Stochastic Modeling of Gene Expression and Post-transcriptional Regulation

Tao Jia

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Physics

Rahul Kulkarni, Chair
Michel Pleimling
Kyungwha Park
Randy Heflin

July 21, 2011
Blacksburg, Virginia

Keywords: Stochastic gene expression, post-transcriptional regulation, regulatory sRNA, transcriptional bursting, queueing theory

Stochastic Modeling of Gene Expression and Post-transcriptional Regulation

Tao Jia

(ABSTRACT)

Stochasticity is a ubiquitous feature of cellular processes such as gene expression that can give rise to phenotypic differences for genetically identical cells. Understanding how the underlying biochemical reactions give rise to variations in mRNA/protein levels is thus of fundamental importance to diverse cellular processes. Recent technological developments have enabled single-cell measurements of cellular macromolecules which can shed new light on processes underlying gene expression. Correspondingly, there is a need for the development of theoretical tools to quantitatively model stochastic gene expression and its consequences for cellular processes.

In this dissertation, we address this need by developing general stochastic models of gene expression. By mapping the system to models analyzed in queueing theory, we derive analytical expressions for the noise in steady-state protein distributions. Furthermore, given that the underlying processes are intrinsically stochastic, cellular regulation must be designed to control the 'noise' in order to adapt and respond to changing environments. Another

focus of this dissertation is to develop and analyze stochastic models of post-transcription regulation. The analytical solutions of the models proposed provide insight into the effects of different mechanisms of regulation and the role of small RNAs in fine-tunning the noise in gene expression. The results derived can serve as building blocks for future studies focusing on regulation of stochastic gene expression.

Acknowledgments

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Rahul Kulkarni, who has supported me throughout my Ph.D study with his excellent knowledge, caring and patience. He is one of the most hard-working professors I knew but he never ever pushed me. He allowed me the room to work on a different major in Industrial Engineering. He taught me on how to make a successful presentation and how to write a scientific article. This dissertation, as well as all the achievements I obtained would not have been possible without his help and encouragement. My choice of pursuing a post-doctoral position after my Ph.D is mainly inspired by his passion and integrity in science. I am extremely fortunate to have him as my mentor.

I would like to show my gratitude to my committee members: Dr. Michel Pleimling, Dr. Kyungwha Park and Dr. Randy Heflin. I really appreciate the time and effort they put in reading and commenting on my work. My special thanks to Dr. Michel Pleimling, who reads my dissertation in great details and provides me extremely helpful suggestions.

I would like to acknowledge other people I worked with: Vlad Elgart, Andrew Fenley, Charles Baker and Thierry Platini. This dissertation includes results from some projects that we did together.

I would like to thank my wife, Xiaoqiu Gong. She was always there cheering me up and stood by me through the good times and bad.

Finally, I would also like to thank all the friends I have in Blacksburg. I could not list all of their names here but they indeed made my life in Blacksburg more enjoyable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For decades, biology was widely considered to be (relatively) a data-poor science: the enormous complexity of a living system overwhelmed the analytical tools available. Much of the research in the field was done by ignoring complexity and focusing on the component parts. While this remains a powerful strategy, the situation has changed due to advances in technology; biology has now been transformed into a data-rich science [95]. Current technology enables scientists to probe the previous "black boxes" by taking a complex process apart and more closely examining the interactions between components. An important example is the process of gene expression which is now being analyzed by single-cell and single-molecule experiments [17, 76, 99, 8, 12].

1.1 Gene Expression

The process of gene expression is at the core of all known living systems: eukaryotes (including multicellular organisms), prokaryotes (bacteria and archaea) and viruses. It is the process in which the information stored in DNA is used to generate functional molecules (usually proteins) required for the life processes. In a simple point of view, gene expression consists of two processes: transcription and translation. Transcription is the first stage in gene expression during which messenger RNA (mRNA) is synthesized with the genetic information transcribed from DNA. In the translation process that follows, ribosomes bind to mRNA, decode the information stored and generate a specific amino acid chain that later folds into an active protein. Through the process of gene expression, genotypes are translated into different phenotypes.

1.2 Genotype and Phenotype

Gene expression is the fundamental process that connects genotype with phenotype. The genotype is the "blueprint" that a living system is created with and the phenotype is how it looks like on the outside. As we can imagine, phenotype is to a large extent determined by the underlying genotype. For example, different eye colors in Persians are caused by differences at the level of genes.

While different genotypes can give rise to different phenotypes, the external environment can also affect the expressed phenotype. For example, flamingos can have pink or white color depending on the food they eat; a person can have a much darker skin color if extensively exposed to sunshine. The question that follows is: what about genetically identical organisms grown in homogeneous environments? Strikingly, researchers have observed that even genetically identical individuals growing in the same environments can be very different [60]. One fundamental source of this variability is the random fluctuation in the gene expression process, which will be the focus of this dissertation.

1.3 Stochasticity in Gene Expression

It is known that the chemical reactions taking place in the cell are inherently stochastic. We can not deterministically predict when one reaction will happen or which reaction will occur first. Furthermore, the key molecules involved in the gene expression are usually present in small numbers. For example, there are usually one or two copies of DNA at any given point in the cell cycle and the mRNA levels are usually low due to low transcription rate. The small numbers of these molecules implies that the fluctuations cannot, in general, be averaged away. The Law of Mass Action which is based on statistical averaging over very large numbers of molecules does not apply. Instead, gene expression has to be modeled as a stochastic process [91, 77].

Moreover, it has been observed that many cellular regulation pathways are triggered by signals that occurs stochastically [92, 93, 4]. The intrinsic fluctuation in these regulators can thus bring about different phenotypes for isogenic cells in a homogeneous environment [16, 68, 38, 9, 38, 41, 58, 77, 79]. For example, the HIV virus can live in the infected cells in two different states: latent infection and productive infection. The switch between these two states is controlled by the key regulatory protein Tat [97]. When Tat levels are low, the HIV virus stays in the latent infection state. However, due to stochasticity in the underlying processes a high concentration of Tat protein can be reached, following which a feedback loop is triggered. The HIV virus then enters the productive infection state by producing a large amount of virus leading to the death of the cell infected.

Different phenotypes caused by such intrinsic noise provide a "bet-hedging survival strategy" to respond and adapt to the changing environment [98, 45, 57, 55]. One example is the development of persister cells [43, 52]. Persister cells are a small fraction of the population (*e.g* 10^{-5} in *E. coli*) that are highly tolerant to the antibiotics. They are not mutants. Instead, they are phenotypic variants of the wild types due to the intrinsic noise in gene expression and genetically identical to the regular cells. When exposed to antibiotics, persister cells will not grow or die and when antibiotics are removed, they can regenerate the whole population again.

Given these interesting phenomena, the topic of stochastic gene expression

and its regulation has drawn much attention recently. Both experimental and theoretical work [81, 75, 100, 69, 83, 80, 78, 3, 16, 23, 33, 44, 64, 67, 68, 72, 83] has been carried out in understanding this process at a fundamental level. While several new discoveries have already been made, it is widely regarded that current research only marks the beginning phase; the steady trends in technological innovations coupled with their continued application to novel cellular processes is expected to produce many more significant results and novel insights in the near future.

In this dissertation, we will focus on the theoretical analysis of the gene expression and its regulation. In the following chapters, we will first review some basic results which are derived in Chapter 2. Then we will consider more complex models in Chapter 3 that include the effects of molecular memory and bursting mechanisms in gene expression. In Chapter 4 and 5, we will study models of post-transcriptional regulation of gene expression.

Chapter 2

Stochastic Modeling of Gene Expression

In this chapter, we will review the basic results for models of gene expression and derive results that, for most part, have been obtained previously (sometimes using different approaches). We will first introduce the master equation for the stochastic gene expression model and discuss the useful results derived from it. Then we will extend the analysis to models with translation bursting and transcriptional bursting. Finally we will discuss a specific computational method that was used in this work. It is named the Gillespie algorithm [22], and it is broadly used as an efficient tool for stochastic simulations.

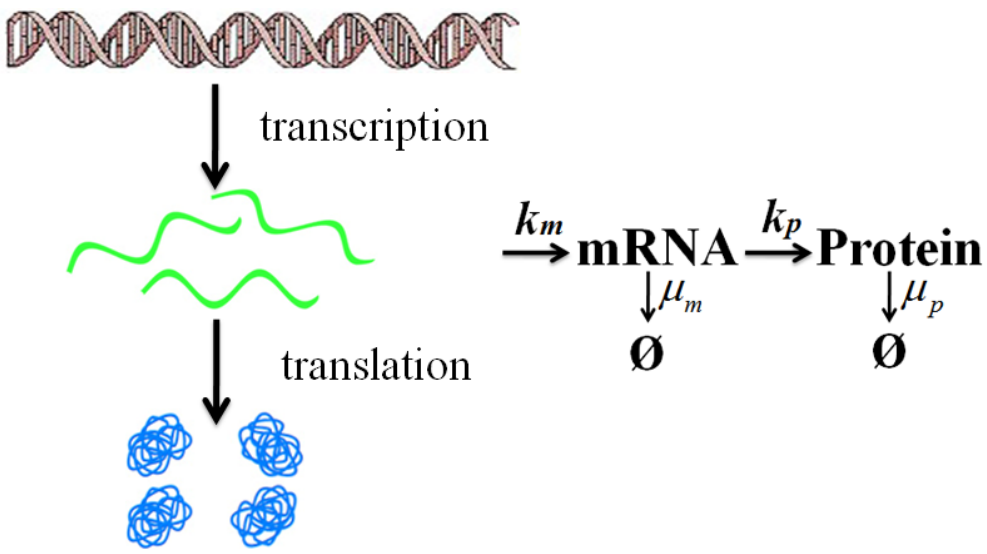

Figure 2.1: Figure of the gene expression process and the reaction scheme studied. The reaction scheme considers the transcription, translation and decaying of the molecules occurring at constant probability per unit time with rates indicated.

2.1 Stochastic Model and Master Equation

2.1.1 Deterministic Rate Equation

The reaction scheme of the simplest gene expression model is shown in Fig.(2.1) [39]. It considers only the most fundamental processes: transcription, translation and decay of the molecules. The synthesis of mRNAs happens at rate k_m during transcription and proteins are generated at rate k_p during translation. The mRNA and protein lifetimes are given as τ_m and τ_p respectively. This implies that mRNA and protein degrade respectively at rates $\mu_p = 1/\tau_p$ and $\mu_m = 1/\tau_m$.

Different approaches can be applied to analyze this reaction scheme. Per-

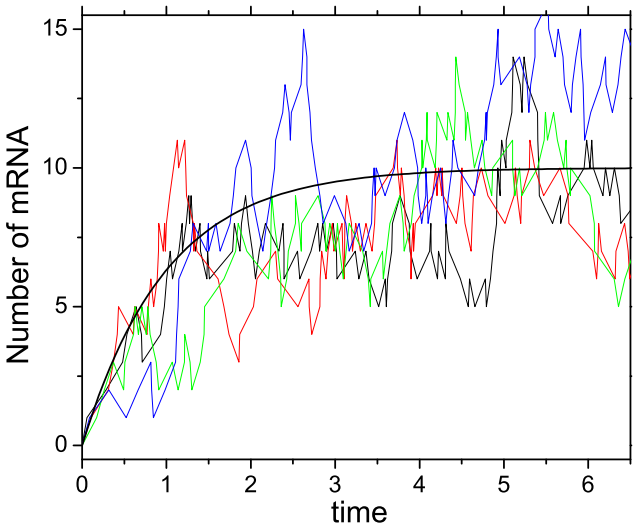

Figure 2.2: The mRNA evolution based on rate equation and simulation. The bold line corresponds to Eq.(2.2) and other curves are the simulation results by assuming each process is Poisson process. The parameters are chosen as $k_m = 10$ and $\mu_m = 1$.

haps the most simple one is to obtain the rate equations from the Law of Mass Action. Let $m(t)$ denote the number of mRNAs in the cell at time t. Correspondingly, we have the equation describing the evolution of mRNAs as

$$\frac{dm(t)}{dt} = k_m - m(t)\mu_m. \tag{2.1}$$

The solution to this equation can be found, by assuming the initial condition that $m(0) = 0$, as:

$$m(t) = \frac{k_m}{\mu_m}(1 - e^{-t}). \tag{2.2}$$

Eq.(2.2) gives a smooth curve as shown in Fig.(2.2). This is due to the fact that the Law of Mass Action considers the reaction to be macroscopic

and deterministic, which corresponds to the ensemble average of the system evolution. However, each realization of the evolution will randomly fluctuate around the values determined from the rate equations and not be smooth, as also shown in Fig.(2.2). As stated in Chapter 1, the key molecules such as DNAs are present in the cell in very low copy numbers. Given this is the case, the ensemble average will not accurately reflect the evolution for a single system. Instead, a probabilistic approach quantifying the chance that the system will stay at a certain state will be more helpful. As the result, the stochastic approach is needed.

2.1.2 Master Equation and Generating Function

The connection between probability functions and the rates is not as straight forward as it seems to be. The rates (*e.g.* mRNA production rate) are macroscopic quantities that are the measurements of ensemble averages. On another hand, the probability that one reaction happens next instance depends on microscopic properties of the reaction itself. Generally one can not infer a microscopic quantity (probability) based on a macroscopic quantity (rate).

The bridge across this gap is to assume that these fundamental processes (*e.g.* transcription, translation and decay of molecules) are essentially Poisson processes. With this assumption, the probability that one reaction with rate k will take place during the small time interval t to $t+\Delta t$ is $k\Delta t$. Furthermore,

if there are N of these reactions each with rate k, the probability that one reaction will occur during the small time interval Δt is $Nk\Delta t$ (this is discussed further later). Denote $P(m,t)$ as the probability that there are m mRNAs at time t, we can have the following equation describing the evolution of mRNA:

$$\begin{aligned} P(m,t+\Delta t) &= k_m \Delta t P(m-1,t) + (m+1)\mu_m \Delta t P(m+1,t) \\ &+ (1 - k_m \Delta t - m\mu_m) P(m,t). \end{aligned} \tag{2.3}$$

The first two terms in Eq.(2.3) correspond to the probability that some reaction takes the system to the m mRNAs state and the last term is the probability that the system stays in the m mRNAs state.

In the limit that $\Delta t \sim dt \to 0$, Eq.(2.3) can be simplified as a partial differential equation as:

$$\begin{aligned} \frac{\partial P(m,t)}{\partial t} &= k_m P(m-1,t) + (m+1)\mu_m P(m+1,t) \\ &- (k_m + m\mu_m) P(m,t). \end{aligned} \tag{2.4}$$

Eq.(2.4) that describes the time evolution of probability function $P(m,t)$ is also called the Master equation. Eq.(2.4) is solvable but the solution is somewhat complicated. In many cases, what is of primary interest is the system steady-state behavior, *i.e.* by taking $t \to \infty$. The steady-state solution to Eq.(2.4) is a Poisson distribution as

$$P(m) = \lim_{t\to\infty} P(m,t) = \frac{(k_m/\mu_m)^m}{m!} e^{-k_m/\mu_m}. \tag{2.5}$$

Following a similar procedure, we can derive the master equation for the gene expression model in Fig.(2.1) as:

$$\begin{aligned}\frac{\partial P(m,n,t)}{\partial t} &= k_m\big(P(m-1,n,t)-P(m,n,t)\big)\\ &+ \mu_m\big((m+1)P(m+1,n,t)-mP(m,n,t)\big)\\ &+ k_p\big(mP(m,n-1,t)-mP(m,n,t)\big)\\ &+ \mu_p\big((n+1)P(m,n+1,t)-nP(m,n,t)\big),\end{aligned} \tag{2.6}$$

where $P(m,n,t)$ is the transient distribution of the probability that m mRNAs and n proteins are present in the system at time t.

Correspondingly, the equation for the steady-state distribution $P(m,n) = \lim_{t\to\infty} P(m,n,t)$ is

$$\begin{aligned}0 &= k_m\big(P(m-1,n)-P(m,n)\big)\\ &+ \mu_m\big((m+1)P(m+1,n)-mP(m,n)\big)\\ &+ k_p\big(mP(m,n-1)-mP(m,n)\big)\\ &+ \mu_p\big((n+1)P(m,n+1)-nP(m,n)\big).\end{aligned} \tag{2.7}$$

The typical way to solve Eq.(2.7) is to apply generating functions to convert the discrete form of equation to the differential form. Specifically, let us define function $G(z_1,z_2)$ as

$$G(z_1,z_2)=\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} z_1^m z_2^n P(m,n). \tag{2.8}$$

In simplification of Eq.(2.7), we need to use some properties of the function $G(z_1, z_2)$ that are listed below:

$$\begin{aligned}
\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} z_1^m z_2^n P(m-1,n) &= z_1 G(z_1,z_2) \\
\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} z_1^m z_2^n m P(m,n) &= z_1 \frac{\partial G(z_1,z_2)}{\partial z_1} \\
\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} z_1^m z_2^n (m+1) P(m+1,n) &= \frac{\partial G(z_1,z_2)}{\partial z_1}.
\end{aligned} \tag{2.9}$$

Following this, we can covert Eq.(2.7) into a differential equation:

$$\begin{aligned}
0 &= k_m(z_1-1)G + \mu_m(1-z_1)\partial_{z_1}G \\
&+ k_p(z_2-1)z_1\partial_{z_1}G + \mu_p(1-z_2)\partial_{z_2}G.
\end{aligned} \tag{2.10}$$

To simplify the expression, the notation $\partial_{z_i}G$ is used to denote the partial derivative of function $G(z_1, z_2)$ with respect to z_i $(i = 1, 2)$.

Eq.(2.10) can be solved by the method of characteristics [86]. While this approach is somewhat complicated and the solution is non-trivial, it is relatively easy to derive the mean and variance of mRNA and protein level from Eq.(2.10) based on the basic properties of generating function, *e.g.*:

$$\begin{aligned}
\langle m \rangle &= \partial_{z_2} G|_{z_1=z_2=1} \\
\langle m(m-1) \rangle &= \partial^2_{z_2^2} G|_{z_1=z_2=1}
\end{aligned} \tag{2.11}$$

From there, we will have the following results:

$$\begin{aligned}
\langle m_s \rangle &= \frac{k_m}{\mu_m} \\
\sigma^2_{m_s} &= \frac{k_m}{\mu_m} = \langle m \rangle \\
\langle p_s \rangle &= \frac{k_m k_p}{\mu_m \mu_p} \\
\sigma^2_{p_s} &= \frac{k_m k_p}{\mu_m \mu_p}(1 + \frac{k_p}{\mu_m + \mu_p}) = \langle p_s \rangle (1 + \frac{k_p}{\mu_m + \mu_p}),
\end{aligned} \tag{2.12}$$

where the symbols $\langle .. \rangle$ and σ^2 are used to denote the mean and variance, the variable m_s and p_s are steady-state mRNA and protein number, respectively.

2.1.3 Noise in Protein steady-state Distribution

To quantify the noise in gene expression, we use the squared coefficient of variance. From Eq.(2.12), we obtain

$$\begin{aligned}
\frac{\sigma^2_{m_s}}{\langle m_s \rangle^2} &= \frac{1}{\langle m_s \rangle} \\
\frac{\sigma^2_{p_s}}{\langle p_s \rangle^2} &= \frac{1}{\langle p_s \rangle} + \frac{1}{\langle m_s \rangle} \frac{\mu_p}{\mu_m + \mu_p}.
\end{aligned} \tag{2.13}$$

We can see from Eq.(2.13) that the noise in mRNA steady-state distribution is the same as that of a Poisson distribution, which vanishes when the mean level is high (*e.g.* with large transcription rate k_m). This is because the mRNA production and decaying are assumed to be Poisson processes that leads to a Poisson distribution for the steady-state. On the other hand, the

noise in protein steady-state distribution is greater than that of a Poisson distribution: there is another term that contributes to the noise. This term is significant as the mRNA mean level is typically low in the cell. This indicates that the noise in protein steady-state distribution is mainly due to the fluctuation (or small number) of mRNAs.

2.2 Burst Synthesis Approximation

Several recent single-molecule experimental studies [7, 20, 26, 28, 42, 46, 47, 59, 64, 69, 85, 91, 90, 92, 93] have seen protein production occurring in random and short bursts, presumably occurring upon the production of mRNA. This observation is explained by the fact that the mRNA liefime is typically much shorter than the protein lifetime (*i.e.* $\mu_m \gg \mu_p$ and $\tau_p \gg \tau_m$). When analyzing the system at short time scales, the protein number increases linearly with time when mRNA is present. This is because protein production happens during the mRNA lifetime. On the other hand, when analyzing the system at large time scales, system dynamics during the mRNA lifetime can be neglected and proteins are approximated to be produced instantaneously right after the mRNA production. Under this approximation, the proteins are assumed to be produced by independent bursts with random size arising from the underlying mRNA production. This approximation is shown in Fig.(2.3) by simulation results.

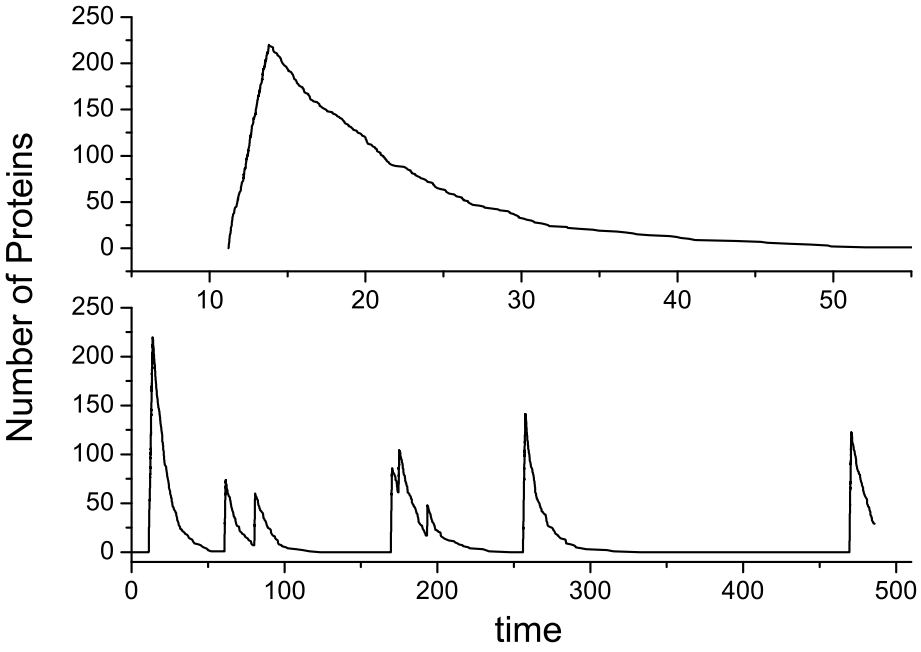

Figure 2.3: The time evolution of protein numbers based on simulation when $\mu_m \gg \mu_p$. The graph on top is in small time scale and the one on bottom is in large time scale. When the system is analyzed at large time scale, the protein evolution can be characterized by the processes of burst creation of proteins and the decaying of proteins alone and the mRNA evolution can be neglected.

Motivated by the experimental observation of translational bursts, we can apply the so-called "burst synthesis approximation" to analytically solve the stochastic gene expression model. In the limit that $\mu_m \gg \mu_p$ (and $\tau_p \gg \tau_m$) which is valid in many cellular systems, the original reaction scheme can be simplified as a process of protein burst production and degradation alone [20, 86, 13]. Denoting p_b as the number of proteins produced in one translational burst and $P_{pb}(n)$ as its probability density function, the Master

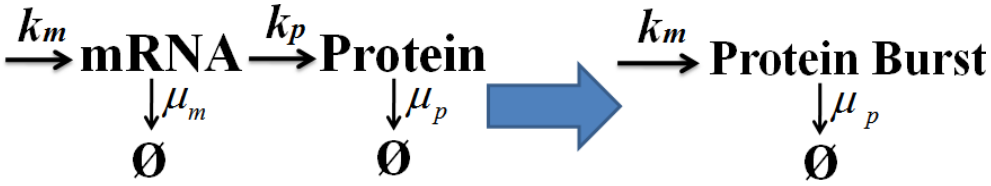

Figure 2.4: The original reaction scheme of gene expression can be approximated by the process of protein burst production and protein decaying. The simplified reaction scheme is used to find the steady-state solution once the protein burst size distribution is given.

equation for the protein steady-state distribution $P(n)$ gives

$$\begin{aligned} 0 &= k_m\Big(\sum_{i=0}^{n} P(n-i)P_{pb}(i) - P(n)\Big) \\ &+ \mu_p\big((n+1)P(n+1) - nP(n)\big). \end{aligned} \tag{2.14}$$

Eq.(2.14) can be solved by applying generating functions as introduced in the preceding section. Define $G(z) = \sum_{n=0}^{\infty} z^n P(n)$ and $G_{pb}(z) = \sum_{n=0}^{\infty} z^n P_{pb}(n)$, Eq.(2.14) can be rewritten in terms of the generating functions as:

$$\mu_p(z-1)\frac{dG(z)}{dz} = k_m G(z)(G_{pb}(z) - 1). \tag{2.15}$$

The solution of $G(z)$ is given by

$$G(z) = \frac{k_m}{\mu_p} \times \exp\Big(\int_1^z \frac{G_{pb}(x) - 1}{x - 1} dx\Big). \tag{2.16}$$

Based on Eq.(2.16), $G(z)$ can be found once the expression of $G_{pb}(z)$ is known. Equivalently, once the protein burst size distribution is $P_{pb}(n)$ is known, the protein steady-state distribution can be determined. We can see that

Eq.(2.16) serves as a bridge connecting the protein burst and steady-state distribution.

It is also note worthy that though we only take the basic gene expression in the above discussion, the "burst synthesis approximation" can be used to probe more complicated problems with regulations and molecular memory considered (as will be discussed in the following chapters) as long as each burst is independent. The burst approximation is equivalently integrating all processes affecting protein production into a "black-box" that gives rise to different burst size distribution $P_{pb}(n)$. In analyzing other problems, we can separate the process into two different steps: first obtain the burst size distribution based on the reaction scheme and then derive the steady-state protein distribution from Eq.(2.16).

The protein burst size distribution $P_{pb}(n)$ depends on the specific reaction scheme. For the basic gene expression model introduced here, the burst size distribution follows a geometric distribution, which is also confirmed by experiments [99, 8]. A simple argument can be given as follows. The protein burst size distribution measures how many proteins one mRNA can produce before it degrades. When one mRNA is present, the next reaction taking place can have only two options, which is either the creation of a protein or the degradation of the mRNA. This is the similar to the process when we toss a coin and can have either a heads or a tails. As the result, the probability that n proteins are produced before the mRNA decays is equivalent to the

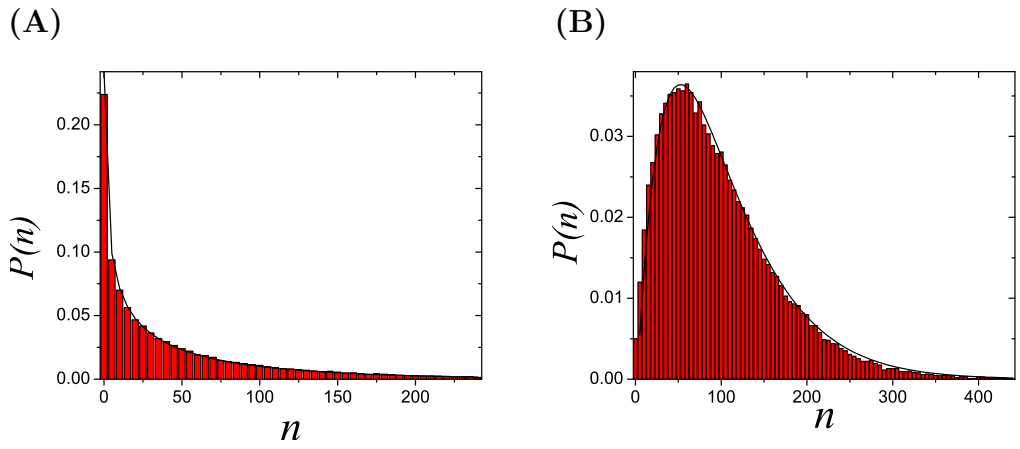

Figure 2.5: Protein steady-state distributions $P(n)$ with different parameters. (A) $k_m = 0.05$, $\mu_m = 1$, $k_p = 100$ and $\mu_p = 0.1$ (B) $k_m = 0.2$, $\mu_m = 1$, $k_p = 50$ and $\mu_p = 0.1$. The lines correspond to the binomial distribution derived in Eq.(2.19) that agree with simulation results.

probability of n heads appearing before the appearance of a tails when tossing the coin. For the latter problem, it is known that probability follows the geometric distribution. Correspondingly, the protein burst size distribution will be geometric as well.

The analytical approach (which will be introduced in Chapter 4) also yields the same result. The detailed expression of $P_{pb}(n)$ and $G_{pb}(z)$ are given in below:

$$\begin{aligned} P_{pb}(n) &= \left(\frac{k_p}{\mu_m + k_p}\right)^n \frac{\mu_m}{\mu_m + k_p} \\ G_{pb}(z) &= \frac{\mu_m z}{\mu_m + k_p(1-z)} \end{aligned} \tag{2.17}$$

By taking the expression of $G_{pb}(z)$ into Eq.(2.16), we can see that

$$G(z) = \big(\frac{\mu_m z}{\mu_m + k_p(1-z)}\big)^{\frac{k_m}{\mu_p}}. \tag{2.18}$$

Correspondingly, the protein steady-state distribution is a negative binomial distribution given by

$$P(n) = \binom{n + \frac{k_m}{\mu_p} - 1}{n} \big(\frac{k_p}{\mu_m + k_p}\big)^n \big(\frac{\mu_m}{\mu_m + k_p}\big)^{\frac{k_m}{\mu_p}}. \tag{2.19}$$

The result derived agrees with the experimental observation [99, 8] and numerical simulations (in Fig.(2.5)). Furthermore, we can analytically derive the mean and variance and compare them with those in Eq.(2.12). The mean and variance for the basic gene expression model under burst synthesis approximation are

$$\begin{aligned} \langle p_s \rangle &= \frac{k_m k_p}{\mu_m \mu_p} \\ \sigma^2_{p_s} &= \frac{k_m k_p}{\mu_m \mu_p}(1 + \frac{k_p}{\mu_p}) = \langle p_s \rangle (1 + \frac{k_p}{\mu_p}), \end{aligned} \tag{2.20}$$

By comparing Eq.(2.12) and Eq.(2.20), we notice that the results based on basic reaction scheme will be identical to those with the burst approximation when $\mu_m \gg \mu_p$, which is the basic assumption the burst approximation is based on.

For a more general case, such that the expression of $G_{pb}(z)$ is not given, we

can still find the mean and variance from Eq.(2.16) as

$$\begin{aligned}
\langle p_s \rangle &= \frac{k_m}{\mu_p}\langle p_b \rangle \\
\frac{\sigma^2_{p_s}}{\langle p_s \rangle^2} &= \frac{1}{\langle p_s \rangle} + \frac{\mu_p}{2k_m}\big(1 + \frac{\sigma^2_{p_b}}{\langle p_b \rangle^2} - \frac{1}{\langle p_b \rangle}\big) \\
&= \frac{1}{\langle p_s \rangle} + \frac{1}{\langle m_s \rangle}\big(1 + \frac{\sigma^2_{p_b}}{\langle p_b \rangle^2} - \frac{1}{\langle p_b \rangle}\big) \times \frac{\mu_p}{\mu_m},
\end{aligned} \tag{2.21}$$

where p_s and m_s are respectively the protein and mRNA steady state number and p_b quantifies the protein burst size. Eq.(2.21) shows how different gene expression processes with different burst size distribution can change the noise in steady-state protein distribution. A more detailed study will be conducted in Chapter 3.

2.3 mRNA Transcriptional Bursting

In the preceding section, we analyzed the effect of translational bursting as observed in experiments. On the other hand, it is also observed in experiments that mRNA can be produced in transcriptional burst [75]. This observation is explained by a DNA two-state model that the DNA can switch randomly between transcription active (on) and inactive (off) states [71]. The mRNA can only be created when DNA is in the on state. In the limit that mRNA degradation rate is much smaller than the DNA off rate, the mRNA production can be considered to happen in bursts.

Mathematically, the translational and transcriptional bursting models have many features in common. In comparing the two, the DNA on and off state is analogous to the state with and without a single mRNA present. Similar to the protein bursting, the two-state model will give rise to the mRNA burst size as a geometric distribution and the steady state distribution will be a negative binomial distribution. This result provides a useful way to identify the existence and the degree of the transcriptional burst by measuring the Fano factor (the ratio of variance to mean) of mRNA steady state distribution. If the transcriptional process is a simple Poisson process, the steady-state mRNA distribution will be Poisson with Fano factor exactly equal to one. Otherwise, if the transcription happens in bursts, the Fano factor equals the mean number of mRNA during a burst [75].

The effect of transcriptional bursting on protein steady-state distribution can be analyzed by studying the master equation. It is noteworthy that we can do this even without knowing the form of mRNA burst size distribution. Here we focus on the basic gene expression model described at the beginning of the chapter and assume that transcriptional bursting occurs at rate k_m. Denote by m_b the number of mRNA produced in one burst and $P_{mb}(m)$ as its probability density function. We can write the Master equation of mRNA

and protein joint steady-state distribution $P(m,n)$ as

$$\begin{aligned} 0 &= k_m\big(\sum_{i=0}^{m} P(m-i,n)P_{mb}(i) - P(m,n)\big) \\ &+ \mu_m\big((m+1)P(m+1,n) - mP(m,n)\big) \\ &+ k_p\big(mP(m,n-1) - mP(m,n)\big) \\ &+ \mu_p\big((n+1)P(m,n+1) - nP(m,n)\big). \end{aligned} \tag{2.22}$$

Eq.(2.22) considering transcriptional bursting is similar to Eq.(2.7) without bursting except for the first term. Similarly, we can apply the generating function approach to analyze it. Defining $G_{mb}(z) = \sum_{m=0}^{\infty} z^m P_{mb}(m)$ as the generating function of the mRNA burst size distribution and recalling the definition of $G(z_1, z_2)$ in section 1, we get

$$\begin{aligned} 0 &= k_m(G_{mb} - 1)G + \mu_m(1 - z_1)\partial_{z_1}G \\ &+ k_p(z_2 - 1)z_1\partial_{z_1}G + \mu_p(1 - z_2)\partial_{z_2}G. \end{aligned} \tag{2.23}$$

Following a similar approach as in section 1, we can find the mean and vari-

ance of steady-state mRNA and protein distributions as

$$\begin{aligned}
\langle m_s \rangle &= \frac{k_m \langle m_b \rangle}{\mu_m} \\
\sigma^2_{m_s} &= \frac{k_m}{2\mu_m}(\sigma^2_{m_b} + \langle m_b \rangle^2 + \langle m_b \rangle) \\
\langle p_s \rangle &= \frac{k_m k_p \langle m_b \rangle}{\mu_m \mu_p} \\
\sigma^2_{p_s} &= \frac{k_m k_p}{2\mu_m \mu_p}\Big(\frac{k_p(\sigma^2_{m_b} + \langle m_b \rangle + \langle m_b \rangle^2)}{\mu_m + \mu_p} + 2\langle m_b \rangle\Big) \\
&= \langle p_s \rangle + \frac{\langle p_s \rangle^2}{2}\frac{\mu_m}{k_m}\Big(\frac{\sigma^2_{m_b}}{\langle m_b \rangle^2} + \frac{1}{\langle m_b \rangle} + 1\Big)\frac{\mu_p}{\mu_m + \mu_p} \\
&= \langle p_s \rangle + \frac{\langle p_s \rangle^2}{2}\frac{\langle m_b \rangle}{\langle m_s \rangle}\Big(\frac{\sigma^2_{m_b}}{\langle m_b \rangle^2} + \frac{1}{\langle m_b \rangle} + 1\Big)\frac{\mu_p}{\mu_m + \mu_p}. \qquad (2.24)
\end{aligned}$$

Furthermore, the noise term of steady-state protein level can be found as

$$\frac{\sigma^2_{p_s}}{\langle p_s \rangle^2} = \frac{1}{\langle p_s \rangle} + \frac{\langle m_b \rangle}{2\langle m_s \rangle}\Big(\frac{\sigma^2_{m_b}}{\langle m_b \rangle^2} + \frac{1}{\langle m_b \rangle} + 1\Big)\frac{\mu_p}{\mu_m + \mu_p}. \qquad (2.25)$$

Eq.(2.25) demonstrates how transcriptional bursting can contribute to the noise in protein levels. We can see that when there is no transcriptional bursting, $\langle m_b \rangle = 1$ and $\sigma_{m_b} = 0$, the noise term expressed in Eq.(2.25) will be identical to previously derived Eq.(2.13). However, the noise will change when transcriptional bursting happens. First, different transcription mechanism can give rise to different mRNA burst size distributions with different values of $\sigma^2_{m_b}/\langle m_b \rangle^2$. More importantly, transcriptional bursting can be associated with a reduction in the transcription frequency, while keeping the same level of mRNA and protein. Typically, infrequent transcription will give rise to

greater noise. As shown in Eq.(2.25), the second term will be scaled by the mean transcriptional burst size. As a result, transcription bursting is an important source of noise in gene expression [88].

2.4 Simulation Algorithm

To test the analytical results derived, we use numerical simulations and thus need good simulation algorithms that can provide accuracy and efficiency. One successful method for the numerical simulations of the stochastic gene expression is the Gillespie algorithm.

To start with, let us first introduce the discrete time Monte Carlo simulation algorithm. In our model we typically consider all processes are essentially Poisson processes. As a basic property, a Poisson process with rate k will have probability $k\Delta t$ to take place during the infinitesimally small time interval Δt. Naturally, one can develop a simulation algorithm based on this property by dividing the time trace into very small discrete time intervals. At every time instance, a random number r uniformly distributed between [0,1] is generated. If r is no greater than $k\Delta t$, the process happens, otherwise we move to the next time step and repeat the above process.

This discrete time simulation is accurate if the time interval is properly chosen to be small enough. However, for small Δt value, the probability that the process occurs will also be small and the total number of time divisions will

be large. This means that in order to analyze the system from time 0 to time T, we need to check for a large number of discrete time instances (about $T/\Delta t$). Furthermore, most of the computation is wasted on time intervals without the process occurring, given the small probability of its occurrence. In the other word, the rejection rate is very high. In such cases, the discrete time Monte Carlo simulation is not very efficient.

2.4.1 Waiting-Time Distribution

From the above discussion, we see that the main drawback of the discrete time Monte Carlo simulation is that a lot of computation is wasted in determining when the process will happen. On the other hand, if we could directly determine when the process happens and which process happens (if there are multiple Poisson processes), the efficiency can be greatly improved.

To do so, we need further properties of Poisson processes. We first look into the waiting-time distribution for the Poisson process. The waiting time, denoted by T, is a random variable measuring the time elapsed from the last occurrence of the reaction until the time that it happens again. For the Poisson process, the distribution of T is given by an exponential distribution. This can be proved as follows. Consider a Poisson process with rate k and divide the time trace into very small time intervals Δt. The probability that

the reaction occurs during the time interval $(t, t + \Delta t)$ is

$$P(T \in (t, t + \Delta t)) = (1 - k\Delta t)^n k\Delta t, \tag{2.26}$$

where $n = t/\Delta t$. The first term in above equation corresponds to the probability that the reaction does not happen before $T \leq t$ and the latter term is the probability that reaction occurs during the interval $(t, t + \Delta t)$. In the limit that $\Delta t = dt \to 0$ and $n \to \infty$, we have

$$\begin{aligned} P(T \in (t, t + dt)) &= k \lim_{n\to\infty} (1 - \frac{kt}{n})^n dt \\ &= ke^{-kt} dt, \end{aligned} \tag{2.27}$$

which means that the waiting-time distribution is exponential.

2.4.2 Memoryless Distribution

One important property of exponential distribution is that it is memoryless. The mathematical expression for memoryless property is given as

$$P(T > t|T > s) = P(T > t - s), \tag{2.28}$$

for any $t > s$. Assuming the probability density function of T is ke^{-kt} and $t > s$, we can be approved the above relationship as follows.

$$\begin{aligned} P(T > t|T > s) &= \frac{P(T > t \& T > s)}{P(T > s)} \\ &= \frac{P(T > t)}{P(T > s)} = \frac{e^{-kt}}{e^{-ks}} \\ &= e^{-k(t-s)} = P(T > t - s). \end{aligned} \tag{2.29}$$

Literally this means that the probability that the reaction happens in the future does not depend on any other information from the past. Here is one example to better illustrate the concept of memorylessness.

Consider you are measuring the waiting-time distribution of a Poisson process using a clock. The problem is that the clock is old and it will randomly stop and you have to reset it to have it run again. To overcome this problem, you apply the following strategy: if the clock stops and the reaction takes place, you have an exact waiting-time measurement; otherwise, if the clock stops and the reaction does not happen yet, you reset the clock immediately and the time counting will start from zero. The question is, will you have the correct measurement of waiting-time distribution with this clock?

The answer is yes, reseting the clock will not affect the measurement because Poisson process has exponential waiting-time distribution that is memoryless. The reseting of the clock will not affect the distribution of time for the reaction to take place.

2.4.3 Multiple Poisson Processes

Another property of the exponential distribution is that the minimum of two exponential random variable is still exponentially distributed. Consider two Poisson process with rate k_1 and k_2, the waiting-time is denoted by T_1 and T_2 respectively. Let $T = \min(T_1, T_2)$, then the distribution of T follows an exponential distribution with mean $1/(k_1 + k_2)$. This can be derived by the following calculation:

$$\begin{aligned} P(T > t) &= P(T_1 > t \& T_2 > t) = P(T_1 > t)P(T_2 > t) \\ &= e^{-k_1 t} e^{-k_2 t} = e^{-(k_1+k_2)t}, \end{aligned} \tag{2.30}$$

which proves the statement above.

We can also calculate the probability that T_1 is greater than T_2. This is given as follows:

$$\begin{aligned} P(T_1 > T_2) &= \int_0^\infty k_1 e^{-k_1 t} dt_1 \int_0^{t_1} k_2 e^{-k_2 t} dt_2 \\ &= \frac{k_1}{k_1 + k_2}. \end{aligned} \tag{2.31}$$

Similarly, the probability that $T_1 < T_2$ is $k_2/(k_1 + k_2)$.

The properties stated above have very useful applications in simulations. Now we know that if there are two Poisson processes with rate k_1 and k_2, then the time elapsed from now till some reaction happens follows an exponential distribution with mean $1/(k_1 + k_2)$. The probability that the reaction taking

place is the one with rate k_1 is $k_1/(k_1 + k_2)$ and vice versa. This reasoning can be extended to the case that there are n Poisson processes each with rate k_i where $i = 1$ to n. The waiting-time distribution that one among these n reactions happens is exponential with mean $1/\sum_{i=1}^{n} k_i$. The probability that it is the i^{th} reaction that occurs is $k_i/\sum_{i=1}^{n} k_i$.

2.4.4 Gillespie Algorithm

The Gillespie algorithm improves the efficiency by directly determining when the reaction happens and which reaction occurs using the properties of Poisson processes discussed above [22]. In the Gillespie algorithm, we first sum up all the rates of Poisson processes that may occur. Then we generate an exponential random variable based on this total rate. The random variable generated is the waiting-time for the occurrence of one of these reactions. We can also determine which reaction it corresponds to based on the probability that each reaction happens. After this, we will repeat the preceding procedure. Note that the reactions that do not happen will not be affected due to the property of memorylessness.

Now the only detail left is how to generate an exponential random variable. The exponential random variable is generated based on the property that its cumulative distribution function varies uniformly from 0 to 1. If we pick a random number, denoted as r, from a uniformly distributed interval [0, 1], then $t = -\frac{1}{k}ln(r)$ follows an exponential distribution with mean $1/k$.

With this information, we can run the simulation for the basic gene expression problem (the reaction scheme shown in Fig.(2.1)) with Gillespie algorithm as follows:

0.) at time T, there are m mRNA's and n proteins.

1.) Find the rate $k = k_m + m\mu_m + mk_p + n\mu_p$.

2.) Generate a random variable r_1 from a uniform distribution in [0,1]. $t = -\frac{1}{k}ln(r_1)$ is the waiting-time that one reaction takes place.

3.) Generate a random variable r_1 from a uniform distribution in [0,1]. If $r_1 \leq k_m/k$, it is the transcription that occurs and mRNA number increases by 1. If $k_m/k \leq r_1 \leq (k_m + m\mu_p)/k$, protein will decay and protein number decreases by 1, and so on.

4.) Update the protein and mRNA number and set $T = T + t$.

5.) Repeat from step 0.) (because exponential distribution is memoryless).

Chapter 3

Noise with Molecular Memory and Bursting

As discussed in the preceeding chapter, in order to carry out a stochastic analysis of the gene expression process, one has to make certain assumptions about the underlying molecular mechanisms. The typical assumption is to consider the basic processes of molecular creation and degradation as Poisson processes with memoryless exponential waiting time distributions [41, 70, 80, 92]. One benefit of this assumption is that we can then easily relate the probability function with the macroscopic rates that are experimentally measurable. From there we can apply the mathematical framework already established for stochastic Poisson process and solve the problem analyzed.

Nevertheless, it is worthwhile to consider the validity of this assumption and to examine cases where it breaks down. It has been observed that transcription events in *Escherichia coli* often follow exponential waiting time distri-

bution [99, 8, 23]. However, the creation and decaying of molecules typically involve multiple microscopic substeps and are controlled by complicated reaction networks. This can generate non-exponential waiting time distribution for the events studied. Unless one sub-step is rate limiting, it is inadequate, in general, to model transcription, translation and degradation as elementary Poisson processes [72]. Though we can expand the analysis to all the sub-processes involved and consider them as Poisson process, this approach requires detailed knowledge of all sub-processes in the gene expression network and hard to practice in general.

Motivated by these observations, a model was introduced by Pedraza and Paulsson (PP model)[72] which considers *arbitrary* waiting-time distributions for processes governing arrival and decay of mRNAs (corresponding to 'gestation' and 'senescence' effects). The transcriptional bursting that gives rise to mRNA burst production is also considered in the PP model. For this model, analytical results were derived for the noise in steady-state protein distributions. However the results derived are approximate expressions. Furthermore, the PP model only considers the most basic translation process that each mRNA gives rise to a geometric protein burst distribution; however more general schemes of gene expression (e.g. involving post-transcriptional regulation [35]) can give rise to bursts that deviate significantly from a geometric distribution. Given the diversity of cellular regulatory mechanisms for gene expression, it is of interest to consider models with arbitrary burst distributions for protein production from mRNAs, acting in combination with

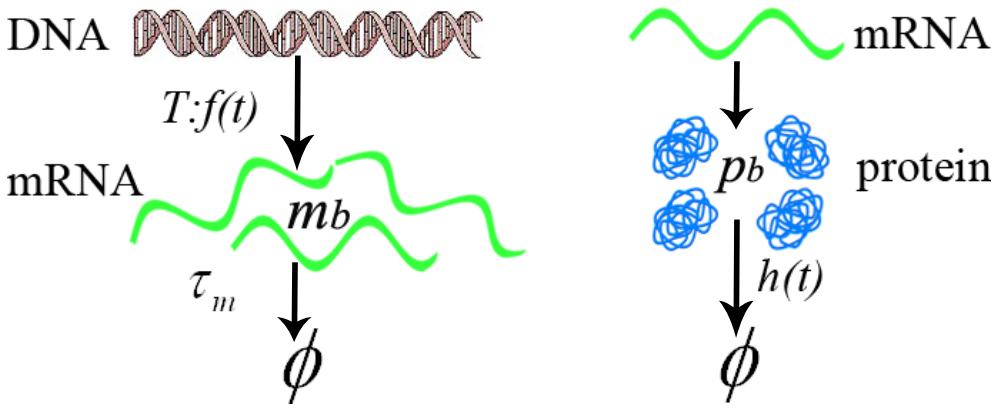

Figure 3.1: Reaction scheme for the underlying gene expression model. Production of mRNAs occurs in bursts (characterized by random variable m_b with arbitrary distribution) and each mRNA gives rise to a burst of proteins (characterized by random variable p_b with arbitrary distribution) before it decays (with lifetime τ_m). The waiting-time distributions for transcriptoinal burst and decay of proteins are characterized by the functions $f(t)$ and $h(t)$ respectively.

arbitrary distributions characterizing gestation and senescence effects. The following sections focus on analyzing such general stochastic models of gene expression and on deriving analytical expressions for the corresponding noise in protein distributions [34].

3.1 Model and Method

The underlying reaction scheme for the models analyzed in this work is shown in Fig.(3.1). Production of mRNAs occurs in independent bursts and the number of mRNAs produced in a single burst is characterized by the random variable m_b. Each mRNA independently gives rise to a random number of proteins (characterized by random variable p_b) during its lifetime τ_m. The

time interval between the arrival of consecutive bursts is characterized by random variable T with corresponding probability density function (p.d.f) $f(t)$. Proteins are degraded independently and the waiting-time distribution for protein decay is characterized by the p.d.f $h(t)$ with mean protein lifetime τ_p.

As discussed in Chapter 2, in the limit that the mRNA lifetime (τ_m) is much shorter than the protein lifetime (τ_p), i.e. $\tau_m \ll \tau_p$ (which holds for many cellular systems), the evolution of cellular protein concentrations can be modeled by processes governing arrival and decay of proteins alone [20, 86]. Unless otherwise stated, the analysis in this chapter will focus on this 'burst' limit, in which proteins are considered to arrive in independent instantaneous bursts arising from the underlying mRNA burst. In this limit, we have shown [14] that the processes involved in gene expression can be mapped to the problems of interest in queueing theory.

Queueing theory is a branch of applied mathematics that studies properties of waiting lines in diverse industrial applications and has been developed further in fields such as operations research (a branch of industrial engineering). The results of queueing theory are often used to determine the resources needed to provide service or to optimize the efficiency of the service center. However, by applying proper mapping, the results in Queueing theory can also be utilized to study gene expression models. In this mapping, individual proteins are the analogs of customers in queueing models. The burst synthesis of proteins

then corresponds to the arrival of customers in 'batches', whereas the protein decay-time distribution is the analog of the service-time distribution for each customer. Given that degradation of each protein is independent of others in the system, the process maps on to queueing systems with infinite servers. Correspondingly, the gene expression model in Fig.(3.1) maps on to what is known as a $GI^X/G/\infty$ system in the queueing literature. In this notation, the symbol G refers to the general waiting-time distribution and I^X indicates that the customers arrive in batches of random size X, where X is drawn independently each time from an arbitrary distribution.

The $GI^X/G/\infty$ system has been analyzed in previous work in queueing theory [53]. In the following, we briefly review the notation and relevant results from the queueing theory analysis. Similar to our model shown in Fig.(3.1), $f(t)$ and $h(t)$ denote the probability density function (p.d.f) for the customer arrival time and service time respectively, with $F(t)$ and $H(t)$ as the corresponding cumulative density functions (c.d.f). The distribution of batch size X has the corresponding generating function $A(z)$, defined as $A(z) = \sum_{i=1}^{\infty} P(X = i)z^i$. The kth factorial moment of batch size X, denoted by A_k, is given by $A_k = (d^k A(z)/dz^k)|_{z=1}$. The number of customers in service at time t is denoted by $N(t)$. The analytical expressions have been derived for the r^{th} binomial moment $B_r(t)$ of $N(t)$ [53]. These results can be used to derive expressions for all the moments of $N(t)$, for example $E[N(t)] = B_1(t)$ and $Var[N(t)] = 2B_2(t) + B_1(t) - B_1^2(t)$.

In the following, we will focus on two general subcategories of the $GI^X/G/\infty$ system for which closed-form analytical expressions can be derived for the mean and variance of steady-state protein distributions across the population of cells. These correspond to the two cases: A) arbitrary distributions for bursting with gestation effect in mRNA production and a Poisson process governing protein degradation and B) arbitrary distributions for bursting with senescence effect in protein decaying and a Poisson process governing burst arrival. These systems are analyzed further below.

3.2 Results

3.2.1 Effect of Gestation and Bursting in mRNA Production

Consider first case A, for which arbitrary gestation and bursting effects are included. In this case, the random variable T characterizing the time interval between bursts is drawn from an arbitrary p.d.f. $f(t)$. The protein decay-time distribution $h(t)$ is taken to be an exponential function with $h(t) = \mu_p e^{-\mu_p t}$. The protein degradation rate is denoted as μ_p mean protein lifetime is given by $\tau_p = 1/\mu_p$.

The corresponding queueing system is $GI^X/M/\infty$ where M indicates that the process of customer departure, which is the analog of protein decay, is Markovian. Recall the definition that $A(z)$ corresponds to the generating function of customer batch size, A_k is the factorial moment of customer batch

size and $N(t)$ denotes the number of customers at time t. The previous analysis [53] has derived expressions for the steady-state mean and variance corresponding to $N = \lim_{t\to\infty} N(t)$ for the $GI^X/M/\infty$ queue as [54]:

$$\begin{aligned} E[N] &= \frac{1}{\mu_p \langle T \rangle} A_1 \\ Var[N] &= E[N](1 + \frac{f_L(\mu_p)}{1 - f_L(\mu_p)} A_1 - E[N] + \frac{A_2}{2A_1}), \end{aligned} \tag{3.1}$$

where $\langle T \rangle$ is the mean of p.d.f $f(t)$ and $f_L(s)$ is the Laplace transform of $f(t)$.

To translate the result Eq.(3.1) into an expression for the noise in protein distributions, we need to derive expressions for A_1 and A_2 in terms of variables characterizing mRNA and protein burst distributions. In gene expression problem, $A(z)$ is the generating function of burst size distribution. In general, each mRNA will produce a random number of proteins (p_b) and furthermore the number of mRNAs in the burst is also a random variable (m_b). The number of proteins produced in a single burst is thus a sum of a random number of random variables, which is also named as compound random variable. Correspondingly, using standard results from probability theory [84], we derive the following equations for burst size parameters (A_1 and A_2) in

terms of m_b and p_b:

$$\begin{aligned} A_1 &= \langle m_b \rangle \langle p_b \rangle \\ A_2 &= \langle m_b \rangle (\sigma_{p_b}^2 + p_b^2 - p_b) + (\sigma_{m_b}^2 + \langle m_b \rangle^2 - \langle m_b \rangle) \langle p_b \rangle^2 \\ &= \langle m_b \rangle (\sigma_{p_b}^2 - \langle p_b \rangle) + (\sigma_{m_b}^2 + \langle m_b \rangle^2) \langle p_b \rangle^2, \end{aligned} \tag{3.2}$$

where the symbols $\langle .. \rangle$ and σ represent the mean and standard deviation respectively.

Using Eq.(3.2), in combination with identification of the random variable N with the corresponding variable characterizing the protein steady-state distribution (p_s), we obtain the following expressions for the mean and variance of the protein steady-state distribution:

$$\begin{aligned} \langle p_s \rangle &= \frac{\tau_p}{\langle T \rangle} \langle m_b \rangle \langle p_b \rangle \\ \sigma_{p_s}^2 &= \langle p_s \rangle \Big(1 + A_1 (\frac{f_L(\mu_p)}{1 - f_L(\mu_p)} - \frac{1}{\mu_p \langle T \rangle} + \frac{1}{2}) - \frac{1}{2} A_1 + \frac{A_2}{2A_1}) \\ &= \langle p_s \rangle + \langle p_s \rangle^2 \frac{\mu_p \langle T \rangle}{2} \Big(K_g - 1 + \frac{\langle m_b \rangle (\sigma_{p_b}^2 - \langle p_b \rangle) + (\sigma_{m_b}^2 + \langle m_b \rangle^2) \langle p_b \rangle^2}{\langle m_b \rangle^2 \langle p_b \rangle^2} \Big) \\ &= \langle p_s \rangle + \langle p_s \rangle^2 \frac{\langle T \rangle}{2\tau_p} \Big(K_g + \frac{\sigma_{m_b}^2}{\langle m_b \rangle^2} + \frac{\sigma_{p_b}^2 / \langle p_b \rangle^2 - 1/\langle p_b \rangle}{\langle m_b \rangle} \Big) \end{aligned} \tag{3.3}$$

where

$$K_g = 2 \Big(\frac{f_L(\mu_p)}{1 - f_L(\mu_p)} - \frac{1}{\mu_p \langle T \rangle} \Big) + 1, \tag{3.4}$$

is denoted as the *gestation factor*. Correspondingly, the noise in protein

steady-state distribution can be found as

$$\frac{\sigma_{p_s}^2}{\langle p_s \rangle^2} = \frac{1}{\langle p_s \rangle} + \frac{\langle T \rangle}{2\tau_p} \times \Big(K_g + \frac{\sigma_{m_b}^2}{\langle m_b \rangle^2} + \frac{\sigma_{p_b}^2/\langle p_b \rangle^2 - 1/\langle p_b \rangle}{\langle m_b \rangle} \Big), \tag{3.5}$$

Different contributions to the noise in protein distributions are highlighted in Eq.(3.5): gestation effects, mRNA transcriptional bursting, and translational bursting from a single mRNA, which correspond to the terms K_g, $\sigma_{m_b}^2/\langle m_b \rangle^2$ and $\sigma_{p_b}^2/\langle p_b \rangle^2$, respectively. The first two terms can be modified by transcriptional regulation and the last term can be tuned by post-transcriptional regulation. It is noteworthy that each source contributes additively to the overall noise in the steady-state distribution. Moreover, while the noise due to gestation effects is independent of the degree of transcriptional bursting, the noise contribution from translation bursting is effectively reduced when transcriptional bursting occurs (corresponding to large $\langle m_b \rangle$ values).

While Eq.(3.5) is valid for general gestation effects and bursting, it is of interest to consider specific examples. First we can check if Eq.(3.5) can give the same results introduced in the preceding chapter. Consider first the basic transcription process such that $\langle m_b \rangle = 1$ and $\sigma_{m_b} = 0$. With the relationship that $\langle m_s \rangle = \langle m_b \rangle \tau_m / \langle T \rangle$ and $K_g = 1$ without gestation effect, Eq.(3.5) becomes identical to Eq.(2.21) in Chapter 2. Furthermore, consider the basic translation process that p_b follows a geometric distribution, Eq.(3.5) will give a similar result as shown in Eq.(2.25). The difference arising is because Eq.(3.5) is under the burst synthesis approximation while Eq.(2.25)

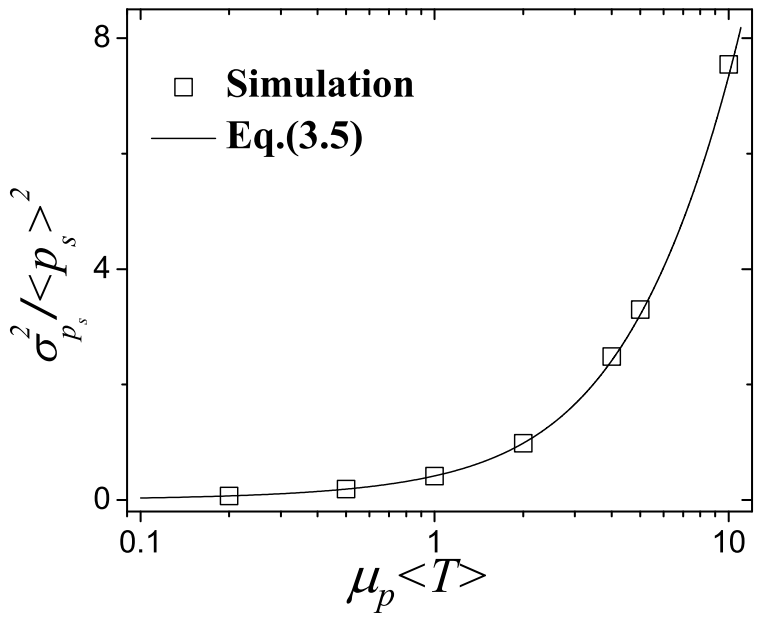

Figure 3.2: The noise *vs* $\mu_p\langle T\rangle$ from analytical expressions and stochastic simulations. The time between consecutive bursts is fixed and only 1 mRNA is produced each burst. The protein production is under post-transcriptional regulation such that $\sigma^2_{p_b} = 0.67\langle p_b\rangle^2 + \langle p_b\rangle$ [35]. The mRNA and protein lifetime are chosen as $\tau_m/\tau_p \approx 0.02$ such at the condition for the burst synthesis approximation is satisfied.

is not. For the basic translation process without gestation effect, one can go beyond the burst synthesis approximation by scaling the terms in the bracket in Eq.(3.5) with $\frac{\tau_p}{\tau_m+\tau_p}$. This term is often denoted as the time averaging factor [70].

We can also consider another case such that there is a constant delay T_d between arrival of consecutive mRNA bursts, i.e. the waiting-time distribution is $f(t) = \delta(t - T_d)$. In this case, the gestation factor is given by $K_g = 2e^{-\mu_p T_d}/(1 - e^{-\mu_p T_d}) - 2/\mu_p T_d + 1$. The corresponding expression for the noise in protein distributions Eq.(3.5), considering a general case which

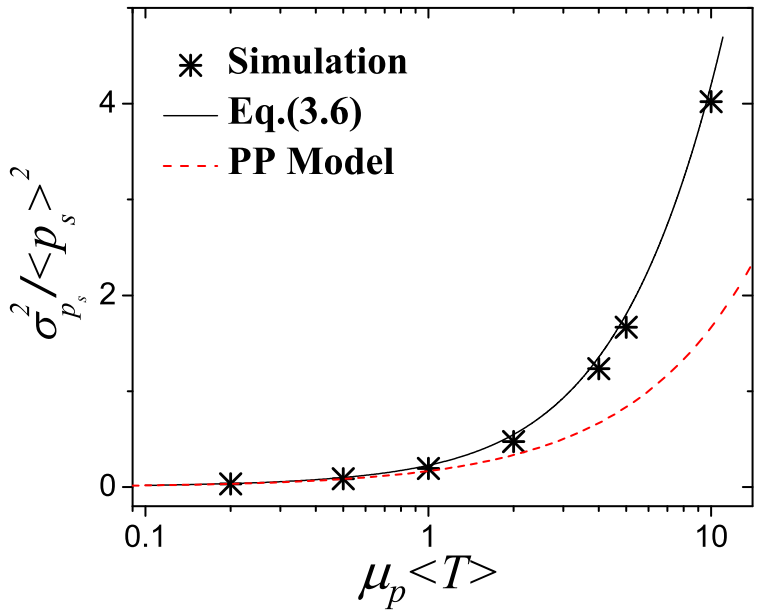

Figure 3.3: The noise *v.s.* $\mu_p\langle T\rangle$ from analytical expressions and stochastic simulations. The time interval between bursts is drawn from a Gamma distribution and the number of mRNAs created in one burst is drawn from a Poisson distribution. It includes the basic translation process such that the number of proteins created by each mRNA follows a geometric distribution. The parameters are $\langle m_b\rangle = 10$, $\sigma^2_{m_b}/\langle m_b\rangle^2 = 0.1$ and $\sigma^2_T/\langle T\rangle^2 = 0.2$. The mRNA and protein lifetime are chosen as $\tau_m/\tau_p = 0.2$ to go beyond the burst synthesis approximation. While Eq.(3.6) agrees with simulations, the result from Ref. [72] is less accurate when $\mu_p\langle T\rangle$ is large.

also includes the effects of post-transcriptional regulation [35], is in excellent agreement with results from stochastic simulations, as shown in Fig.(3.2). It is noteworthy that K_g can be nonvanishing even though the time interval between consecutive bursts is fixed (i.e. $\sigma^2_T = 0$). In contrast to previous work [72], which suggests that the contribution of gestation effects to the noise vanishes when $\sigma^2_T = 0$, our result shows that K_g can be tuned from 0 to 1 as $\mu_p T_d$ is varied.

While the results derived above are exact in the limit $\tau_m \ll \tau_p$, an exact

expression for the noise in the general case (i.e. without invoking the condition $\tau_m \ll \tau_p$ and for general gestation and bursting distributions) is difficult to obtain. However, a useful approximation can be obtained by considering that the time-averaging factor $\frac{\tau_p}{\tau_m+\tau_p}$ for the basic case without gestation is the same for general gestation and bursting distributions. We then obtain

$$\frac{\sigma^2_{p_s}}{\langle p_s \rangle^2} \approx \frac{1}{\langle p_s \rangle} + \frac{\langle T \rangle}{2\tau_p} \times \left(K_g + \frac{\sigma^2_{m_b}}{\langle m_b \rangle^2} + \frac{\sigma^2_{p_b}/\langle p_b \rangle^2 - 1/\langle p_b \rangle}{\langle m_b \rangle} \right) \times \frac{\tau_p}{\tau_m + \tau_p}. \quad (3.6)$$

It is instructive to compare Eq.(3.6) with the result derived in previous work for the PP model [72]. The PP model assumes a specific protein production reaction scheme such that each mRNA gives rise to a burst of proteins drawn from a geometric distribution. Considering Eq.(3.6) for the specific case of a geometric distribution (i.e. $\sigma^2_{p_b} = \langle p_b \rangle^2 + \langle p_b \rangle$), we note that Eq.(3.6) is identical to the previous result [72] apart from the terms corresponding to the gestation factor K_g, as discussed below.

The correspondence between Eq.(3.6) and the previous result can be further analyzed as follows. The Laplace transform, $f_L(\mu_p)$, can be written as:

$$f_L(\mu_p) = 1 - \mu_p \langle T \rangle + \frac{\mu_p^2 \langle T^2 \rangle}{2} + O(\mu_p^3 \langle T^3 \rangle). \quad (3.7)$$

Assuming that the higher order terms in Eq.(3.7) can be ignored (e.g. $\mu_p \langle T \rangle$ is small and $\langle T^n \rangle$ scales as the n^{th} power of $\langle T \rangle$ or less), substituting for $f_L(\mu_p)$ in Eq.(3.4) shows that K_g can be approximated by $K_g \approx \sigma^2_T / \langle T \rangle^2$ which corresponds to the previous result. Since the parameter $1/(\mu_p \langle T \rangle)$

measures the mean number of bursts occuring during the protein lifetime, this indicates that the previous result [72] is valid for the case of frequent bursting during a protein lifetime, and breaks down when bursts occur over larger time intervals, as demonstrated in Fig.(3.3).

3.2.2 Effect of Senescence in Protein Decaying

We now consider case B, which corresponds to arbitrary distributions for bursting and senescence effects along with exponential waiting-time distributions for burst arrival. Previous work [72] has focused on the case that the mRNA decay-time distribution is a Gamma distribution. Since different regulation schemes can impact degradation, it is of interest to consider the effects of more general waiting-time distributions for protein decay. For the general problem, we take the waiting-time for protein degradation to be drawn from an arbitrary distribution characterized by p.d.f $h(t)$ and c.d.f $H(t)$. The waiting-time between consecutive bursts is characterized by an exponential distribution with $f(t) = \lambda e^{-k_m t}$. The corresponding system, following the mapping to queueing theory, is the $M^X/G/\infty$ queue. The steady-state mean and variance of N for this queue has been obtained in previous work [53]:

$$\begin{aligned} E[N] &= k_m A_1 \int_0^\infty [1 - H(t)]dt \\ Var[N] &= E[N] + k_m A_2 \int_0^\infty [1 - H(t)]^2 dt. \end{aligned} \tag{3.8}$$

By taking Eq.(3.2) and the relation $\langle T\rangle = 1/k_m$ into account, the mean and the variance for arbitrary senescence and bursting distribution can be derived as:

$$\begin{aligned}\langle p_s\rangle &= \frac{A_1}{\langle T\rangle}\int_0^\infty [1-H(t)]dt = \frac{\tau_p}{\langle T\rangle}\langle m_b\rangle\langle p_b\rangle \\ \sigma^2_{p_s} &= \langle p_s\rangle + \langle p_s\rangle^2\frac{\langle T\rangle}{2\tau_p}\Big(1+\frac{\sigma^2_{m_b}}{\langle m_b\rangle^2}+\frac{\sigma^2_{p_b}/\langle p_b\rangle^2 - 1/\langle p_b\rangle}{\langle m_b\rangle}\Big)\times K_s \quad (3.9)\end{aligned}$$

where

$$K_s = \frac{2\int_0^\infty [1-H(t)]^2 dt}{\tau_p} = 2 - \frac{2\int_0^\infty H(t)[1-H(t)]dt}{\tau_p}, \quad (3.10)$$

is denoted as the *senescence factor*. The noise in protein steady-state can also be found as

$$\frac{\sigma^2_{p_s}}{\langle p_s\rangle^2} = \frac{1}{\langle p_s\rangle} + \frac{\langle T\rangle}{2\tau_p}\times\Big(1+\sigma^2_{m_b}/\langle m_b\rangle^2 + \frac{\sigma^2_{p_b}/\langle p_b\rangle^2 - 1/\langle p_b\rangle}{\langle m_b\rangle}\Big)\times K_s, \quad (3.11)$$

It is noteworthy that Eq.(3.11) and Eq.(3.5) have multiple terms in common. The terms characterizing the noise from transcriptional and translational bursting remain unchanged. For the case without senescence or gestation effects, i.e. $K_g = K_s = 1$, Eq.(3.11) and Eq.(3.5) become identical. However, unlike the gestation factor that contributes to the total noise *additively*, the senescence factor serves as a *scaling* factor for the total noise. While there is no obvious upper limit on the value of K_g, the upper bound for K_s is 2 as is evident from Eq.(3.10). In general, the K_s value is lower for functions $h(t)$

that vary slowly with t, and higher for narrowly peaked distribution $h(t)$. When $h(t)$ becomes a delta funtion, K_s reaches its maximum value.

3.3 Summary

The general results derived in this chapter will serve as useful inputs for the analysis and interpretation of diverse experimental studies of gene expression. Some examples are: 1) Recent experiments on single-cell studies of HIV-1 viral infections have focused on the frequency and degree of transcriptional bursting [89]. For such studies, the derived results can be used to relate measurements of inter-arrival waiting-time distributions and burst distributions to the noise in protein distributions. 2) Experimental data and computational models of the cell-cycle in yeast indicate that modeling the basic processes of gene expression as Poisson processes gives rise to unrealistically large noise in protein distributions [40], thereby suggesting that regulatory schemes which change distributions with reduced noise are employed by the cell. The analytical expressions derived highlight different contributions to noise and can thus provide insight into how different regulatory schemes can lead to noise reduction. 3) More generally, the results derived can be used in the analysis of inverse problems, i.e. using experimental measurements of intrinsic noise to determine parameters of the underlying kinetic models. Such efforts, in turn, can lead to further insights into cellular factors that impact gene regulation, based on experimental observations of noise in gene expression.

In summary, we have analyzed the noise in protein distributions for general stochastic models of gene expression. The present work extends previous analysis by deriving analytical results for the noise in protein distributions for arbitrary gestation, senescence and bursting mechanisms. The expressions obtained provide insight into how different sources contribute to the noise in protein levels which can lead to phenotypic heterogeneity in isogenic populations. The results derived will thus serve as useful inputs for the analysis and interpretation of experiments probing stochastic gene expression and its phenotypic consequences.

At a broader level, though powerful analytical approaches have been developed in modeling stochastic processes, the application of some of these tools to cellular processes has been limited to date. Our work is one of the first studies that applied queueing theory in analyzing the noise in protein distributions for stochastic gene expression models. This work demonstrates the benefits of developing a mapping between models of stochastic gene expression and queueing systems which has potential applications for research in both fields. The extensive analytical approaches and tools developed in queueing theory can now be employed to analyze stochastic processes in gene expression. It is also anticipated that future analysis of regulatory mechanisms for gene expression will lead to new problems and challenges for queueing theory.

Chapter 4

Post-transcriptional Regulation of Noise I

The intrinsic stochasticity of biochemical reactions involved in gene expression can lead to large variability of protein levels across a clonal population of cells. In order to adapt and respond to changing environments, cellular systems must employ regulating mechanisms to control the variability (or "noise") in gene expression. An important question that arises is: what are the roles of different ways of regulations in controlling the noise in protein steady state distributions?

The results in preceding chapters have brought some insights on how transcriptional, post-transcriptional and post-translational regulation control the noise in the gene expression process. In the following two chapters, we will focus on models of post-transcriptional regulation. Two broad classes of reaction schemes are introduced: (I) for global regulators present in large

amounts for which fluctuations of the regulator can be neglected and (II) for regulators that undergo coupled degradation for which fluctuations of the regulator cannot be neglected. We will discuss the case (I) in this chapter and case (II) in the following chapter.

4.1 Introduction

Given the fact that the gene expression is inherently stochastic, the need to regulate this intrinsic variability in protein level places important constraints on cellular regulatory pathways; in particular those that bring about global changes in gene expression. In such pathways, it has been shown that a central component of several global regulatory networks is post-transcriptional control by regulatory proteins and by small RNAs (sRNAs) in bacteria and MicroRNAs (miRNAs) in higher organisms. Recent research points towards an increasingly important role for this mode of regulation in fine-tuning the noise in gene expression and in regulating important cellular processes. In bacteria, this trend is highlighted by the continuing discoveries of sRNAs which play central roles in global regulatory pathways [1, 24, 27]. In higher organisms, microRNAs are known to play key roles in the regulation of critical processes such as development, stem cell pluripotency and cancer [6, 29, 32].

Although noise in gene expression can have deleterious effects in some cases and thus needs to be limited, in other cases such noise is utilized and in-

deed required by the cell e.g. for processes leading to cell-fate determination [56, 30]. Furthermore, it has been argued that noise in gene expression could be advantageous under conditions of high stress, since variability in a population provides a bet-hedging strategy that can enable survival [19, 11]. Regulation of the noise in gene expression is thus essential for the proper functioning of several cellular processes. Since sRNAs regulate critical cellular processes, understanding their role in fine-tuning the noise in gene expression is of fundamental importance [30].

In this chapter, we develop a framework for modeling post-transcriptional regulation with regulators present in high abundance. The analytical results obtained provide insight into how different mechanisms of post-transcriptional regulation modulate noise in protein distributions. The results also show how to quantify the degree of transcriptional bursting based on observations of regulated protein burst distributions. Furthermore, the analytical approach can also be applied in studying the model that the mRNA target is regulated by multiple regulators.

4.2 Stochastic Modeling

4.2.1 Model

The post-transcriptional regulation of gene expression is a process through which regulators can bind to mRNA and control protein production by al-

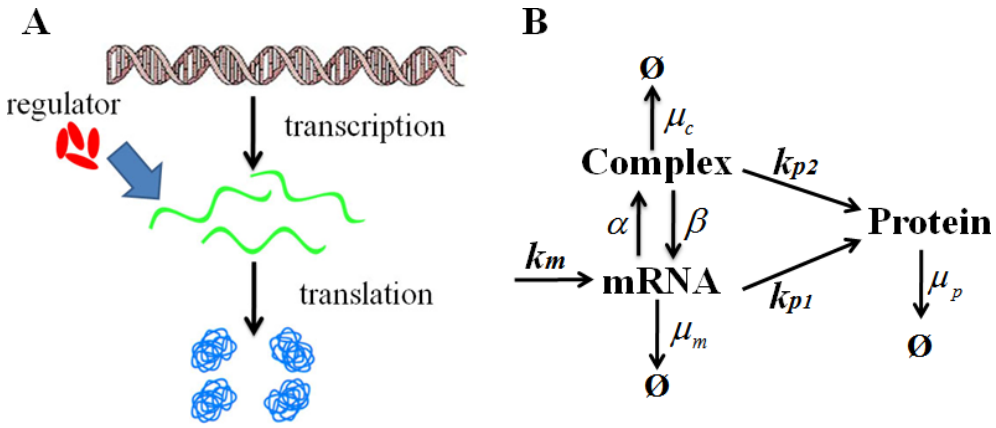

Figure 4.1: (A) Figure of post-transcriptional regulation. (B) Kinetic scheme for regulation of protein production by a mRNA-binding protein (denoted as Complex).$k_{p1}(k_{p2})$ are the translation rates in the free (bound) states and $\mu_m(\mu_c)$ are the corresponding decay rates.

tering mRNA stability or by regulating translational efficiency [96], as shown in Fig.(4.1A). Since many regulator (*e.g.* regulatory sRNAs) are known to repress gene expression, most previous models have focused on regulation by irreversible stoichiometric degradation [50, 66, 62, 21]. However, sRNAs can affect not only mRNA degradation rates but also protein production rates and the corresponding biochemical reactions are, in general, reversible [18, 96]. Furthermore, not all regulators repress gene expression; there are sRNAs which are known to activate gene expression and even some which can switch from activating to repressing in response to cellular signals [18, 94]. To quantify the corresponding effects on stochastic gene expression, a general model which includes the different mechanisms of sRNA-based regulation needs to be analyzed.

In the limit that global regulators are present in large numbers, cellular con-

centrations of the regulator are not significantly affected by binding to target mRNAs. For example, during exponential growth phase a cell typically has $\sim$ 20,000 copies of regulatory protein CsrA. Hence it is a good approximation to assume that fluctuations in regulator concentration can be neglected and furthermore that the binding/unbinding rates can be taken to be constant. Following that, we propose a general reaction scheme in analyzing the process of post-transcriptional regulation as shown in Fig.(4.1B). The regulator binds mRNA to form a complex with rate α; the dissociation rate for the complex is β. The parameters k_{p1} and k_{p2} are the rates of protein production from the mRNA in free and bound states and μ_m and μ_c are the corresponding decay rates.

With the rates of mRNA creation and protein degradation, we can write down the master equation for the protein steady state distribution following the procedures introduced in Chapter 2. The master equation is solvable in principle, but complicated in practice. However, as discussed in Chapter 2, the gene expression can be approximated by the processes of protein creation and decaying only in the limit that protein life time is much longer than that of mRNA. In this burst synthesis approximation, the protein steady state distribution can be derived using Eq.(2.16) once the protein burst size distribution is known. As also discussed in Chapter 3, for the situation that transcription and protein decay processes are not Poisson processes, we can still find the noise in protein steady state distribution once the noise in protein burst size distribution (equivalently the mean and variance of p_b) is given.

For these reasons, we wish to find the protein burst size distribution under the regulation. The results derived can be used to derive corresponding analytical expressions for quantities of interest in steady state protein level over a population of cells.

4.2.2 Method

We denote the protein burst size distribution by $P_{pb}(n)$ which corresponds to the number of proteins translated from a single mRNA before it decays. During this process, the mRNA can exist in two states: either free or bound in a complex with the post-transcriptional regulator. Correspondingly, we define the functions $f_1(n,t)$ and $f_2(n,t)$ (generalizing the approach outlined in [5]) which denote the probabilities of finding the mRNA in free and bound states respectively at time t, having produced a burst of n proteins. The initial condition corresponds to creation of the mRNA in its free state at $t = 0$, *i.e.* $f_1(0,0) = 1$ and $f_2(n,0) = 0$. Now, the burst distribution $P_{pb}(n)$ can be obtained from $f_1(n,t)$ and $f_2(n,t)$ as

$$P_{pb}(n) = \int_0^\infty f_1(n,t)\mu_m \, dt + \int_0^\infty f_2(n,t)\mu_c \, dt, \tag{4.1}$$

which is found by the probability of having n proteins by time t and the mRNA / complex degrades at the next time interval t to $t + \Delta t$. Furthermore, the time evolution of $f_1(n,t)$ and $f_2(n,t)$ is determined by the following

Master equations:

$$\begin{aligned}\frac{\partial f_1(n,t)}{\partial t} &= k_{p1}(f_1(n-1,t)-f_1(n,t))-(\mu_m+\alpha)f_1(n,t)+\beta f_2(n,t)\\ \frac{\partial f_2(n,t)}{\partial t} &= k_{p2}(f_2(n-1,t)-f_2(n,t))-(\mu_c+\beta)f_2(n,t)+\alpha f_1(n,t)\end{aligned} \quad (4.2)$$

The above equations can be analyzed further using a combination of generating functions and Laplace transforms. Specifically, defining generating function of the protein burst size distribution $G_{pb}(z) = \sum_n z^n P_{pb}(n)$ and $F_{1,2}(z,s) = \sum_n z^n \int_0^\infty e^{-st} f_{1,2}(n,t)\,dt$, we obtain

$$\begin{aligned}G_{pb}(z) &= \lim_{s\to 0}\Big(\mu_m F_1(z,s)+\mu_c F_2(z,s)\Big)\\ sF_1(z,s)-1 &= \Big(k_{p1}(z-1)-(\mu_m+\alpha)\Big)F_1(z,s)+\beta F_2(z,s)\\ sF_2(z,s) &= \Big(k_{p2}(z-1)-(\mu_c+\beta)\Big)F_2(z,s)+\alpha F_1(z,s)\end{aligned} \quad (4.3)$$

4.2.3 Results

Evaluating these equations then leads to the exact expression for $G_{pb}(z)$, which can be written as

$$G_{pb}(z) = X\frac{1-S_1}{z-S_1}+(1-X)\frac{1-S_2}{z-S_2}, \quad (4.4)$$

where

$$
\begin{aligned}
X &= \frac{\sqrt{\Delta} - \big(k_{p1}(\beta+\mu_c) + k_{p2}(\alpha-\mu_m)\big)}{2\sqrt{\Delta}} \\
S_{1\,2} &= 1 + \frac{k_{p1}(\beta+\mu_c) + k_{p2}(\alpha+\mu_m) \pm \sqrt{\Delta}}{2k_{p1}k_{p2}} \\
\Delta &= \big(k_{p1}(\beta+\mu_c) - k_{p2}(\alpha+\mu_m)\big)^2 + 4\alpha\beta k_{p1}k_{p2}.
\end{aligned} \tag{4.5}
$$

The above expression indicates that the number of proteins produced in one burst can be expressed as a weighted sum of two random variables, each of which corresponds to the geometric distribution. While the complete expression for $P_{pb}(n)$ can thus be derived from the results obtained, in some cases the primary interest is in derived quantities charaterizing the noise in protein distributions. For the protein burst distribution $P_{pb}(n)$ derived above, both the mean and the coefficient of variance can be readily obtained from the generating function as

$$
\begin{aligned}
\langle p_b \rangle &= \frac{k_{p1}(\mu_c+\beta) + k_{p2}\alpha}{\mu_m(\mu_c+\beta) + \mu_c\alpha} \\
\frac{\sigma_{p_b}^2}{\langle p_b \rangle^2} &= 1 + \frac{1}{\langle p_b \rangle} + \frac{2\alpha k_{p2}(k_{p2}\mu_m - k_{p1}\mu_c)}{(\alpha k_{p2} + k_{p1}\beta + k_{p1}\mu_c)^2},
\end{aligned} \tag{4.6}
$$

The above expression is tested and in excellent agreement with result from stochastic simulations using the Gillespie algorithm.

It is noteworthy that Eq. (4.4) is valid for the most general choice of parameters. To gain additional insight, let us consider specific parameter choices of

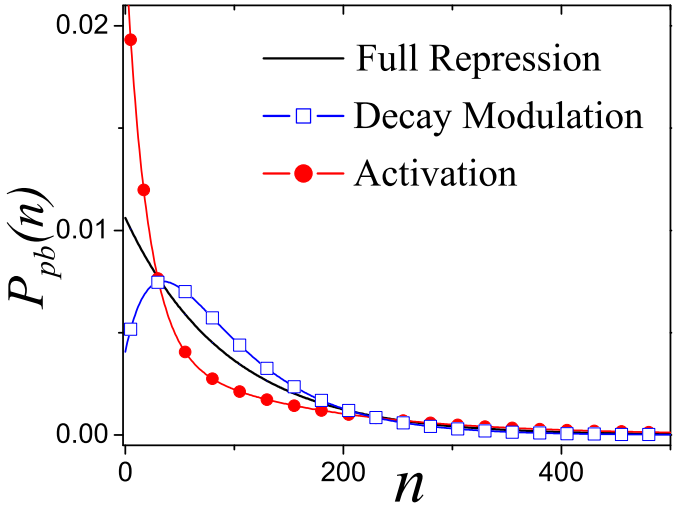

Figure 4.2: The protein burst distribution $P_{pb}(n)$ for full repression, decay modulation and activation from Eq.(4.6)). In all three cases, the mean of the protein burst distribution is kept the same. The burst distribution for full repression is identical to the geometric distribution with the same mean, whereas $P_{pb}(n)$ for decay modulation and activation deviate significantly from the geometric distribution. Parameters for decay modulation and activation are chosen such that $\frac{\alpha}{\mu_m} = 5$, $\frac{\beta}{\mu_m} = 1$, $\frac{\mu_c}{\mu_m} = 5$, $p_m = 1$ and $\frac{\alpha}{\mu_m} = \frac{\beta}{\mu_m} = 1$, $\frac{kp_2}{kp_1} = 4$, $p_m = 1$ respectively.

interest. For example, taking the limit $\alpha \to 0$ corresponds to the unregulated protein burst distribution. In this case, we obtain

$$G_{pb}(z) = \frac{\mu_m}{\mu_m + k_{p1}(1-z)}, \tag{4.7}$$

which corresponds to the generating function of a geometric distribution in agreement with previous studies [99, 20, 31, 86] and the discussion in Chapter 2.

Of greater interest is the effect of different modes of regulation. While a generally accepted model is that regulator binding prevents ribosome access

(i.e. $k_{p2} = 0$), recent studies have shown that small RNAs can also repress gene expression by binding in the coding region significantly downstream of the ribosome binding site [73]. In the latter case, regulator binding is not expected to affect the translation rate, but instead alters the mRNA decay rate. To explore the effects of these different regulatory mechanisms on the noise in protein distributions, we consider two special cases for the general results derived above: 1) full repression ($k_{p2} = 0$) and 2) decay modulation ($k_{p2} = k_{p1}$, $u_m < u_c$).

For full repression (taking the limit $k_{p2} \to 0$), the generating function is given by

$$G_{pb}(z) = \frac{\mu_m + \frac{\mu_c \alpha}{\mu_c + \beta}}{\mu_m + \frac{\mu_c \alpha}{\mu_c + \beta} - k_{p1}(1 - z)}. \tag{4.8}$$

The result is identical to Eq.(4.7) provided that the mRNA degradation rate is rescaled from μ_m to $\mu_m + \frac{\mu_c \alpha}{\mu_c + \beta}$. Thus the protein burst distribution remains a geometric distribution but with a reduced mean due to lowering of the effective mRNA lifetime. This implies that regulation by full repression results in a protein burst distribution that is identical to that of an unregulated burst distribution with the same mean.

On the other hand for regulation by decay modulation, the burst distribution shows deviations from a geometric distribution, as displayed in Fig.(4.2). To analyze this further, let us focus on the noise $\sigma^2_{p_b}/\langle p_b \rangle^2$ in Eq.(4.6) which, for

decay modulation, is given by

$$\begin{aligned}\frac{\sigma_{p_b}^2}{\langle p_b \rangle^2} &= 1 + \frac{1}{\langle p_b \rangle} + \frac{2\alpha k_{p1}(1-\theta_1)}{(\alpha+\beta+\theta_1\mu_m)^2} \\ &= 1 + \frac{1}{\langle p_b \rangle} + Q, \end{aligned} \tag{4.9}$$

where $\theta_1 = \mu_c/\mu_m > 1$ and the term Q quantifies the deviation from the geometric distribution ($Q = 0$ for a geometric distribution). Thus for regulation by decay modulation, the noise strength can be tuned by the parameter θ_1 resulting in a burst distribution with reduced variance when compared with an unregulated burst distribution with the same mean. Eq. (4.9) indicates that this reduction can be significant since the maximum magnitude for Q is 0.5. Such a narrowing of the variance relative to the mean has been previously proposed as a potential function for small RNAs with important implications for canalization of gene expression during development [29].

The previous results for repression mechanisms can be contrasted with the effect of post-transcriptional activation of gene expression. The burst distribution for activation also shows significant deviations from a geometric distribution with the same mean, as displayed in Fig.(4.2). For activation due to increased protein production (with $\mu_c = \mu_m$), the deviation Q is given by:

$$Q = \frac{2\alpha\theta_2 p_m \mu_m(\theta_2 - 1)}{(\alpha\theta_2 + \beta + \mu_m)^2}. \tag{4.10}$$

where $\theta_2 = k_{p2}/k_{p1}$. As $\theta_2 > 1$ for activation, the noise will be greater than

that of an unregulated burst distribution with the same mean. The value of Q, depending on the choice of θ_2 and α, can be made arbitrarily large. Our results thus indicate that activation of gene expression by small RNAs can potentially lead to a large variance in protein distribution, which in turn can give rise to phenotypic heterogeneity that is often beneficial for the organism [19].

4.2.4 Application

The above analysis can be extended to study the problem wherein transcriptional bursting is considered such that multiple mRNAs can be produced in a single burst. In this case, the probability of having m mRNAs in one transcriptional burst, $P_{mb}(m)$, is given by a geometric distribution, conditional on the production of at least 1 mRNA [31]

$$P_{mb}(m) = (1 - p_m)^{m-1} p_m. \tag{4.11}$$

Eq.(4.11) serves as a general formula for the mRNA burst distribution. The case $p_m = 1$ correspond to a single mRNA produced in every transcription event; whereas if $p_m < 1$, the mean number of mRNAs produced per burst $\langle m_b \rangle = 1/p_m$ is greater than 1. In general, each mRNA will produce a random number of proteins drawn from the distribution $P_{pb}(n)$ (the corresponding generating function $G_{pb}(z)$ is given by Eq.(4.4)) and furthermore the number of mRNAs in the burst is also a random variable defined by the distribution

Eq.(4.11). The total number of protein produced is thus a compound random variable [84]. The corresponding generating function $G'_{pb}(z)$ is given by

$$G'_{pb}(z) = \frac{G_{pb}(z)p_m}{1 - G_{pb}(z)(1 - p_m)}. \tag{4.12}$$

From Eq.(4.12), we can identify the conditions such that the value of p_m can be determined. If $G'_{pb}(z)$ corresponds to the generating function of the geometric distribution, it is completely determined by its mean value. For example, in the case without regulation, the mean protein burst size is given by $\langle p'_b \rangle = \frac{1}{p_m}(\frac{k_{p1}}{\mu_m})$. Since there is effectively one measurable quantity ($\langle p'_b \rangle$) for the burst distribution, p_m cannot be determined given that $\frac{k_{p1}}{\mu_m}$ is not known [31].

However, with post-transcriptional regulation which gives rise to burst size distribution different from geometric, it is possible to distinguish the degree of transcriptional bursting. Based on Eq.(4.12), the Fano Factor of the protein burst size distribution under decay modulation can be found as

$$\begin{aligned} \frac{\sigma'^2_{p_b}}{\langle p'_b \rangle} &= 1 + \langle p'_b \rangle + \frac{2\alpha k_{p1}(1-\theta)}{(\alpha + \beta + \theta\mu_m)(\beta + \theta(\alpha + \mu_m))} \\ &= 1 + \langle p'_b \rangle + D, \end{aligned} \tag{4.13}$$

where the term D quantifies the deviation from the geometric distribution ($D = 0$ for a geometric distribution). It is of interest to note that D depends on $\frac{k_{p1}}{\mu_m}$ but is independent of p_m. Thus, measurements of D and $\langle p_b \rangle'$ can, in

principle, be used to determine both $\frac{k_{p1}}{\mu_m}$ and p_m and thereby to discriminate if transcriptional bursting exists.

The argument above provides a means of determining p_m provided the interaction parameters such as α, β and θ_1 are known. In general, these parameters are not known, however for regulators such that the dissociation rate $\frac{\beta}{\mu_m} \to 0$, the following protocol can be used to determine p_m: (i) Obtain the mean protein burst levels without regulation, denoted by $\langle p'_{b0} \rangle$. (ii) Choose a certain regulator concentration. Obtain the mean protein burst level $\langle p'_{b1} \rangle$ and the corresponding variance. Determine the deviation from a geometric distribution as defined in Eq.(4.13), which is denoted by D_1 and let $n_1 = \langle p'_{b1} \rangle / \langle p'_{b0} \rangle$. (iii) Change the concentration of the regulator, which effectively changes the regulator binding rate α. Repeat step (ii) and obtain the corresponding quantities denoted by D_2 and $n_2 = \langle p'_{b2} \rangle / \langle p'_{b0} \rangle$. Given the five quantities $n_{0,1,2}$ and $D_{1,2}$, the mean transcriptional burst size $\langle m_b \rangle (= 1/p_m)$ is given by:

$$\begin{aligned} \langle m_b \rangle &= -2n_0 n_1 n_2 \frac{D_1(1-n_2) - D_2(1-n_1)}{D_1 n_1 (1-n_2) - D_2 n_2 (1-n_1)} \\ &\times \frac{(1-n_1)(1-n_2)(n_1-n_2)}{D_1 n_1 (1-n_2)^2 - D_2 n_2 (1-n_1)^2} \end{aligned} \tag{4.14}$$

Using stochastic simulations, we have verified that the above expression accurately predicts the degree of transcriptional bursting. It should be noted that experimental approaches have been developed recently for direct measurements of mRNA burst distributions [77, 78] and it would be informative to compare results from these direct approaches with estimates from the

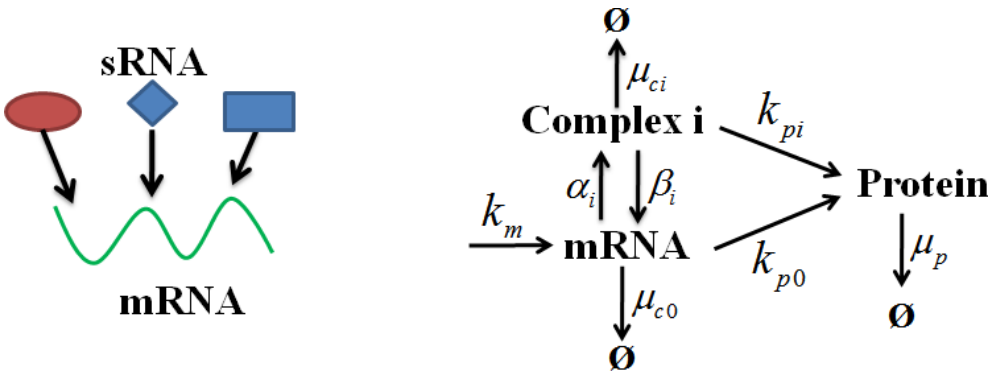

Figure 4.3: Schematic illustration of regulation of gene expression by multiple sRNAs. In the full reaction scheme, there are N different regulators and the kinetic scheme is shown for the i^{th} sRNA regulator. The association and dissociation rates for binding to the mRNA are denoted by α_i and β_i respectively. Association results in a complex which produces proteins with rate k_{p_i} and is degraded with rate μ_{c_i}. Note that only one regulator can bind to mRNA so the transition from one complex to another requires the mRNA returning to its unbound state before forming a new complex.

above protocol.

4.3 Regulation by Multiple Regulators

The above work can be generalized to analyze the case of multiple sRNAs regulating a single mRNA target. We begin by considering protein production from a single mRNA regulated by N independent sRNAs [2]. We assume that the mRNA can only be bound by one regulator at a time. The corresponding reaction scheme is shown in Fig.(4.3). The mRNA has $N+1$ possible states due to the regulation, with the states $i = 1, .., N$ denoting mRNA bound to the i^{th} sRNA regulator to form complex i. For notational simplicity, we denote the unbound mRNA state as complex 0. An unbound

mRNA forms complex i with rate α_i and the complex can either dissociate with a rate β_i, decay with a rate μ_{c_i}, or initiate protein production with a rate k_{p_i}. We assume the sRNA regulators are present in large amounts such that fluctuations in their concentration can be ignored; correspondingly the rates α_i are taken to be constant.

The protein burst size distribution from a single mRNA (interacting with N sRNAs) is denoted as $P_{pb,N}(n)$. We further define the function $f_i(n,t)$ which denotes the probability that n proteins have been produced and the mRNA is in state i at time t. Similar to the previous deduction, the protein burst distribution, $P_{pb,N}(n)$, can be obtained from

$$P_{pb,N}(n) = \sum_{i=0}^{N} \int_0^\infty f_i(n,t)\mu_{c_i} dt. \tag{4.15}$$

The time-evolution of the probabilities $f_i(n,t)$ is governed by the Master equation

$$\begin{aligned}
\frac{\partial f_0(n,t)}{\partial t} &= k_{p_0}(f_0(n-1,t) - f_0(n,t)) - (\mu_{c_0} + \sum_{i=1}^{N} \alpha_i) f_0(n,t) \\
&+ \sum_{i=1}^{N} \beta_i f_i(n,t) \\
\frac{\partial f_i(n,t)}{\partial t} &= k_{p_i}(f_i(n-1,t) - f_i(n,t)) - (\mu_{c_i} + \beta_i) f_i(n,t) \\
&+ \alpha_i f_0(n,t).
\end{aligned} \tag{4.16}$$

The initial condition corresponds to creation of a single unbound mRNA and

no proteins in the system at time $t = 0$, i.e. $f_0(0, 0) = 1$.

The procedure outlined in preceeding section can now be applied to obtain the generating function of protein burst distribution. Define $G_{pb,N}(z) = \sum_n z^n P_{pb,N}(n)$ and $F_i(z, s) = \sum_n z^n \int_0^\infty e^{-st} f_i(n, t)\, dt$, we have

$$G_{pb,N}(z) = \lim_{s\to 0} \sum_{i=0}^{N} \Big(\mu_{c_i} F_i(z, s) \Big) \tag{4.17}$$

To present the results in a compact form, it is convenient to define the dimensionless variables $\xi_i = \frac{k_{p_i}}{\beta_i + \mu_{c_i}}$ and $\omega_i = \frac{\alpha_i}{\beta_i + \mu_{c_i}} \left(\frac{\mu_{c_i}}{\mu_{c_0}} \right)$ for $i > 0$ and $n_i = \frac{k_{p_i}}{\mu_{c_i}}$ for $i \geq 0$. Now, by setting $\omega_0 = 1$ and $\xi_0 = 0$ we further define the 'weight functions' $\omega_i(z) = \omega_i \frac{1}{1+\xi_i(1-z)}$. Note that $\frac{1}{1+\xi_i(1-z)}$ is the generating function of a geometric distribution with mean ξ_i.

Using the above definitions, we obtain the following exact expression for the generating function of the protein burst distribution

$$G_{pb,N}(z) = \frac{\sum_{i=0}^{N} \omega_i(z)}{\sum_{i=0}^{N} \omega_i(z) + \sum_{i=0}^{N} \omega_i(z) n_i (1 - z)} \tag{4.18}$$

For $N = 0$, i.e. the unregulated case, the generating function reduces to

$$G_{pb,0}(z) = \frac{\mu_m}{\mu_m + k_p(1 - z)} \tag{4.19}$$

in agreement with previous discussion.

For the general case, the generating function can be recast in a form that

shows that the protein burst distribution is a mixture of $N+1$ geometric distributions [35]. However, the corresponding expression, even for the case of $N = 2$, is too complex to be reproduced here. On the other hand, using Eq. (4.18), compact analytic expressions for the mean, $\langle p_{b,N} \rangle$, and squared coefficient of variance, $\sigma^2_{p_{b,N}} / \langle p_{b,N} \rangle^2$ can be derived. The mean (scaled by the unregulated mean) is given by

$$\frac{\langle p_{b,N} \rangle}{\langle p_{b,0} \rangle} = 1 + F_N \tag{4.20}$$

and the noise strength (squared coefficient of variance) is given by

$$\frac{\sigma^2_{p_{b,N}}}{\langle p_{b,N} \rangle^2} = 1 + \frac{1}{\langle p_{b,N} \rangle} + Q_N \tag{4.21}$$

where

$$\begin{aligned} F_N &= \frac{\sum_{i=0}^{N} \omega_i \left(n_i - n_0 \right)}{\sum_{i=0}^{N} \omega_i n_0} \\ Q_N &= \frac{\sum_{i,j=0}^{N} \omega_i \omega_j (\xi_i - \xi_j)(n_i - n_j)}{\left(\sum_{i=0}^{N} \omega_i n_i \right)^2} \end{aligned}$$

Note that the signs of F_N and Q_N characterize the impact of the sRNAs on the regulated protein distribution. Specifically, the unregulated case has mean $\langle p_{b,0} \rangle$; thus $F_N < 0$ corresponds to repression whereas $F_N > 0$ corresponds to activation. Similarly, an unregulated protein burst distribution with mean

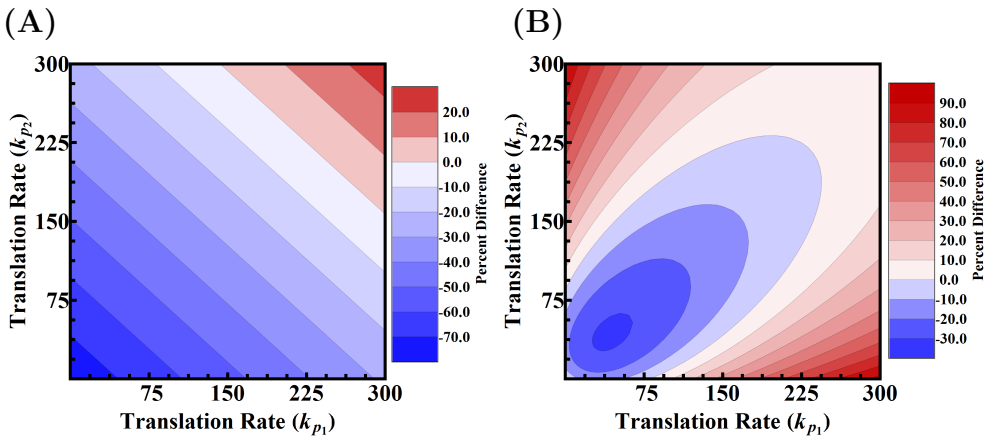

Figure 4.4: Contour plots for the percent change in the mean and noise strength of a two regulator system from the corresponding unregulated values as a function of k_{p_1} and k_{p_2}. (a) Mean: Plot of $f(k_{p_1}, k_{p_2}) = \frac{\langle p_{b,2}\rangle - \langle p_{b,0}\rangle}{\langle p_{b,N}\rangle} \cdot 100\%$. Note that along the contour $f(k_{p_1}, k_{p_2}) = -20\%$ the noise strength changes from less than -5% to over 70%. (b) Noise Strength: Plot of $g(k_{p_1}, k_{p_2}) = \frac{Q_2}{1+1/\langle p_{b,2}\rangle} \cdot 100\%$. Note that $g(k_{p_1}, k_{p_2})$ contains contours that sweep out a large portion of the plotted (k_{p_1}, k_{p_2}) state space. By proportionally changing the k_p values corresponding to the two regulators, the noise strength can be varied while maintaining the same mean value. The parameters used were $k_{p_0} = 50$, $\mu_{c_0} = 1$, $\mu_{c_1} = 4.5$, $\mu_{c_2} = 4.5$, $\beta_1 = 1$, $\beta_2 = 0.5$, $\alpha_1 = 2$ and $\alpha_2 = 2$.

$\langle p_{b,N}\rangle$ has a squared coefficient of variance $1+1/\langle p_{b,N}\rangle$; thus, when $Q_N < 0$ we have noise reduction whereas $Q_N > 0$ corresponds to increased noise strength (relative to an unregulated burst distribution with the same mean).

We now focus on using Eq. (4.20) and Eq. (4.21), to elucidate interesting features for the case of regulation by a single sRNA, i.e. $N = 1$. Note that all of the variables in the expressions for the mean and noise strength are always positive (or zero) except for the term $(n_1 - n_0)$. Thus, the sign of F_1 and Q_1 is determined completely by $\Delta_{10} = n_1 - n_0$. When $\Delta_{10} > 0$ both the mean and the noise strength are higher than their unregulated values. Similarly, when $\Delta_{10} < 0$ both the mean and the noise strength are lower

than the corresponding unregulated values (except for the case $\xi_i = 0$ for which the noise strength is identical to an unregulated distribution with the same mean). In either case, we note that, for a single sRNA regulator, there is a coupling between the mean and noise strength of the regulated burst distribution such that both cannot be tuned independently, e.g. a decrease in the mean cannot be associated with an increase in the noise strength.

In contrast to the case of regulation by a single sRNA, in the case of regulation by multiple sRNAs, the mean and noise of the protein distribution can be tuned independently. The deviation of the mean from its unregulated value depends solely upon terms of the form $\Delta_{i0} = n_i - n_0$. On the other hand, considering the general form of the noise strength for $N > 1$, we have terms of the form $\tilde{\Delta}_{ij} = (\xi_i - \xi_j)(n_i - n_j)$ that contribute to the deviation from the corresponding unregulated value. Thus, for appropriately chosen parameters, two sRNAs can be used to tune both the mean and variance of the regulated protein distribution as discussed below.

Consider the case of regulation by 2 sRNAs that are maintained at some fixed cellular concentrations. A new mRNA target for these sRNAs can arise from the evolution of appropriate sRNA binding sites on the mRNA sequence. For the new target, we assume that the parameters k_{p_1} and k_{p_2} can be tuned based on changes in the sequence and location of the sRNA binding sites. The corresponding variation in the mean and noise strength is shown in Fig.(4.4). Note that by maintaining a linear relationship between k_{p_1} and k_{p_2}, the mean

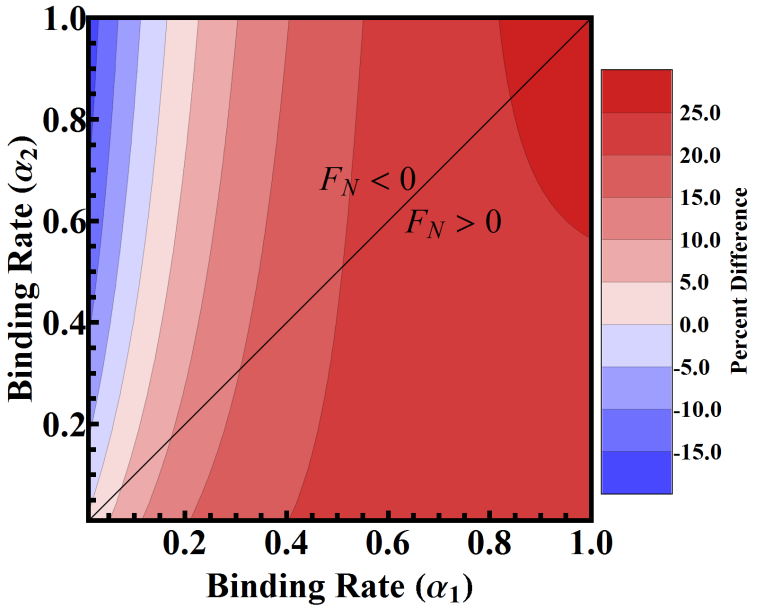

Figure 4.5: Contour plot of the percent change in noise strength of a two regulator pathway from its corresponding unregulated value as a function of α_1 and α_2, i.e. $f(\alpha_1, \alpha_2) = \frac{Q_2}{1+\langle p_{b,2}\rangle} \cdot 100\%$. The change in mean from the unregulated to regulated pathways, F_N, is positive below and negative above the line $\alpha_1 = \alpha_2$. The parameters used were $k_{p_0} = 50$, $k_{p_1} = 200$, $k_{p_2} = 72.5$, $\mu_{c_0} = 1$, $\mu_{c_1} = 2.725$, $\mu_{c_1} = 2.725$, $\beta_1 = 0.15$ and $\beta_2 = 0.15$.

of the regulated protein distribution can be left unchanged; however, the noise strength can be tuned over a large range. For example, for some choices of the parameters, the mean can be fixed and the noise strength can be varied by over 100% relative to the unregulated distribution (see Fig.(4.4)). In this context, it is interesting to note that it has been observed that several sRNAs have a minimal effect on the mean levels of their regulatory targets. For such targets, sRNAs could be functioning primarily as modulators of noise while giving rise to only a minimal change in mean levels due to regulation [30]. Our results provide quantitative insight into how such regulation can be implemented using multiple sRNA regulators.

The results obtained also illustrate how changing sRNA concentrations can be used to modulate the noise in gene expression, as shown in Fig.(4.5). For our model, changes in the concentration of the sRNA regulators effectively alter the binding rates (α_i) to the mRNA. From Eq. (4.22), we see that for 2 regulators, by choosing one of the regulators to be a repressor and the other to be an activator, the mean of the regulated protein distribution can be increased ($F_N > 0$) or decreased ($F_N < 0$) by adjusting the relative concentrations of the two regulators. Furthermore, by changing the concentrations of the regulators such that their relative concentration is fixed, the mean of the regulated protein distribution is left unchanged, whereas the variance can be tuned over a range of values. This insight is particularly relevant, given that noise can be advantageous to a cell, especially in response to stress.

Chapter 5

Post-transcriptional Regulation of Noise II

In the previous chapter, we have studied the model of post-transcriptional regulation in gene expression when regulators (*e.g.* regulatory sRNA) are present in large amount such that the fluctuations in regulator concentration can be neglected. On the other hand, there is also the case such that we have to take the fluctuations of the regulators into consideration. In such situation, the typical model that is intensively analyzed is to consider sRNA regulation via stoichiometric degradation of the target mRNA [44, 49, 51, 65, 63, 101]. Previous theoretical studies have primarily focused on mean-field approaches and on steady-state distributions using expansions around mean-field solutions. However, mean-field approaches will not be accurate when we have a combination of nonlinear reaction rates (due to interaction with small RNAs) and low mRNA/sRNA levels, thereby pointing to the need for better

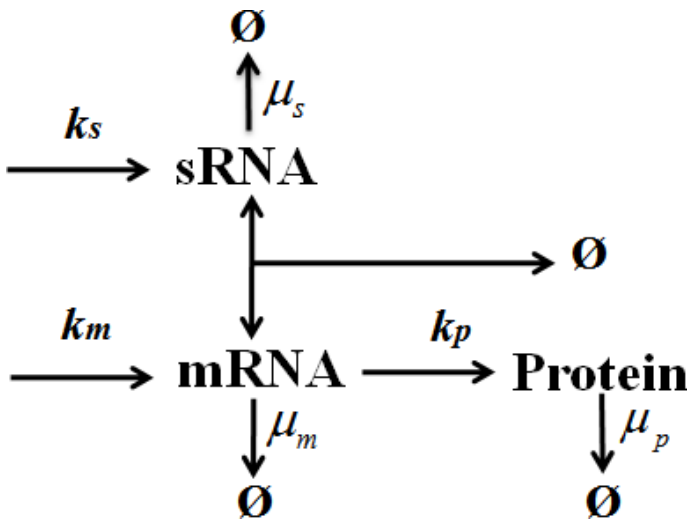

Figure 5.1: The reaction scheme of post-transcriptional regulation by sRNA. The regulation is via a stoichiometric degradation at rate γ.

approximation approaches. We will discuss the insufficiency of mean-field approach [74] and introduce some approximate approaches in the following.

5.1 Mean-Field Approach

The reaction scheme of the problem analyzed is shown in Fig.(5.1). The regulatory sRNA is produced at rate k_s and degraded at rate μ_s. The coupled degradation rate is γ. Denoting the steady state mRNA and sRNA probability distribution by $P(m, s)$, we can write the corresponding master equation

based on the reaction scheme as:

$$\begin{aligned} 0 &= k_m(P(m-1,s)-P(m,s)) + \mu_m\big((m+1)P(m+1,s)-mP(m,s)\big) \\ &+ k_s(P(m,s-1)-P(m,s)) + \mu_s\big((s+1)P(m,s+1)-sP(m,s)\big) \\ &+ \gamma\big((m+1)(s+1)P(m+1,s+1)-msP(m,s)\big). \end{aligned} \tag{5.1}$$

Note that on the third line of the equation, there are nonlinear terms that make the exact analytical solution intractable.

Following the methods introduced in Chapter 2, we can find the equation for the mean mRNA and sRNA levels via generating functions. Define m_s and s_s to the random variables characterizing the number of mRNA and sRNA in the steady-state, we have

$$\begin{aligned} 0 &= k_m - \mu_m\langle m_s\rangle - \gamma\langle m_s s_s\rangle \\ 0 &= k_s - \mu_s\langle s_s\rangle - \gamma\langle m_s s_s\rangle. \end{aligned} \tag{5.2}$$

The mean-field approach ignores the correlation between m_s and s_s such that $\langle m_s s_s\rangle = \langle m_s\rangle\langle s_s\rangle$. By taking this condition into Eq.(5.2), we will have quadratic equations of variable $\langle m_s\rangle$ and $\langle s_s\rangle$:

$$\begin{aligned} \frac{\mu_m}{k_m}\langle m_s\rangle - \frac{\gamma}{k_m}\langle m_s\rangle\langle s_s\rangle - 1 = 0 \\ \frac{\mu_s}{k_s}\langle s_s\rangle - \frac{\gamma}{k_s}\langle m_s\rangle\langle s_s\rangle - 1 = 0. \end{aligned} \tag{5.3}$$

Eq.(5.3) can be further simplified by introducing the following dimensionless

variables

$$\begin{aligned} n_m &= \frac{k_m}{\mu_m}, \quad & \epsilon_m &= \frac{k_s\gamma}{\mu_m\mu_s} \\ n_s &= \frac{k_s}{\mu_s}, \quad & \epsilon_s &= \frac{k_m\gamma}{\mu_m\mu_s} \\ X &= \frac{\langle m_s\rangle}{n_m}, \quad & Y &= \frac{\langle s_s\rangle}{n_s}, \end{aligned} \tag{5.4}$$

which leads to simple equations

$$\begin{aligned} \epsilon_m XY + X - 1 &= 0 \\ \epsilon_s XY + Y - 1 &= 0. \end{aligned} \tag{5.5}$$

Eq.(5.5) can be solved easily. The solution of X and Y lead to steady state mean mRNA and sRNA levels. Eq.(5.5) implies that X and Y depend on variables ϵ_m and ϵ_s only. As the result, we can test the validity of the mean-field approximation by checking the relationship between X, Y and other dimensionless variables (*e.g.* n_m and n_s). Furthermore, another indicator of the accuracy of mean-field results is the ratio $C = \langle m_s s_s\rangle / \langle m_s\rangle\langle s_s\rangle$ which equals 1 under the approximation.

The simulation data of X and C is plotted in Fig.(5.2). We can see that the value of X depends on n_m and n_s, which is inconsistent with the mean-field approximation. Furthermore, in the biological important parameter region (small n_s and n_m values), the value of C also deviates from 1. This indicates that mean-field approach is insufficient in solving the regulation

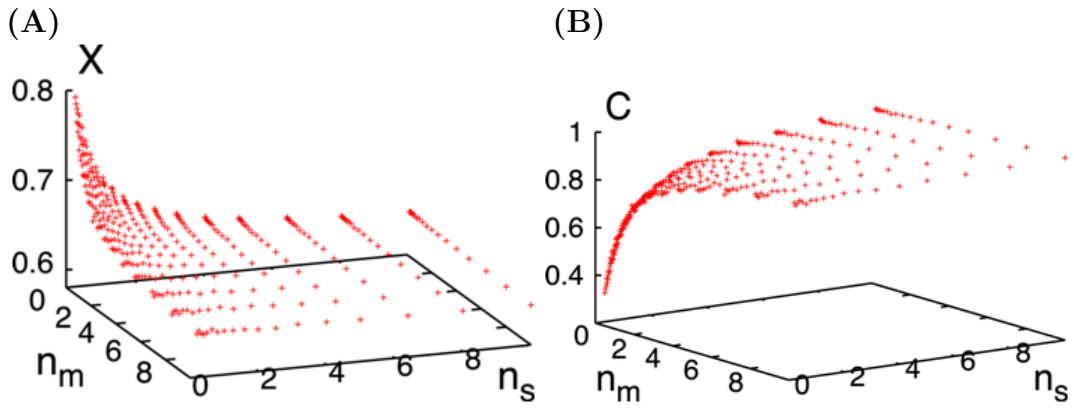

Figure 5.2: (A) The simulation data of $X = \langle m_s \rangle / n_m$ plotted as a function of parameter n_m and n_s. While the mean-field approximation predicts X is independent of n_m and n_s, the simulation data shows dependency of n_m and n_s. (B) The simulation data of $C = \langle m_s s_s \rangle / \langle m_s \rangle \langle s_s \rangle$ plotted as a function of parameter n_m and n_s. While the mean-field approximation assumes $C = 1$, the actual value deviates from 1 when n_m and n_s is small.

model. Better approximation approaches are needed to solve the problem. These approaches will be discussed further below.

5.2 Approximation for Infrequent Transcription

Given the nonlinearity of the problem, the exact analytical solution is hard to find. However, in some cases, we can apply approximations to get reasonably good estimates. In this section, we will discuss conditions under which such analytical solutions can be derived.

Similar to discussion in Chapter 4, we will focus on the "burst limit" that protein lifetime is much longer than that of mRNA (*i.e.* $\tau_p \gg \tau_m$). In this limit, we can approximate the gene expression process as the independent production of bursts of proteins with random sizes. We will mainly concen-

trate in how the regulation will change the protein burst size distribution. Furthermore, we assume low transcription rate k_m such that sRNA distribution *prior* to each burst can be well approximated by the unregulated small RNA distribution, which corresponds to a Poisson distribution with mean $n_s = \frac{k_s}{\mu_s}$. With these approximations, it is possible to derive an expression for the regulated protein burst distribution due to interaction with sRNAs as shown below.

Let us begin with the initial condition ($t = 0$) corresponding to the creation of a mRNA. The protein burst distribution corresponds to the number of proteins produced from this single mRNA until the time it is degraded, either naturally or due to interaction with small RNAs. Our approach will focus on first deriving an expression for the survival probability of the mRNA at time t ($S(t)$). Let us define $\tilde{P}(n,t)$ as the probability that the mRNA exists at time t (i.e. it has not been degraded) and the number of sRNAs is n. Then, the mRNA survival probability is given by $S(t) = \sum_{n=0}^{\infty} \tilde{P}(n,t)$.

The master equation for the probability function $\tilde{P}(n,t)$ can be written as:

$$\begin{aligned} \partial_t \tilde{P}(n,t) &= k_s\big(\tilde{P}(n-1,t) - \tilde{P}(n,t)\big) \\ &+ \mu_s\big((n+1)\tilde{P}(n+1,t) - n\tilde{P}(n,t)\big) \\ &- \mu_m \tilde{P}(n,t) - \gamma n \tilde{P}(n), \end{aligned} \tag{5.6}$$

with initial condition that

$$\tilde{P}(n, t = 0) = e^{-n_s} \frac{n_s^n}{n!}, \tag{5.7}$$

where $n_s = (k_s/\mu_s)$ (i.e., Poisson distribution of sRNAs at time $t = 0$) as discussed above.

Following the procedures discussed previously, we first introduce the generating function for $\tilde{P}(n, t)$ as $G(z, t) = \sum_{n=0}^{\infty} z^n \tilde{P}(n, t)$. Based on the similar properties introduced in Eq.(2.9), we can write a partial differential equation based on Eq.(5.7) as

$$\begin{aligned} \partial_t G(z,t) &= k_s(z-1)G(z,t) - \mu_s(z-1)\partial_z G(z,t) \\ &- \mu_m G(z,t) - \gamma z \partial_z G(z,t), \end{aligned} \tag{5.8}$$

with the corresponding initial condition

$$G(z, 0) = \exp\left(n_s(z-1)\right). \tag{5.9}$$

The value of the generating function $G(x, t)$ at point $z = 1$ corresponds to $\sum_{n=0}^{\infty} \tilde{P}(n, t)$, i.e., the survival probability $S(t)$ of the mRNA molecule at time t. This survival probability can be obtained by solving Eq.(5.8) using the method of characteristics (see Appendix). We obtain

$$S(t) = \exp\left[-a\left(1 - e^{-\tau}\right) - b\tau\right], \tag{5.10}$$

where the dimensionless parameters are defined as

$$\begin{aligned} \tau &= (\mu_s + \gamma)t \\ a &= \left(n_s - \frac{k_s}{\mu_s + \gamma}\right)\frac{\gamma}{\mu_s + \gamma} \\ b &= \frac{\mu_m}{\mu_s + \gamma} + \frac{\gamma k_s}{(\mu_s + \gamma)^2}. \end{aligned} \tag{5.11}$$

We can now proceed and calculate the generating function $G_{pb}(z)$ of the protein burst distribution. Since protein production occurs at a constant rate k_p during the mRNA lifetime, the number of proteins produced by a surviving mRNA in time t is given by the Poisson distribution, with the corresponding generating function given by $e^{k_p(z-1)t}$. Since the difference $S(t) - S(t+\delta t)$ of survival probabilities is the probability that the mRNA degrades within the time interval $\{t, t+\delta t\}$, we obtain the burst generating function as

$$G_{pb}(z) = -\int_0^\infty \mathrm{d}t\, \partial_t S(t) e^{k_p(z-1)t}. \tag{5.12}$$

Rewriting the burst size distribution in terms of dimensionless parameters results in the following integral form

$$G_{pb}(z) = 1 - k(1-z)\int_0^1 \mathrm{d}x\, x^{k(1-z)+\beta-1} e^{\alpha(x-1)}, \tag{5.13}$$

where k is yet another dimensionless parameter

$$k = \frac{k_p}{\mu_s + \gamma}. \tag{5.14}$$

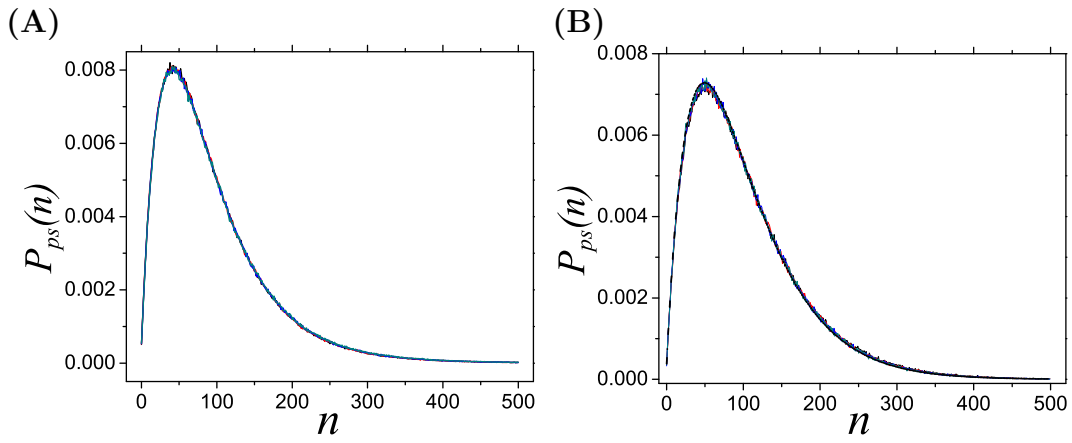

Figure 5.3: Protein steady-state distributions with sRNA regulation based on simulation. (A).Parameters are chosen from Table 5.1. Distributions corresponding to different choice of parameters all collapse to a single curve. (B). Parameters are chosen from Table 5.2. Distributions corresponding to different choice of parameters all collapse to a single curve which coincides with the theoretical predication (bold dashed line).

The burst distribution with sRNA regulation, introduced by Eq.(5.13), has some interesting scaling features. First of the all, we note that the burst size distribution studied here depends on five variables k_s, μ_s, μ_m, γ and k_p. By scaling one to all others, we have four dimensionless variables. Eq.(5.13), on the other hand, predicts that the burst distribution depends on three dimensionless parameters, a, b, and k (defined in Eq.(5.11) and (5.14)) and the steady-state distribution (see Eq.(2.16)) only adds a dependence on k_m/μ_p. Thus the modulation of any of the kinetic parameters in this problem (for fixed k_m/μ_p) should result in the same steady-state distribution so long as the modifications occur in such a way that a, b, and k remain constant (and model assumptions/approximations are valid). As shown in Table 5.1, we can choose very different kinetic parameters that give rise to the same values for

Table 5.1: The values of the parameters used in the numeric simulations shown in Fig.(5.3A). For all simulations, $\alpha \simeq 4.76$, $\beta \simeq 1.34$, and $k \simeq 243.9$. Also, $\mu_m = 1$, $k_m = 0.01$, $\mu_p = 0.005$.

Simulation #	k_p	k_s	μ_s	γ
1	250	0.400313	0.072619	0.952381
2	300	0.717708	0.122308	1.107690
3	400	1.378120	0.217778	1.422222
4	500	2.055630	0.310870	1.739130

a, b and k and the prediction is that the burst and steady-state distributions for these different parameter choices should collapse onto a single curve. As can be seen in Fig.(5.3A), the simulation results are consistent with the scaling prediction since the curves with different parameter choices all collapse onto a single curve.

Moreover, we can do further approximations to the survival probability in Eq.(5.10). By expanding $S(t)$ in a power series, we can see that when the value of b is large, $S(t)$ can be approximated as

$$S(t) = e^{-(a+b)\tau}. \tag{5.15}$$

By taking the definitions of the parameters a, b and τ into the equation, we can express the mRNA survival probability as

$$S(t) = e^{-(\mu_m + \frac{k_s}{\mu_s}\gamma)t}. \tag{5.16}$$

This tells us that mRNA has a constant degradation rate $\mu_m + \gamma k_s/\mu_s$. The first term is the mRNA self degradation rate and the latter is the effective

Table 5.2: The values of the parameters used in the numeric simulations shown in Fig.(5.3B). For all simulations, $k_m = 0.1$, $\mu_m = 1.0$ and $\mu_p = 0.05$. The mean burst size $\langle p_b \rangle$ based on Eq.(5.17) is fixed to be 50.

Simulation #	k_p	k_s	μ_s	γ	a	b
1	200	6.0	0.067	0.033	10	30
2	250	12	0.075	0.025	10	40
3	300	7.5	0.12	0.08	10	20
4	400	5.25	0.086	0.114	20	20

degradation rate due to sRNA regulation. Eq.(5.16) indicates that the fluctuation of sRNA can be ignored when b is large and we effectively have a "mean-field" degradation rate. Following the survival probability in Eq.(5.16), the protein burst size distribution will be geometric. Similar to Eq.(4.7) and Eq.(4.8), the corresponding generating function can be found as

$$G_b(z) = \frac{\mu_m + \frac{k_s}{\mu_s}\gamma}{\mu_m + \frac{k_s}{\mu_s}\gamma + k_p(1-z)}. \tag{5.17}$$

Similarly, we carry out simulations to test our predication. We choose a wide range of parameters as shown in Table 5.2, with large b value and fixed mean protein burst size $\langle p_b \rangle = 50$ (which gives the same steady-state mean for all sets of parameters). All the curves (as shown in Fig.(5.3B)) with different choices of parameters all collapse onto a single curve that can be well fitted by the distribution given by Eq.(5.17).

5.3 Approximation for Infrequent Transcription with mRNA Bursts

The discussion so far is focused on Poisson production of mRNA, meaning that only one mRNA can be created in a transcription event. This is the most typical situation that was studied extensively [51, 63, 44, 65, 63, 101, 36]. However, not much work has been done considering transcriptional bursting with regulatory sRNAs fluctuating with time. Besides the nonlinearity of the problem, the difficulty in solving this problem also lies in the fact that the mRNAs created in each burst is not independent anymore: when one mRNA degrades with sRNA, the regulatory concentration will change for the rest. Due to this correlation, we can no longer treat the proteins produced by all mRNAs as a compound random variable and the result developed in the preceeding session can not be applied directly.

Despite these difficulties, it is noteworthy that in case of infrequent transcription and strong regulation, an approximate solution of the problem can be obtained. In the following discussion, we will start with some extreme conditions under which the exact solution can be found. We will gradually release the constraints on our model, move out from the extreme parameter regions that we start with and eventually come up with an approximate solution that is valid for a wide range of parameters.

5.3.1 Basic Result

The model analyzed is the same as that is shown in Fig.(5.1) except that mRNAs are now produced in bursts with the burst size distribution $P_{mb}(n)$ as

$$P_{mb}(m) = (1 - p_m)^{m-1} p_m, \tag{5.18}$$

which serves as a general formula for the mRNA burst distribution as studied in Chapter 4 (Eq.(4.11)). The parameter p_m controls the degree of transcriptional burst with the mean burst size $\langle m_b \rangle = 1/p_m$. We will still concentrate on the "burst limit" that $\tau_p \gg \tau_m$ under which the gene expression can be simplified as processes of creation and decaying of proteins. We focus on the case that transcription rate k_m is low such that the bursts of proteins are typically well-separated in time and can be considered as independent events. Following that, finding the protein burst size distribution, which is what we will analyze next, will be sufficient to derive the protein steady-state distribution.

We start with the following extreme conditions that

1). mRNA degrades immediately on the appearance of sRNA.

2). There is no synthesis of new sRNAs *during* a burst.

3). The distribution of sRNAs prior to a burst is the steady-state distribution of sRNAs in the absence of mRNAs, which is a Poisson distribution with the mean $n_s = \frac{ks}{\mu_s}$.

Given that the assumption 1) is valid, the regulation by sRNA is an instant modification of mRNA transcriptional burst level. Denoting m and s as the number of mRNA and sRNA at the beginning of the burst respectively, the distribution of s is a Poisson distribution by assumption 3) that can be defined as

$$\rho(s) = \frac{(n_s)^s}{s!} e^{-n_s}. \tag{5.19}$$

Proteins are produced only when $m > s$ and under this circumstance, the rest $m - s$ mRNA will proceed into translation process as if there is no regulation at all, based on assumption 2). The total protein in one burst will be a compound random variable depending on the the number of surviving mRNAs after the regulation. Define $G'_{pb}(z)$ as the generating function corresponding to the total number of proteins produced in one burst and $G_{pb}(z)$ as the generating function of protein burst size distribution by a single mRNA without regulation, we have

$$G'_{pb}(z) = \sum_{s,m} \big(Prob(m \leq s) + Prob(m > s)(G_b(z))^{m-s} \big). \tag{5.20}$$

In combination of Eq.(5.18) and Eq.(5.19) that state the mRNA burst size

distribution and sRNA distribution prior to the burst, Eq.(5.20) will lead to

$$\begin{aligned} G'_{pb}(z) &= \sum_{i=j}^{\infty}\sum_{j=1}^{\infty}\rho(i)P_{mb}(j) + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty}\rho(i)P_{mb}(i+j)G_{pb}^{j}(z) \\ &= 1 - \sum_{i=0}^{\infty}(1-p_m)^i\rho(i)\sum_{j=1}^{\infty}p_m(1-p_m)^{j-1} \\ &+ \sum_{i=0}^{\infty}(1-p_m)^i\rho(i)\sum_{j=1}^{\infty}p_m(1-p_m)^{j-1}G_{pb}^{j}(z) \\ &= 1 - e^{-n_s p_m} + e^{-n_s p_m}\frac{p_m G_{pb}(z)}{1-G_{pb}(z)(1-p_m)}.. \end{aligned} \tag{5.21}$$

Note that the generating function (defined as $\tilde{G}'_{pb}$) of total number of proteins without regulation can be expressed as (according to Eq.(4.12))

$$\tilde{G}'_{pb} = \frac{p_m G_{pb}(z)}{1-G_{pb}(z)(1-p_m)}. \tag{5.22}$$

Eq.(5.21) can be rewritten as

$$G'_{pb}(z) = 1 - e^{-n_s p_m} + e^{-n_s p_m}\tilde{G}'_{pb}. \tag{5.23}$$

Eq.(5.23) demonstrates the threshold for the gene expression under the sRNA regulation[50, 48, 49]. The first two terms in Eq.(5.23) will contribute to the probability of producing zero protein. All the factorial moments will be rescaled by the factor $e^{-n_s p_m}$. As the result, only the transcriptions with mean burst size $\langle m_b \rangle = 1/p_m > n_s$ can be effectively expressed. We can further derive the mean, variance and the noise (squared coefficient of variance) based

on Eq.(5.23). Denote p'_b and $\tilde{p}'_b$ as the random variable characterizing the number of proteins produced with and without regulation, respectively, we have

$$\begin{aligned} \langle p'_b \rangle &= e^{-n_s p_m} \langle \tilde{p}'_b \rangle \\ \sigma^2_{p'_b} &= e^{-n_s p_m} \sigma^2_{\tilde{p}'_b} + (e^{n_s p_m} - 1) \langle p'_b \rangle^2 \\ \frac{\sigma^2_{p'_b}}{\langle p'_b \rangle^2} &= e^{n_s p_m} \frac{\sigma^2_{\tilde{p}'_b}}{\langle \tilde{p}'_b \rangle^2} + (e^{n_s p_m} - 1). \end{aligned} \tag{5.24}$$

We can see that while the mean protein level under regulation decreases by a factor $e^{-n_s p_m}$, the noise in protein burst size distribution increases. Though more accurate expressions of mean and variance can be derived based on the following discussion, this conclusion does not change. This is mainly due to the fact that regulation will give rise to large probability of no protein (or very few if γ is reasonably large) produced in the burst.

5.3.2 Advanced Analysis

The assumptions that Eq.(5.23) is based on are only valid for a limited range of parameters. Further analysis is needed for moving beyond them. Let us first take assumption 1). The purpose of this assumption is to give the coupled degradation the most priority to occur among all reactions such that the reduction of sRNAs and mRNAs takes place instantaneously. Ideally this requires infinitely large γ value. However, since this assumption is to

freeze the system dynamics during the coupled degradation, it can be approximately achieved by choosing γ value much greater than μ_m and k_s. $\gamma \gg \mu_m$ is to make sure that the regulation is stronger than natural degradation and the number of proteins created during the coupled degradation can be neglected. $\gamma \gg k_s$ prevents the sRNA dynamics during the coupled degradation (the sRNA number is typically greater than 1 so $\gamma \gg \mu_s$ when $\gamma \gg k_s$). Based on simulation results shown in Fig.(5.4), we can see that when $\gamma \geq 10 \max[k_s, \mu_m]$, the mean and variance will not change very much as γ increases and we can consider assumption 1) is approximately satisfied. While previous work considering sRNA regulation under Poisson production of mRNA often uses small γ value, it is also possible that the binding rate can be large [66].

Assumption 2) implies that there is no sRNA creation during mRNA life time. While this can be achieved by taking small $\frac{k_s}{\mu_m}$ values, problems would occur if we want to study a broader parameter region. So a modification of Eq.(5.21) has to be made. Let us consider the case when mRNAs outnumber sRNAs and there are $m' = m - s$ mRNAs left when all sRNAs die out though the initial mutual degradation. For the m' mRNA's, we rank them by the *inverse* order of their degradations. Specifically, the mRNA 1 is the last mRNA degraded whereas the mRAN m' is the one that degrades first. Denote T_i as the time interval between the degradation of mRNA i and mRNA $i+1$ and $T_{m'}$ is the time elapsed until the degradation of the first mRNA. According to the basic knowledge in stochastic process, we know that

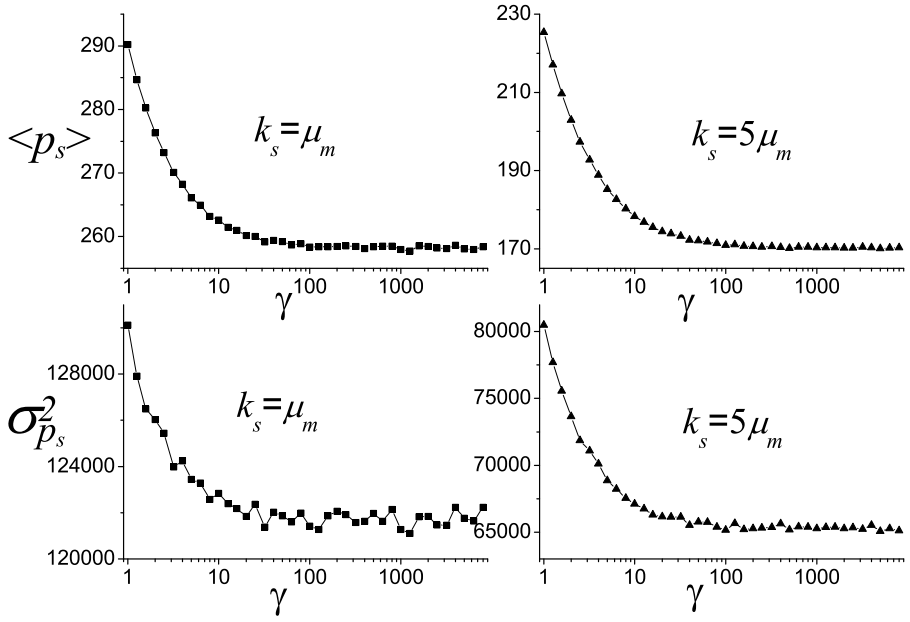

Figure 5.4: The mean and variance of protein steady-state distributions based on simulation. In one case, $k_s = \mu_m$ and in the other case $\frac{k_s}{\mu_m} = 5$. For both cases, $\frac{k_s}{\mu_s} = 5$, $p_m = 0.1$, $\frac{k_m}{\mu_m} = \frac{\mu_p}{\mu_m} = 0.01$, $\frac{k_p}{\mu_m} = 50$ and $\mu_m = 1$. Both mean and variance become steady when γ is large ($\gamma > 10 \max[\mu_m, k_s]$).

the T_i will follow the exponential distribution with mean $1/i\mu_m$ when there is no regulation. The overall protein creation rate during T_i is ik_p. We further notice that based on assumption 1), the creation of one sRNA will result in the immediate degradation of one mRNA. Following that, T_i will still be with exponential distribution but the mean will be $1/(i\mu_m + k_s)$.

Define $G_i(z)$ as the generating function of proteins created during time period T_i. Based on results introduced previously (*e.g.* Eq.(4.7)), we have

$$G_i(z) = \frac{i\mu_m + k_s}{i\mu_m + k_s + ik_p(1-z)}. \tag{5.25}$$

Because the time interval T_i's are independent, we can multiply the $G_i(z)$'s together to get the generating function of total number of proteins in one burst. Conditioning on m' left after the initial coupled degradation, we have

$$G'_{pb}|m'(z) = \prod_{i=1}^{m'} \frac{i\mu_m + k_s}{i\mu_m + k_s + ik_p(1-z)}. \tag{5.26}$$

By taking the probability that m' mRNAs survive into account, we have

$$\begin{aligned} G'_{pb}(z) &= \sum_{i=j}^{\infty}\sum_{j=1}^{\infty} \rho(i)P_{mb}(j) \\ &+ \sum_{l=0}^{\infty}\sum_{j=1}^{\infty} \rho(i)P_{mb}(l+j) \prod_{i=1}^{j} \frac{i\mu_m + k_s}{i\mu_m + k_s + ik_p(1-z)} \\ &= 1 - e^{-n_s p_m} + e^{-n_s p_m} \sum_{j=1}^{\infty} p_m(1-p_m)^{j-1} \prod_{i=1}^{j} \frac{i\mu_m + k_s}{i\mu_m + k_s + ik_p(1-z)}. \end{aligned}$$

. (5.27)

Unfortunately we can not simplify Eq.(5.27) further. From the generating function, we can have the mean protein burst size as

$$\langle p'_b \rangle = e^{-n_s p_m} \sum_{j=1}^{\infty} p_m(1-p_m)^{j-1} \sum_{i=1}^{j} \frac{ik_p}{i\mu_m + k_s}. \tag{5.28}$$

Eq.(5.28) can be simplified further by expanding the last term as a series of

k_s,

$$\begin{aligned}
\langle p_b' \rangle &= e^{-n_s p_m} \sum_{j=1}^{\infty} p_m (1-p_m)^{j-1} \sum_{i=1}^{j} \sum_{l=0}^{\infty} (\frac{-k_s}{i\mu_m})^l \\
&= e^{-n_s p_m} \frac{k_p}{\mu_m} \sum_{l=0}^{\infty} (\frac{-k_s}{\mu_m})^l \sum_{j=0}^{\infty} \sum_{i=1}^{\infty} p_m (1-p_m)^{i+j-1} \frac{1}{i^l} \\
&= e^{-n_s p_m} \frac{k_p}{\mu_m (1-p_m)} \sum_{s=0}^{\infty} (\frac{-k_s}{\mu_m})^s Li_s(1-p_m), \qquad (5.29)
\end{aligned}$$

where $Li_s(z)$ is the polylogarithmic function defined as

$$Li_s(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^s}. \qquad (5.30)$$

Finally, let us come to assumption 3) which says that the number of sRNA prior to a transcription event follows Poisson distribution with mean $n_s = k_s/\mu_s$. Without the appearance of mRNA, sRNA evolves according to the standard birth and death process which gives a Poisson distribution in steady-state. The assumption 3) will be true when each transcription occurs very infrequently (very small k_m value). However, when k_m value is not that small, we have to consider the transient behavior of the sRNA evolution.

Define $G_s(z,t)$ as the generating function of sRNA distribution at time t. $t = 0$ is the time that the translational burst ends, *i.e.* all mRNAs created in one transcriptional burst are degraded. In the problem analyzed, we are more interested in situations that the regulation tunes the protein level but does not fully repress the translation. This corresponds to the parameter

region that $1/p_m \geq n_s$. In this parameter region, the mRNAs produced in transcriptional burst usually outnumber the sRNAs and typically there is no sRNA left after the regulation. This provides the initial condition that the number of sRNA is zero at $t = 0$. Based on the results of birth and death process [61], we have

$$G_s(z,t) = \exp[-\frac{k_s}{\mu_s}(1-e^{-\mu_s t})(1-z)]. \tag{5.31}$$

The waiting time distribution that the next burst occurs is exponential with mean $1/k_m$. Then the generating function of sRNA distribution prior to the burst (distribution $\rho(s)$) will be

$$G_s(z) = \int_0^\infty \exp[-\frac{k_s}{\mu_s}(1-e^{-\mu_s t})(1-z)] \times k_m e^{-k_m t} dt. \tag{5.32}$$

Note that in the above deviation, the term that contributes to the final result is $\sum_{i=0}^{\infty}(1-p_m)^i\rho(i)$ (see Eq.(5.21)). By recalling the definition of generating function $G_s(z) = \sum_{i=0}^{\infty} z^i\rho(i)$, we notice that the term $\sum_{i=0}^{\infty}(1-p_m)^i\rho(i)$ equals $G_s(1-p_m)$. Thus a more accurate expression of the results presented (Eq.(5.27)-Eq.(5.29)) is to replace the term $e^{-n_s p_m}$ by $G_s(1-p_m)$. On the other hand, the form of $G_s(1-p_m)$ is complicated. Based on numerical evaluation, we can approximate $\rho(i)$ by a Poisson distribution with the mean given by $G_s(z)$. This only requires replacing the term $n_s = k_s/\mu_s$ in $e^{-n_s p_m}$ by $n_s' = k_s/(\mu_s + k_m)$, which gives a more simple form and very close to $G_s(1-p_m)$.

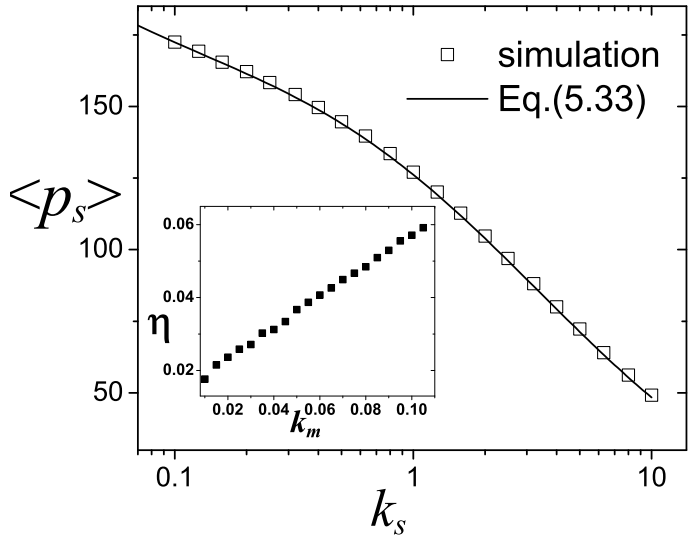

Figure 5.5: The steady-state mean protein number *vs* k_s. The calculation based on Eq.(5.33) is very close to simulation result. The parameters are chosen as $\mu_m = 1$, $\frac{k_s}{\mu_s} = 2$, $k_m = \mu_p = 0.01$, $p_m = 0.2$, $\gamma = 100$ and $k_p = 50$. In the inset, we show the relative error η *vs* k_m. The error increases as k_m increases. The parameters used in the inset are $\mu_m = \mu_s = 1$, $k_s = 2$, $\mu_p = 0.01$, $p_m = 0.2$, $\gamma = 50$ and $k_p = 50$.

Given the protein burst size distribution, we can connect it to the protein steady-state level via Eq.(2.20), which gives

$$\langle p_s \rangle = \frac{k_m}{\mu_p} e^{-\frac{k_s p_m}{\mu_s + k_m}} \frac{k_p}{\mu_m(1-p_m)} \sum_{s=0}^{\infty} \left(\frac{-k_s}{\mu_m}\right)^s Li_s(1-p_m). \tag{5.33}$$

The result in Eq.(5.33) is tested by simulation for a range of parameters with $\frac{k_s}{\mu_m} \in [0.1, 10]$ and $\frac{k_m}{\mu_m} \in [0.01, 0.1]$. The plot is shown in Fig.(5.5). Eq.(5.33) shows a perfect match with the simulation when k_m value is small. While the error does not depend on k_s, it increases when k_m becomes large. This is because our deduction is based on the infrequent transcription that the bursts are clearly separated in time. Increasing k_m will violate this assumption and

bring more error. However, even take this into consideration, Eq.(5.33) still accurately quantifies the steady-state mean protein level (within 6% error) for the range of parameters tested.

5.4 Quantifying mRNA Synthesis and Decay rates Using sRNA

This subsection considers some applications of the results derived to quantifying synthesis and decay rates for mRNAs. The traditional approach for measuring mRNA lifetimes involves quantification of mRNAs remaining at different times following inhibition of transcription, e.g., by the addition of rifampicin [10]. This procedure requires multiple measurements during time intervals of the order of the mRNA lifetime, hence high temporal resolution is required for short-lived mRNAs. More significantly, the procedure for inhibition of transcription can give rise to secondary effects which influence mRNA decay [10], hence it is of interest to consider alternative approaches. With our study of sRNA-based regulation, we outline a novel proposal for quantifying mRNA decay [15].

The experimental setup in our proposal for quantifying mRNA decay is as follows. Consider three different strains as shown in Fig.(5.6): two unregulated strains (i.e., with either sRNA or mRNA deleted) and the wild type (WT) strain. In the WT strain, both mRNA and sRNA are present and regulate

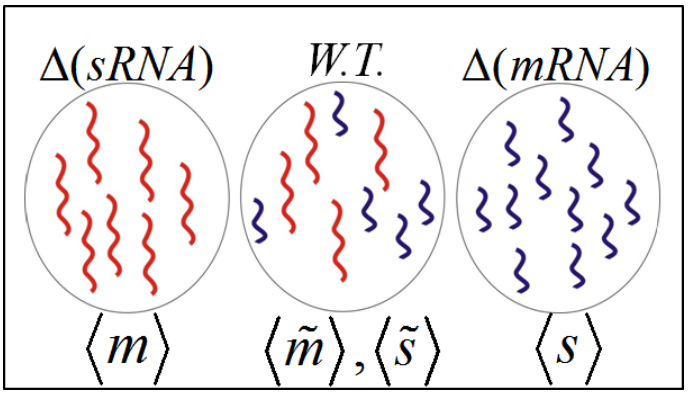

Figure 5.6: The proposed setup involves steady-state measurements for three strains: Δ(sRNA), WT and Δ(mRNA). Unregulated steady-state mean levels of mRNAs and sRNAs along with regulated levels of these molecules are measured. The measured quantities allow determination of the average mRNA transcription rate k_m and decay rate τ_m relative to the sRNA production rate k_s. If k_s is held fixed, and the conditions are varied, the proposed scheme leads to simultaneous determination of fold-changes in the rate of transcription and the rate of mRNA decay. Note that the mRNA/sRNA interaction parameter can be arbitrary.

each other. In steady-state, we derive the following exact relations connecting mRNA/sRNA lifetimes and transcription rates to the mean abundances

$$\begin{aligned} k_m\ \tau_m &= \langle m \rangle \\ k_s\ \tau_s &= \langle s \rangle \end{aligned} \tag{5.34}$$

for the unregulated strains, and

$$\frac{\tau_m}{\tau_s} = \frac{\langle m \rangle - \langle \tilde{m} \rangle}{\langle s \rangle - \langle \tilde{s} \rangle}. \tag{5.35}$$

Here $\langle m \rangle$ and $\langle s \rangle$ are the mean mRNA(sRNA) abundances in strains lacking the sRNA(mRNA), and $\langle \tilde{m} \rangle$ and $\langle \tilde{s} \rangle$ are the mean mRNA and sRNA levels

in WT strain where both are present.

The above relations suggest an alternative approach (Fig.(5.6)) for quantifying decay times for mRNAs that either have a naturally occurring small RNA regulator or for which an antisense RNA regulator can be designed. Consider an experimental setup expressing the sRNA from an inducible promoter such that its transcription rate is primarily controlled by inducer concentration. The mean transcription rate k_s can, in principle, be determined using single-molecule methods [99]. Now the basic parameters for the coupled system are k_m, k_s, τ_m and τ_s. If k_s is known, then the values of the other parameters can be determined using experimental measurements of $\langle m \rangle$, $\langle s \rangle$, $\langle \tilde{m} \rangle$ and $\langle \tilde{s} \rangle$ in combination with equations given above. Alternatively, experiments can be designed to keep k_s fixed while factors regulating mRNA decay are changed e.g., by deletion of a protein known to play a role in mRNA decay. The above equations can be used to simultaneously determine the corresponding fold-changes of the mRNA/sRNA lifetimes and the mean mRNA transcription rate.

The derivation of Eq.(5.35) is based on the reaction scheme in Fig.(5.1). Here, we consider the mRNA and sRNA production by *arbitrary* stochastic processes with mean arrival rates given. Degradation of RNA is assumed to be a Poisson process with rate $1/\tau_m$ and $1/\tau_s$,

Let us choose a particular realization of the system evolution during time

interval $t = [0, T]$. For large values of T, we derive

$$x(T) - x(0) = C_x(T) - Y(T) - \tau_x^{-1} \int_0^T \mathrm{d}t\, x(t), \tag{5.36}$$

where $x(t)$ is the number of molecules of the species $x = \{m, s\}$ at the time t.

In Eq.(5.36), $C_x(t)$ is the total number of molecules of the species x created during system evolution until time T, and $Y(T)$ is the total number of molecules of either species that is *mutually* degraded within the time interval $[0, T]$. Finally, using the law of large numbers, the number of molecules degraded naturally in $[0, T]$ is given by the last term in the Eq.(5.36).

Dividing both sides of Eq.(5.36) by T and taking a limit $T \to \infty$ we obtain

$$\lim_{T\to\infty} \frac{x(T) - x(0)}{T} = k_x - \tau_x^{-1} \langle \tilde{x} \rangle - \lim_{T\to\infty} \frac{Y(T)}{T}, \tag{5.37}$$

where $\langle \tilde{x} \rangle$ is average number of molecules in the system. Also by definition, k_x is the mean arrival rate of the species $x = \{m, s\}$. The limit on the left hand side of Eq.(5.37) vanishes in the case of finite degradation rates τ_x^{-1} (number of molecules at any time is finite.) Note that $Y(T)$ is an extensive quantity (it is monotonic increasing function of T) and therefore, the limit on the right hand side of Eq.(5.37) is finite.

Hence, we derive

$$\begin{aligned} k_m - \tau_m^{-1}\langle\tilde{m}\rangle - \lim_{T\to\infty}\frac{Y(T)}{T} &= 0, \\ k_s - \tau_s^{-1}\langle\tilde{s}\rangle - \lim_{T\to\infty}\frac{Y(T)}{T} &= 0, \end{aligned} \tag{5.38}$$

which immediately yields the following expression

$$k_m - \tau_m^{-1}\langle\tilde{m}\rangle = k_s - \tau_s^{-1}\langle\tilde{s}\rangle. \tag{5.39}$$

In the unregulated case $Y(T) = 0$ for any T, since one of the RNA species is deleted and there is no coupled degradation. In this situation one gets

$$0 = k_m - \tau_m^{-1}\langle m\rangle, 0 = k_s - \tau_s^{-1}\langle s\rangle, \tag{5.40}$$

where $\langle x\rangle$, $x = \{m, s\}$ are the average number of molecules during unregulated system evolution. Combining the set of equations above with Eq.(5.39), we derive the results in Eq.(5.35). We note that the derived results are valid even if the binding of mRNA and sRNA is taken to be reversible and the lifetime of the mRNA-sRNA complex is finite. Finally, the time average can be replaced by the ensemble average in the steady-state.

We have validated the derived results using stochastic simulations based on the Gillespie algorithm [22]. Production of RNA molecules was taken to occur in transcriptional bursts [77] that was introduced in previous discussion. The waiting-time between bursts was a random variable drawn from exponential

or Gamma distributions. As expected, the results from the simulations were in excellent agreement with the derived analytical results.

The proposed approach can be used to address several important questions of current interest, some of which are highlighted in the following. By targeted mutagenesis of specific mRNA sequence elements, the induced fold-change in mRNA lifetime, as well as the corresponding change in the transcription rate k_m, can be determined using the same experimental setup. This is an important feature, given that recent experiments have observed coordination between changes in transcription and changes in mRNA degradation [87]. Quantifying the change in mRNA lifetimes induced by mutations to different components of cellular degradation pathways can address such issues as the role of polyadenylation in mRNA decay [37]. It would also be of interest to design high-throughput experiments for different mRNAs which are regulated by corresponding antisense RNAs, all of which are expressed from identical inducible promoters and thus have the same k_s. The proposed procedure can then be used for genome-wide determination of relative transcription rates and lifetimes of mRNAs. These effective parameters, in turn, serve as critical inputs to systems-level models of cellular processes [82].

In summary, we have proved an exact relation for a nonlinear stochastic model of cellular post-transcriptional regulation. The derived results suggest a novel procedure for simultaneous determination of mRNA production rates and mRNA lifetimes. While the focus was on bacterial mRNAs, the procedure

outlined can also be applied to higher organisms and used to systematically explore the sequence determinants and processes involved in regulation of mRNA decay.

Chapter 6

Summary

Stochasticity is a ubiquitous feature of cellular processes and a quantitative analysis of 'noise is essential for understanding many cellular functions. Recent single-cell experiments have carried out such analyses to characterize probability distributions for quantities of interest, e.g. mRNA/protein levels across a population of cells. Correspondingly, there is a need to develop analytical framework for theoretical modeling and interpretation of data obtained from such single-cell experiments.

In this thesis, we address this issue by developing and analyzing general stochastic models of gene expression. We analytically studied the model by mapping it to problems of interest in queueing theory. Our work is one of the first studies that applied this powerful mathematical tool in analyzing the noise in protein distributions for models of stochastic gene expression. This work demonstrates the benefits of developing a mapping between models of

stochastic gene expression and queueing systems which has potential applications for research in both fields. The extensive analytical approaches and tools developed in queueing theory can now be employed to analyze stochastic processes in gene expression.It is also anticipated that the diverse mechanisms of cellular regulation of gene expression will motivate new models for analysis using queuein theory and related approaches.

Another area of focus is models involving post-transcriptional regulation, e.g. by small RNAs (sRNAs) in bacteria, microRNAs (miRNAs) in higher organisms or regulatory proteins. This regulation is known to play a critical role in diverse cellular processes such as development and differentiation. Recent research has increasingly focused on this mode of regulation and its various cellular roles, e.g. global control [25] and fine-tuning of the noise in gene expression [29]. Our results derived for high regulator concentrations provide insight into how different mechanisms of post-transcriptional regulation can be used to fine-tune the noise in stochastic gene expression with potential implications for studies addressing the evolutionary importance of noise in biological systems. In some limits, the modulated burst distributions can be used to infer the degree of transcriptional bursting and hence to determine sources of intrinsic noise in gene expression. The results derived can serve as building blocks for future studies focusing on regulation of stochastic gene expression. In the last chapter, we extended previous work [44, 49, 51, 65, 63, 101] analyzing post-transcriptional modulation of gene expression using expansions around mean-field solutions. The mean-field approaches are not accurate in

biological systems with a combination of nonlinear reaction rates (due to interaction with small RNAs) and low mRNA abundance. For the biological relevant case of infrequent mRNA synthesis events giving rise to protein expression bursts, using appropriate approximations, we derived accurate expressions for mean protein levels and steady-state distributions. The results derived can serve as building blocks for future studies focusing on regulation of stochastic gene expression.

Appendix A

Derivation of survival probability in Chapter5

Solution of the Eq.(5.8)

$$\begin{aligned} \partial_t G(z,t) &= k_s(z-1)G(z,t) - \mu_s(z-1)\partial_z G(z,t) \\ &- \mu_m G(z,t) - \gamma z \partial_z G(z,t), \end{aligned} \tag{A.1}$$

using method of characteristics is given by

$$G_1(z,t) = \exp\left[-b\tau + \frac{k_s(z-1)}{\gamma+\mu_s} + \frac{k_s\gamma}{(\gamma+\mu_s)^2}\right] g(x), \tag{A.2}$$

where b and τ are dimensionless parameters as defined in the main text and the function $g(x)$ needs to be determined from the initial condition in Eq.(5.9)

$$G(z,0) = \exp\left(n_s(z-1)\right). \tag{A.3}$$

Its argument is given by

$$x = \left((z-1) + \frac{\gamma}{\gamma + \mu_s} \right) e^{-\tau}. \tag{A.4}$$

By matching the initial condition one gets

$$g(x) = \exp\left(-\frac{k_s}{\gamma + \mu_s} x \right) \exp\left[n_s x - \frac{\gamma n_s}{\gamma + \mu_s} \right]. \tag{A.5}$$

Finally, since we are interested in the quantity $S(t) = G_1(1,t)$ (survival probability), we obtain

$$x \to \frac{\gamma}{\gamma + \mu_s} e^{-\tau}, \tag{A.6}$$

$$S(t) = \exp\left[-\beta\tau + \frac{k_s \gamma}{(\gamma + \mu_s)^2} \right] g(x). \tag{A.7}$$

from which the Eq.(5.10) from the main text can be obtained.

Bibliography

[1] P. Babitzke and T. Romeo. Csrb srna family: sequestration of rna-binding regulatory proteins. *Curr Opin Microbiol*, 10(2):156–63, 2007.

[2] C. Baker, T. Jia, and R. V. Kulkarni. Stochastic modeling of regulation of gene expression by multiple small RNAs. *ArXiv e-prints*, Jan. 2011.

[3] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, Y. Pilpel, and N. Barkai. Noise in protein expression scales with natural protein abundance. *Nat Genet*, 38(6):636–43, 2006.

[4] A. Becskei, B. Kaufmann, and A. van Oudenaarden. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nature Genetics*, 37(9):937–944, Sep 2005.

[5] O. G. Berg. A model for the statistical fluctuations of protein numbers in a microbial population. *J Theor Biol*, 71(4):587–603, 1978.

[6] I. Buessing, F. J. Slack, and H. Grosshans. let-7 microRNAs in development, stem cells and cancer. *Trend. Mol. Med.*, 14(9):400–409, 2008.

[7] R. Bundschuh, F. Hayot, and C. Jayaprakash. Fluctuations and slow variables in genetic networks. *Biophys J*, 84(3):1606–15, 2003.

[8] L. Cai, N. Friedman, and X. S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–62, 2006.

[9] P. J. Choi, L. Cai, K. Frieda, and S. Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900):442–446, Oct 17 2008.

[10] C. Condon. Maturation and degradation of RNA in bacteria. *Curr. Opin. Microbiol.*, 10:271–278, 2007.

[11] A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173, Sept. 2010.

[12] J. Elf, G.-W. Li, and X. S. Xie. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–1194, May 25 2007.

[13] V. Elgart, T. Jia, A. T. Fenley, and R. Kulkarni. Connecting protein and mrna burst distributions for stochastic models of gene expression. *Physical Biology*, 8(4):046001, 2011.

[14] V. Elgart, T. Jia, and R. V. Kulkarni. Applications of little's law to stochastic models of gene expression. *Phys. Rev. E*, 82(2):021901, Aug 2010.

[15] V. Elgart, T. Jia, and R. V. Kulkarni. Quantifying mRNA synthesis and decay rates using small RNAs. *Biophys. J.*, 98(12):2780–2784, 2010.

[16] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, 2002.

[17] A. Femino, F. Fay, K. Fogarty, and R. Singer. Visualization of single RNA transcripts in situ. *Science*, 280(5363):585–590, Apr 24 1998.

[18] A. S. Flynt and E. C. Lai. Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nature Reviews Genetics*, 9(11):831–42, Nov. 2008.

[19] D. Fraser and M. Kaern. A chance at survival: gene expression noise and phenotypic diversification strategies. *Molecular Microbiology*, 71(6):1333–40, Mar. 2009.

[20] N. Friedman, L. Cai, and X. S. Xie. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett*, 97(16):168302, 2006.

[21] K. S. Fröhlich and J. Vogel. Activation of gene expression by small RNA. *Current Opinion in Microbiology*, 12(6):674–82, Dec. 2009.

[22] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, 1977.

[23] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005.

[24] S. Gottesman. Stealth regulation: biological circuits with small rna switches. *Genes Dev.*, 16(22):2829–2842, 2002.

[25] S. Gottesman. Micros for microbes: non-coding regulatory rnas in bacteria. *Trends Genet.*, 21(7):399–404, 2005.

[26] F. Hayot and C. Jayaprakash. The linear noise approximation for molecular fluctuations within cells. *Phys Biol*, 1(3-4):205–10, 2004.

[27] R. Hershberg, S. Altuvia, and H. Margalit. A survey of small rna-encoding genes in escherichia coli. *Nucl. Acids Res.*, 31(7):1813–20, 2003.

[28] J. E. Hornos, D. Schultz, G. C. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes. Self-regulating gene: an exact solution. *Phys Rev E*, 72:051907, 2005.

[29] E. Hornstein and N. Shomron. Canalization of development by microRNAs. *Nat. Genet.*, 38(Suppl. 6):S20–S24, 2006.

[30] E. Hornstein and N. Shomron. Canalization of development by microRNAs. *Nature Genetics*, 38 Suppl(June):S20–S24, June 2006.

[31] P. J. Ingram, M. P. H. Stumpf, and J. Stark. Nonidentifiability of the Source of Intrinsic Noise in Gene Expression from Single-Burst Data. *PLoS Comp Biol*, 4(10), 2008.

[32] M. Inui, G. Martello, and S. Piccolo. MicroRNA control of signal transduction. *Nat. Rev. Mol. Cell Biol.*, 11(4):252–263, 2010.

[33] S. Iyer-Biswas, F. Hayot, and C. Jayaprakash. Stochasticity of gene products from transcriptional pulsing. *Phys Rev E*, 79, 2009.

[34] T. Jia and R. Kulkarni. Intrinsic noise in stochastic models of gene expression with molecular memory and bursting. *Physical Review Letters*, 106:058102, 2011.

[35] T. Jia and R. V. Kulkarni. Post-Transcriptional Regulation of Noise in Protein Distributions during Gene Expression. *Physical Review Letters*, 105(1):018101, June 2010.

[36] Y. Jia, W. Liu, A. Li, L. Yang, and X. Zhan. Intrinsic noise in post-transcriptional gene regulation by small non-coding rna. *Biophysical Chemistry*, 143(1-2):60 – 69, 2009.

[37] G. Joanny, J. Le Derout, D. Brechemier-Baey, V. Labas, J. Vinh, P. Regnier, and E. Hajnsdorf. Polyadenylation of a functional mRNA controls gene expression in Escherichia coli. *Nucleic Acids Research*, 35(8):2494–2502, Apr 2007.

[38] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet*, 6(6):451–64, 2005.

[39] 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):451–64, June 2005.

[40] S. Kar, W. T. Baumann, M. R. Paul, and J. J. Tyson. Exploring the roles of noise in the eukaryotic cell cycle. *Proceedings of the National Academy of Sciences*, 106(16):6471–6476, APR 21 2009.

[41] B. B. Kaufmann and A. van Oudenaarden. Stochastic gene expression: from single molecules to the proteome. *Curr Opin Genet Dev*, 17(2):107–12, 2007.

[42] T. B. Kepler and T. C. Elston. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys J*, 81(6):3116–36, 2001.

[43] I. Keren, N. Kaldalu, A. Spoering, Y. Wang, and K. Lewis. Persister cells and tolerance to antimicrobials. *FEMS Microbiology Letters*, 230(1):13–18, Jan 15 2004.

[44] M. Komorowski, J. Miekisz, and A. M. Kierzek. Translational Repression Contributes Greater Noise to Gene Expression than Transcriptional Repression. *Biophys J*, 96(2):372–384, 2009.

[45] E. Kussell and S. Leibler. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, 309(5743):2075–2078, SEP 23 2005.

[46] Y. Lan and G. A. Papoian. The interplay between discrete noise and nonlinear chemical kinetics in a signal amplification cascade. *J Chem Phys*, 125(15):154901, 2006.

[47] Y. Lan, P. G. Wolynes, and G. A. Papoian. A variational approach to the stochastic aspects of cellular signal transduction. *J Chem Phys*, 125(12):124106, 2006.

[48] S. Legewie, D. Dienst, A. Wilde, H. Herzel, and I. M. Axmann. Small rnas establish delays and temporal thresholds in gene expression. *Biophysical Journal*, 95(7):3232 – 3238, 2008.

[49] E. Levine and T. Hwa. Small RNAs establish gene expression thresholds. *Curr Opin in Microbiol*, 11(6):574–579, 2008.

[50] E. Levine, Z. Zhang, T. Kuhlman, and T. Hwa. Quantitative characteristics of gene regulation by small RNA. *PLoS Biology*, 5(9):e229, Sept. 2007.

[51] E. Levine, Z. Zhang, T. Kuhlman, and T. Hwa. Quantitative characteristics of gene regulation by small rna. *PLoS Biol*, 5(9):e229, 2007.

[52] K. Lewis. Persister Cells. In *Annual Review of Microbiology, Vol 64, 2010*, volume 64 of *Annual Review of Microbiology*, pages 357–372. 2010.

[53] L. Liu, B. R. K. Kashyap, and J. G. C. Templeton. On the GIX/G/Infinity system. *Jour. Appl. Prob.*, 27(3):671–683, 1990.

[54] The result given in Ref. [53] has a minor error which is corrected here.

[55] R. Losick and C. Desplan. Stochasticity and cell fate. *Science*, 320(5872):65–68, Apr 4 2008.

[56] R. Losick and C. Desplan. Stochasticity and cell fate. *Science*, 320(5872):65–8, Apr. 2008.

[57] H. Maamar, A. Raj, and D. Dubnau. Noise in gene expression determines cell fate in Bacillus subtilis. *Science*, 317(5837):526–529, Jul 27 2007.

[58] N. Maheshri and E. K. O'Shea. Living with noisy genes: how cells function reliably with inherent variability in gene expression. *Annu Rev Biophys Biomol Struct*, 36:413–34, 2007.

[59] H. H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A*, 94(3):814–9, 1997.

[60] E. McCullagh, J. Farlow, C. Fuller, J. Girard, J. Lipinski-Kruszka, D. Lu, T. Noriega, G. Rollins, R. Spitzer, M. Todhunter, and H. El-Samad. Not all quiet on the noise front. *Nat. Chem. Biol.*, 5(10):699–704, 2009.

[61] J. Medhi. *Stochastic Models in Queueing Theory, Second Edition.* Academic Press, Inc., 2002.

[62] P. Mehta, S. Goyal, and N. S. Wingreen. A quantitative comparison of sRNA-based and protein-based gene regulation. *Molecular Systems Biology*, 4(221):221, Jan. 2008.

[63] P. Mehta, S. Goyal, and N. S. Wingreen. A quantitative comparison of sRNA-based and protein-based gene regulation. *Mol Sys Biol*, 4, 2008.

[64] J. T. Mettetal, D. Muzzey, J. M. Pedraza, E. M. Ozbudak, and A. van Oudenaarden. Predicting stochastic gene expression dynamics in single cells. *Proc Natl Acad Sci U S A*, 103(19):7304–9, 2006.

[65] N. Mitarai, A. M. Andersson, S. Krishna, S. Semsey, and K. Sneppen. Efficient degradation and expression prioritization with small rnas. *Phys Biol*, 4(3):164–71, 2007.

[66] N. Mitarai, A. M. C. Andersson, S. Krishna, S. Semsey, and K. Sneppen. Efficient degradation and expression prioritization with small RNAs. *Physical Biology*, 4(3):164–71, Sept. 2007.

[67] J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman. Single-cell proteomic analysis of S-cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[68] 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.

[69] J. Paulsson. Summing up the noise in gene networks. *Nature*, 427(6973):415–8, 2004.

[70] J. M. Paulsson. Models of stochastic gene expression. *Phys Of Life Rev*, 2(2):157–175, 2005.

[71] J. Peccoud and B. Ycart. Markovian modeling of gene-product synthesis. *Theoretical Population Biology*, 48(2):222 – 234, 1995.

[72] J. M. Pedraza and J. Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, 319(5861):339–43, 2008.

[73] V. Pfeiffer, K. Papenfort, S. Lucchini, J. C. D. Hinton, and J. Vogel. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nature Structural & Molecular Biology*, 16(8):840–U63, Aug 2009.

[74] T. Platini, T. Jia, and R. V. Kulkarni. Regulation by small RNAs via coupled degradation: mean-field and variational approaches. *ArXiv e-prints*, Feb. 2011.

[75] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS Biol*, 4(10):e309, 2006.

[76] A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, Oct 2008.

[77] A. Raj and A. van Oudenaarden. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, 135(2):216, 2008.

[78] A. Raj and A. van Oudenaarden. Single-Molecule Approaches to Stochastic Gene Expression. *Ann. Rev. Biophys.*, 38:255–270, 2009.

[79] C. V. Rao, D. M. Wolf, and A. P. Arkin. Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231–7, 2002.

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

[81] T. Romeo. Global regulation by the small rna-binding protein csra and the non-coding rna molecule csrb. *Mol. Microbiol.*, 29(6):1321–1330, 1998.

[82] M. Ronen, R. Rosenberg, B. Shraiman, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences, USA* , 99(16):10555–10560, 2002.

[83] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–5, 2005.

[84] S. M. Ross. *Introduction to Probability Models, Ninth Edition.* Academic Press, Inc., 2006.

[85] M. Sasai and P. G. Wolynes. Stochastic gene expression as a many-body problem. *Proc Natl Acad Sci USA*, 100(5):2374–9, 2003.

[86] V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proc Natl Acad Sci USA*, 105(45):17256–17261, 2008.

[87] O. Shalem, O. Dahan, M. Levo, M. R. Martinez, I. Furman, E. Segal, and Y. Pilpel. Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Molecular Systems Biology*, 4, 2008.

[88] A. Singh, B. Razooky, C. D. Cox, M. L. Simpson, and L. S. Weinberger. Transcriptional Bursting from the HIV-1 Promoter Is a Significant Source of Stochastic Noise in HIV-1 Gene Expression. *Biophysical Journal*, 98(8):L32–L34, Apr 21 2010.

[89] R. Skupsky, J. C. Burnett, J. E. Foley, D. V. Schaffer, and A. P. Arkin. Hiv promoter integration site primarily modulates transcriptional burst size rather than frequency. *PLoS Comput Biol*, 6(9):e1000952, 09 2010.

[90] P. S. Swain. Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J Mol Biol*, 344(4):965–76, 2004.

[91] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci U S A*, 99(20):12795–800, 2002.

[92] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci U S A*, 98(15):8614–9, 2001.

[93] M. Thattai and A. van Oudenaarden. Attenuation of noise in ultrasensitive signaling cascades. *Biophys J*, 82(6):2943–50, 2002.

[94] S. Vasudevan, Y. Tong, and J. Steitz. Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931–4, Dec. 2007.

[95] O. Vukmirovic and S. Tilghman. Exploring genome space. *Nature*, 405(6788):820–822, June 15 2000.

[96] L. S. Waters and G. Storz. Regulatory RNAs in bacteria. *Cell*, 136(4):615–28, Feb. 2009.

[97] L. Weinberger, J. Burnett, J. Toettcher, A. Arkin, and D. Schaffer. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell*, 122(2):169–182, Jul 29 2005.

[98] D. Wolf, V. Vazirani, and A. Arkin. Diversity in times of adversity: probabilistic strategies in microbial survival games. *Journal of Theoretical Biology*, 234(2):227–253, May 21 2005.

[99] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767):1600–3, 2006.

[100] Z. Zhang, W. Qian, and J. Zhang. Positive selection for elevated gene expression noise in yeast. *Molecular Systems Biology*, 5(1):299, Jan. 2009.

[101] V. P. Zhdanov. Bistability in gene transcription: Interplay of messenger RNA, protein, and nonprotein coding RNA. *Biosystems*, 95(1):75–81, 2009.