects of chronic intermittent hypoxia. *Hypertension*, 60(1):179–187, Jul 2012.

[118] Ian Maze, Herbert E Covington, 3rd, David M. Dietz, Quincey LaPlant, William Renthal, Scott J. Russo, Max Mechanic, Ezekiell Mouzon, Rachael L. Neve, Stephen J. Haggarty, Yanhua Ren, Srihari C. Sampath, Yasmin L. Hurd, Paul Greengard, Alexander Tarakhovsky, Anne Schaefer, and Eric J. Nestler.

Essential role of the histone methyltransferase g9a in cocaine-induced plasticity. *Science*, 327(5962):213–216, Jan 2010.

[119] H. Maamar, A. Raj, and D. Dubnau. Noise in gene expression determines cell fate in bacillus subtilis. *Science*, 317(5837):526–9, 2007.

[120] Diana Y. Vargas, Khyati Shah, Mona Batish, Michael Levandoski, Sourav Sinha, Salvatore A E. Marras, Paul Schedl, and Sanjay Tyagi. Single-molecule imaging of transcriptionally coupled and uncoupled splicing. *Cell*, 147(5):1054–1065, Nov 2011.

[121] Dale Muzzey, Carlos A Gmez-Uribe, Jerome T Mettetal, and Alexander van Oudenaarden. A systems-level analysis of perfect adaptation in yeast osmoregulation. *Cell*, 138(1):160–171, Jul 2009.

[122] E. R. Kandel. The molecular biology of memory storage: a dialogue between genes and synapses. *Science*, 294(5544):1030–1038, Nov 2001.

[123] Angel Barco, Craig H Bailey, and Eric R Kandel. Common molecular mechanisms in explicit and implicit memory. *J Neurochem*, 97(6):1520–1533, Jun 2006.

[124] Alcino J Silva, Yu Zhou, Thomas Rogerson, Justin Shobe, and J. Balaji. Molecular and cellular approaches to memory allocation in neural circuits. *Science*, 326(5951):391–395, Oct 2009.

[125] A. E. West, E. C. Griffith, and M. E. Greenberg. Regulation of transcription factors by neuronal activity. *Nature Reviews Neuroscience*, 3(12):921–931, 2002.

[126] Ricardo Dolmetsch. Excitation-transcription coupling: signaling by ion channels to the nucleus. *Sci STKE*, 2003(166):PE4, Jan 2003.

[127] Zilong Qiu and Anirvan Ghosh. A brief history of neuronal gene expression: regulatory mechanisms and cellular consequences. *Neuron*, 60(3):449–455, Nov 2008.

[128] R. E. Dolmetsch, R. S. Lewis, C. C. Goodnow, and J. I. Healy. Differential activation of transcription factors induced by ca2+ response amplitude and duration. *Nature*, 386(6627):855–858, Apr 1997.

[129] R. E. Dolmetsch, K. Xu, and R. S. Lewis. Calcium oscillations increase the efficiency and specificity of gene expression. *Nature*, 392(6679):933–936, Apr 1998.

[130] Filipe Tostevin and Pieter Rein ten Wolde. Mutual information between input and output trajectories of biochemical networks. *Phys Rev Lett*, 102(21):218101, May 2009.

[131] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, second edition, 2006.

[132] Erel Levine, Eshel Ben Jacob, and Herbert Levine. Target-specific and global effectors in gene regulation by microrna. *Biophys J*, 93(11):L52–L54, Dec 2007.

[133] M. Komorowski, J. Miekisz, and A. M. Kierzek. Translational repression contributes greater noise to gene expression than transcriptional repression. *Biophys J*, 96(2):372–84, 2009.

[134] Gene-Wei Li and X Sunney Xie. Central dogma at the single-molecule level in living cells. *Nature*, 475(7356):308–315, Jul 2011.

[135] Uri Moran, Rob Phillips, and Ron Milo. Snapshot: key numbers in biology. *Cell*, 141(7):1262–1262.e1, Jun 2010.

[136] Gordon L Hager, James G McNally, and Tom Misteli. Transcription dynamics. *Mol Cell*, 35(6):741–753, Sep 2009.

[137] Tom Misteli. Physiological importance of rna and protein mobility in the cell nucleus. *Histochem Cell Biol*, 129(1):5–11, Jan 2008.

Biography

Shangying Wang was born in Wuhu, Anhui province, China on December, 02, 1982. She was accepted into Fudan University to study Optical Information Science and Technology in 2002 and got her bachelors degree in Science there on 2006. She was the first student in her class to be admitted to join a research group in the National Key Laboratory of Advanced Photonic Materials and Devices to do research in biophysics. Since then, she became more and more interested in biophysics area. She came to Duke University in 2006 for pursuing her PhD degree. She has been in Duke for eight years and Durham is like a second hometown to her. From 2006-2007, she worked at Free Electron Laser Laboratory by the supervision of Prof. Glenn Edwards. She then continue doing research at Prof. Sridhar Raghavachari's Lab. She is also co-advised by Prof. Nicolas Buchler.