Stony Brook University

Noise and Oscillations in Simple Gene Networks

A Dissertation Presented

by

David Lepzelter

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Physics

Stony Brook University

December 2009

Stony Brook University

The Graduate School

David Lepzelter

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Jin Wang – Dissertation Advisor
Professor, Department of Physics and Astronomy

Peter Stephens – Chairperson of Defense
Professor, Department of Physics and Astronomy

Robert Shrock
Professor, Department of Physics and Astronomy

John Reinitz
Professor, Department of Applied Mathematics and Statistics
Stony Brook University

This dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Noise and Oscillations in Simple Gene Networks

by

David Lepzelter

Doctor of Philosophy

in

Physics

Stony Brook University

2009

Gene networks are a subject of increasingly intense study. Understanding the means by which organisms regulate their cells and the basic production mechanisms of cells is of immense importance. However, the means for studying these systems are still being developed. Monte Carlo simulations are common and accurate, but tend to make answering important and overarching questions impractical. This thesis examines other ways to look at the solutions to the stochastic equations which are involved in gene networks. It examines the intrinsic noise in these networks, and the important theme of coherent biological oscillation.

To my parents, Cary and Teresa.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor, Professor Jin Wang, for his support over these past few years.

Also, this work would not have been possible without any number of friends who have helped to keep me sane (at least, relatively speaking) in my time as a graduate student. Special thanks go to my fellow students in Jin's lab group, including those who ended up elsewhere.

Last but certainly not least, I would like to thank Pat Peiliker for looking out for me and helping me get through all the parts of the program that I wouldn't have had a clue about without her. She is a magician, and possibly a saint (though I doubt I'll get a chance to ask the Pope for confirmation of that last part).

Chapter 1

Introduction

1.1 Objective

The aim of the research project described in this doctoral thesis is to study the roles of intrinsic noise and oscillation in specific genetic networks, and to explore simple methods for examining these networks. Care is taken in selecting simple systems, with relatively few independent variables or parameters, in order to characterize the still-complex behavior of those systems with reasonable completeness.

Specifically, four systems are chosen for study: the Bicoid protein in *Drosophila melanogaster*, the two-gene toggle switch, the three-gene repressilator, and the single-gene self-repressor. These have all been studied previously, but not in the ways or parameter regimes mentioned in this work.

1.2 Genetic Networks

Genetic networks are central to life as we currently understand it. Every organism on Earth shares a general means of taking the information encoded in DNA and using it to carry out the processes it needs to perform in order to live. These processes are extremely diverse; they include everything from the basic breakdown of glucose for energy to cell reproduction to even more complex behavior like neuron firing. All of these processes depend to some degree on the external environment, but there can be no question that the role of genes and the proteins they encode is vital to every one.

Many of the mechanisms for these kinds of functions involve complex networks of interacting proteins and genes. Current research contains a great deal of data on individual genes, but often the overall function of these genes in a cell, through those networks, is poorly understood. The basic processes by

which genes and proteins interact are known. In general, the enzyme RNA polymerase synthesizes mRNA from genes, a process called transcription. Ribosomes then synthesize proteins from the mRNA, a step called translation. Some proteins can then bind to regions of genes called promoters in order to either increase or decrease the rates of mRNA synthesis. While the basic interactions seem simple, complex behavior is quite possible even without additional considerations; a small network of interacting genes and proteins can create a wide range of interesting and useful behavior, and an even wider range of behavior which can be harmful to an organism.

Modeling genetic networks has been a focus of a significant amount of research. Traditionally, researchers have used chemical kinetic equations to represent the relevant interactions. However, while conventional chemical kinetic equations work well under bulk conditions, they do not always give accurate results in the cell. Large statistical fluctuations can be caused by the relatively small number of molecules involved (often hundreds or thousands); these fluctuations are referred to as "intrinsic noise." [2–8].

This noise can sometimes cause surprising behaviors. There are several means by which one can study intrinsic noise, but by far the most useful involve the master equation formulation of chemical dynamics.

1.3 Master Equations

Master equation formulations are built on a simple set of ideas. The probability of n molecules of a given type existing is referred to as $P(n)$. This probability changes in time in ways which account for different processes that may occur: for instance, one of the molecules (e.g., a protein) in the system may be degraded, or another may be synthesized. If the rate of degradation is k and the rate of synthesis is g, this leads to a master equation of the form

$$\frac{dP(n)}{dt} = g\left(P(n-1) - P(n)\right) + k\left((n+1)P(n+1) - nP(n)\right). \qquad (1.1)$$

Noise, in this case, is simply the possibility that, in a given system, the n of the system is significantly different from its expected (mean) value, $< n >= \sum_n nP(n)$. It is important to note the "Master equation" actually represents an infinite number of equations, one for each possible value of n from 0 to ∞. Even so, at this point the system is quite uncomplicated and can in fact be solved analytically. However, if the genetic state is considered (whether the gene is synthesizing at the maximum rate or not, "on" or "off"), the least

complicated version of the equations is given by

$$\frac{dP_{\text{on}}(n)}{dt} = g\left(P_{\text{on}}(n-1) - P_{\text{on}}(n)\right) + k\left((n+1)P_{\text{on}}(n+1) - nP_{\text{on}}(n)\right), \quad (1.2)$$

$$\frac{dP_{\text{off}}(n)}{dt} = g\left(P_{\text{off}}(n-1) - P_{\text{off}}(n)\right) + k\left((n+1)P_{\text{off}}(n+1) - nP_{\text{off}}(n)\right). \quad (1.3)$$

doubling the complication of the system even without consideration of how the gene state switches from "on" to "off" or vice versa. The simplest switching is still solvable analytically, though this can be a difficult problem. It quickly becomes intractable, however, when one considers the possibility of proteins binding to their own genes' promoters (except in very specific cases, e.g. the monomer self-repressor [9]) or multiple kinds of proteins interacting with each others' genes.

To this problem, there are two possible answers: numerical solutions or approximations (or both). Numerical solutions are relatively common, with stochastic Monte Carlo simulations being especially prevalent partly because the Master equations already tend to assume processes are Markov. Approximations are somewhat less common unless one considers the use of bulk chemical equations instead of master equation formulations. Both methods are used in this dissertation.

1.4 Specific Systems

One of the systems we examine in this work requires little approximation and can be solved analytically. For the other three systems, we use a set of approximations to simplify the mathematics to a point where Monte Carlo simulations are no longer necessary for solutions. This is particularly important as explorations of all three of these systems involve extensive parameter searches, very calculation-intensive even without Monte Carlo simulation and almost prohibitively so with such simulation. The last system, however, is also examined using Monte Carlo methods, and the similarities and differences between the two methods is discussed briefly.

The first of the four systems selected for study, Bicoid, involves an analytical solution for the statistical distribution (the most complete means of representing intrinsic noise) of protein number in a spatially non-homogeneous region. Noise in Bicoid has been heavily studied because its effects in the system are so small as to seem anomalous, given the observed fluctuations. It is also a useful study because theoretical spatially-dependent noise calculations (crucial to understanding genetic systems with significant spatial non-

homogeneity) are still in their infancy.

We analyze the second system, the toggle switch, numerically using a few assumptions and approximations regarding the basic form of the statistical distributions of proteins involved. The toggle switch is slightly more complex than Bicoid, involving two genes which mutually repress each other. It is known as a simple, usefully bistable system which stores a value for a specific choice made by a cell (e.g., live or die). This is the first system for which we explore the effects of a wide range of parameters on system behavior.

The third system, the repressilator, we examine in terms of oscillation in that system using the same kind of approximations and assumptions used in the toggle switch calculations. This is the largest system we study, with three genes, and is a representation of oscillation, a kind of behavior that organisms use as clocks. In the chapter devoted to the repressilator, we discuss the effects of noise on oscillation and other aspects of that oscillation.

We explore the fourth system, the self-repressor, numerically using averages in the same way that the toggle switch and repressilator are examined, and using stochastic simulations. The self-repressor is a common theme in regulatory genes, and in spite of its apparent simplicity it is not yet fully understood. In our study of it, we consider the roles of noise, oscillation, and cooperative binding of proteins.

Chapter 2

Bicoid

2.1 Introduction

The first of the systems we examine is *Drosophila melanogaster*, an excellent example of a system in which noise is important to an organism[1].

D. melanogaster is a common organism for genetic and developmental biology for several reasons. It is easy to perform experiments on, and a large amount of background knowledge exists on it. The geometry of the embryo, roughly ellipsoidal with the anterior-posterior axis being significantly longer than the other two axes, is simple and easily accounted for mathematically. Additionally, until about two hours into its development as an embyro, the organism lacks distinct cells; each embryo has a large number of nuclei, but there are no cell membranes to block the diffusion of proteins from one nucleus to another. This last piece of information makes *D. melanogaster* ideal for refining ideas of diffusion-related spatial pattern formation, and the associated noise, in an embryo.

Of the proteins and genes in the embryo, a few stand out for having large effects on development. One of these is the Bicoid protein. It is useful to observe for four main reasons. First, it seems to have a direct regulatory effect on many developmental genes [11, 12]. Second, its production is independent of the presence or absence of other zygotic proteins. Third, its average concentration level at any given spatial point is essentially constant in time for much of the blastoderm stage. Fourth, its concentration and effects on other proteins are very obviously subject to statistical fluctuations, which make the use of stochastics absolutely necessary for a realistic understanding of the sys-

[1]The data and ideas from this chapter, and much of the language in this chapter, were originally co-authored with Jin Wang. Reproduced in part with permission from [10]. Copyright 2008 American Physical Society.

tem [1, 13]. Specifically, the internal noise of the system, the variability due to finite numbers of proteins, has caused significant debate on how the embryo can so accurately determine the spatial location of the sudden jump in the concentration of a protein, called Hunchback, which is dependent on Bicoid concentration. The Hunchback gradient, in turn, is an important regulator of other zygotic proteins, and its spatial precision has been a matter of significant research [1, 13].

These aspects of the protein have inspired numerical calculations using implementations of the chemical master equation for the system [13]. Such calculations have often been in one spatial dimension, in part because there is an easily recognizable gradient in the anterior-posterior direction which has a definite effect on development. The dorsal-ventral axis, in contrast, has a much smaller Bicoid gradient, and therefore Bicoid's direct effect on dorsal-ventral development is less significant than its anterior-posterior effects.

Even these numerical calculations need some assumptions, however. We will show in this chapter that the same simple assumptions which make the problem calculable numerically or using field theory (as in [14]) also make an exact and straightforward analytic solution possible for the Bicoid probability distribution in one spatial dimension. We also offer arguments as to why the same methods should work in other geometries. This is more than simply a continuation of a trend away from the bulk average concentration calculations done in the past, though it is that as well; even with an exact solution already known from [14], this analysis is important because it significantly clarifies our understanding of the system and similar systems. It offers a simple global characterization of the system, as opposed to local approaches or field theoretic characterizations.

The basic assumptions of our approach involve the three processes which govern the protein's behavior. First, production of Bicoid (which is highly localized in the anterior of the embryo) is assumed to be stochastic in nature. Second, movement of the protein through the embryo is assumed to behave according to traditional (random-walk-type) stochastic diffusion. Third, Bicoid degradation is assumed to be a stochastic event, i.e. through a decay reaction $\mathrm{Bcd} \xrightarrow{k} \emptyset$.

These three assumptions lead to a spatial dependent chemical master equation, which is a complete description of the probabilities involved in the system assuming no other effects. The importance of the chemical master equation to a gene-protein network can be compared to that of the Schrödinger equation for an atom: it forms the fundamental basis for further detailed characterization. While the nonlinear chemical rate equations provide quantitative description of the cellular networks on the average level showing complex behavior, the

probabilistic description obeys the linear master equation. So the deterministic kinetics can be chaotic but the corresponding probabilistic description can be quite regular. While the chemical kinetics gives reasonable description in the bulk, the probabilistic description provides the foundation for the mesoscopic intra- (or, in the case of this chapter, inter-) cellular network. The chemical kinetics give the deterministic trajectories with probability one. The probabilistic description provides a distribution of the protein concentrations. In other words, knowing the probability distribution, one knows the weights of individual states in protein concentration space. It is in this sense we can call it a probabilistic landscape in protein concentration space.

Landscape concepts have been introduced to the biology community in the areas of molecular and developmental biology [15] and population dynamics [16, 17]. The landscape is quantified in the areas of protein dynamics [18] and protein folding [19] while the potential energy landscape is known a priori with quasi equilibrium assumptions. For the non-equilibrium cellular networks, the potential landscape is associated with the potential free energy of the system mapped out over possible states of the system. Though it is not known a priori, one can obtain it by finding the probabilistic distribution through solving the master equation. A generalized potential U corresponding to the probabilistic description P for the non-equilibrium networks can be defined as $U = -\ln P$ in analogy with the Boltzmann relationship in equilibrium statistical mechanics [20–27]. Once the landscape can be quantified this way, it can give a global characterization of the network, providing the weight distribution in the protein concentration space and quantifying the importance of each state (in terms of weight). The stability, robustness and function of the network can be now studied in a global and physical way from landscape perspectives [20–27].

2.2 Master Equation

When the concentration has spatial dependence such as the developmental process, the probability distribution in protein concentration space becomes a probabilistic functional of protein concentrations which themselves also depend on space. It is in that sense a statistical probabilistic field theory representation (field being the protein concentrations which depend on space). Therefore by solving the probabilistic functional, we can map out the spatial dependent landscape of the cellular network. This is crucial for unraveling the origin of stability, robustness, and function of spatially-dependent cellular networks.

It should be noted that one complicating factor generally not included in master equation calculations is external noise, which can represent anything

from environmental temperature fluctuations to diffusion from outside the embryo, and is not explicitly accounted for in this model. This chemical master equation is most easily expressed in terms of a vector, $\boldsymbol{n}$, whose components $\boldsymbol{n} = (n_0, n_{\Delta x}, n_{2\Delta x} \ldots) = (\{n_x\})$ correspond to the number of Bicoid proteins at evenly spaced spatial positions $x = 0, \Delta x, 2\Delta x, \ldots$ (which each represent a finite amount of space Δx assumed to be evenly mixed), with Δx constant and essentially arbitrary. The equation is [28, 29]

$$\begin{aligned} \frac{dP(\boldsymbol{n})}{dt} = \; & g\left(P(\boldsymbol{n}-\hat{0}) - P(\boldsymbol{n})\right) \\ & +k\textstyle\sum_x \left((n_x+1)P(\boldsymbol{n}+\hat{x}) - n_x P(\boldsymbol{n})\right) \\ & +D\textstyle\sum_{xy} \left((n_x+1)P(\boldsymbol{n}+\hat{x}-\hat{y}) - n_x P(\boldsymbol{n})\right), \end{aligned} \tag{2.1}$$

where $P(\boldsymbol{n})$ is the probability that number and position of proteins is described exactly by $\boldsymbol{n}$. g is the rate of protein generation, $\hat{0}$ is a unit vector in the 0 space (representing a single protein at the origin, spatial point 0), and the term multiplying g represents the process of a protein being created at the origin. k is the rate of degradation, $\hat{x}$ represents a single protein at point x, and the term multiplying k represents the protein decay at any spatial position. D is the finite-volume diffusion rate, and the term multiplying it gives diffusion from each spatial point to its neighbors. The sums over x are over all space $x = 0, \Delta x, 2\Delta x, \ldots$, and over y are all spatial neighbors of x ($y = x \pm \Delta x$).

2.3 Ansatz

The next step in this process would be to find a time independent steady-state solution, $\frac{dP(\boldsymbol{n})}{dt} = 0$ for all $\boldsymbol{n}$. It should be noted that the deterministic form of this problem can be easily solved; $0 = \frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} + g\delta(x) - kC$ yields $C(x) = (g/\sqrt{kD})e^{-x\sqrt{k/D}}$. This corresponds to the reaction diffusion equation and its associated solution, often used in bulk studies. However, the uncertainties in concentration due to the finite number of molecules can only be found by solving the master equation. While the master equation itself does not immediately suggest a solution, the assumptions made do strongly suggest the use of Green's function techniques often encountered in physics and chemistry. Each individual protein has no interactions of any kind with any other protein; its creation, diffusion, and decay are all completely independent of any other effects. Therefore, we propose an ansatz in a format slightly different from that of the master equation,

$$P = \sum_{n=0}^{\infty} \frac{e^{-g/k}(g/k)^n}{n!} \prod_{m=1}^{n} G(x_m), \tag{2.2}$$

where n is the total number of proteins present in the system, m is a representation of each protein in the system, and $G(x_m)$ is actually a multidimensional generating function describing the chance that protein m is at the spatial point x_m. One can understand the probability expression above as the decomposition of the generation functions in Poisson space.

In order to prove the validity of the ansatz, we must first match its form more closely with the notation used in the master equation. Let us consider the spatial point x. For any given total number of proteins n, there are n proteins each with probability distribution G. Let G_x be the discrete version of $G(x)$. Then for a given n, the probability of n_x proteins existing at point x should be $P(n_x) = \begin{pmatrix} n \\ n_x \end{pmatrix} (G_x)^{n_x}$.

Combining this with the simple Poisson probability of n proteins existing, we find

$$P(\boldsymbol{n}) = \prod_{x}^{\text{all space}} \frac{e^{-gG_x/k}(gG_x/k)^{n_x}}{n_x!}. \tag{2.3}$$

We note that, while we implicitly used a vector $\boldsymbol{n}$ which began at the spatial point 0, the solution takes the form of Eq. 2.3 for other geometries as well. We also note that the form of the solution for the probability of Bicoid concentration $P(\boldsymbol{n})$ is simply that of a Poisson distribution with average value gG_x/k for each point in space, without spatial correlations. This is suggested, but not explored in detail, by a solution to a different problem in [30]. Others (e.g. [14]) state a Poisson solution, but we believe that this approach offers a useful contribution to the understanding of the problem because it is relatively simple and straightforward.

Given the form of the ansatz, we will define the Poisson distribution for the point x, $\mathcal{P}_x(n_x) = \frac{e^{-gG_x/k}(gG_x/k)^{n_x}}{n_x!}$, and note that $\mathcal{P}_x(n_x{+}1) = \frac{e^{-gG_x/k}(gG_x/k)^{(n_x+1)}}{(n_x+1)!} = \frac{gG_x/k}{n_x+1}\mathcal{P}_x(n_x)$.

Then inserting the ansatz into the master equation,

$$\frac{dP(\boldsymbol{n})}{dt} = \begin{bmatrix} g\left(\frac{n_0}{gG_0/k} - 1\right) \\ +k\sum_x \left((gG_x/k) - n_x\right) \\ +D\sum_{xy}\left((n_y\frac{G_x}{G_y}) - n_x\right) \end{bmatrix} \prod_{x'}^{\text{space}} \mathcal{P}_{x'}(n_{x'}).$$

Using $\sum_x G_x = 1$, and rearranging a sum,

$$\frac{dP(\boldsymbol{n})}{dt} = \begin{bmatrix} \frac{kn_0}{G_0} - g + k\frac{g}{k} - k\sum_x n_x \\ +D\sum_{xy}\left((n_x\frac{G_y}{G_x}) - n_x\right) \end{bmatrix} \prod_{x'}^{\text{space}} \mathcal{P}_{x'}(n_{x'}).$$

Again, all space in this geometry is $x = 0, \Delta x, 2\Delta x, \ldots$, and the neighbors x are $y = x \pm \Delta x$, except at $x = 0$ where y can only be Δx. Therefore

$$\frac{dP(\boldsymbol{n})}{dt} = \begin{bmatrix} \frac{kn_0}{G_0} - kn_0 + Dn_0\left(\frac{G_{\Delta x}}{G_0} - 1\right) \\ -k\sum_{x=1}^{\infty} n_x\left[1 + \frac{D}{k}\left(\frac{G_{x+\Delta x}}{G_x}\right.\right. \\ \left.\left. + \frac{G_{x-\Delta x}}{G_x} - 2\right)\right] \end{bmatrix} \prod_{x'=0}^{\infty} \mathcal{P}_{x'}(n_{x'}). \tag{2.4}$$

Since we are interested in the steady state solution, we solve for $\frac{dP(\boldsymbol{n})}{dt} = 0$. As n_x can in theory be any finite number, to ensure that the right-hand side of Eq. 2.4 is 0 we must ensure that the coefficients of each n_x are 0.

$$\frac{k}{G_0} - k + D\left(\frac{G_{\Delta x}}{G_0} - 1\right) = 0.$$

$$-k + D\left(\frac{G_{x+\Delta x}}{G_x} + \frac{G_{x-\Delta x}}{G_x} - 2\right) = 0, \;\; x > 0.$$

Defining for convenience $z \equiv \left(1 + \frac{k}{2D} - \sqrt{\frac{k^2}{4D^2} + \frac{k}{D}}\right)$, the solution,

$$G_0 = 1 - z,$$

$$G_x = zG_{x-\Delta x} = z^{x/\Delta x}(1 - z) = (1 - z)e^{\ln(z)x/\Delta x},$$

is simple. Since the mean of the distribution should be given by gG_x/k, it is reassuring to note that it corresponds to a decaying exponential function ($\ln z < 0$), the same form expected from both experiment and non-stochastic theory. It should be noted that this does not correspond exactly to the expected $e^{-x\sqrt{k/D}}$; this is because the definition of D is not precisely the same for finite-volume spaces, and because within each space the solution is assumed to be well-mixed. However, both of these issues can be avoided by using small enough distances between spatial points.

Substituting G_x into our formulation of $P(\boldsymbol{n})$, we obtain the final analytical expression for the probability:

$$P(\boldsymbol{n}) = \prod_{x}^{\text{all space}} \frac{e^{-gz^{x/\Delta x}(1-z)/k}(gz^{x/\Delta x}(1 - z)/k)^{n_x}}{n_x!}. \tag{2.5}$$

We note that, if this form is used in the original chemical master equation, it does in fact give $\frac{dP}{dt} = 0$, and therefore the ansatz is the correct and exact analytical solution to the steady- state problem.

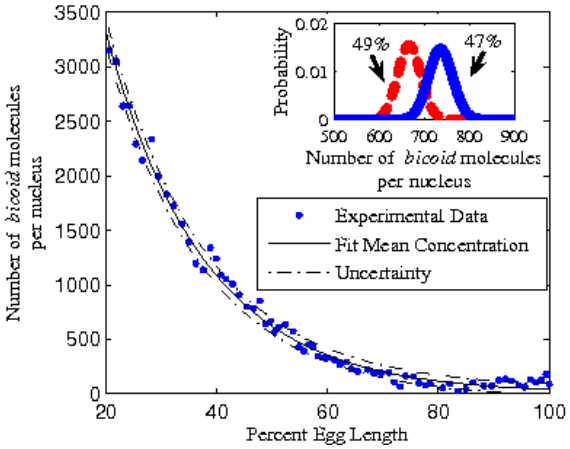

Figure 2.1: Calculation of expected distribution versus data, courtesy of Dr. Thomas Gregor, from an embryo. Error bars include intrinsic Poisson noise from proteins, photon counting noise, both of which are calculated from first principles, and a small constant gaussian noise intended to account for focal plane alignment. Errors from nuclear identification are not included. This fit gives $\chi^2/\text{dof}= 1.26$.

2.4 Experimental Data

Both the mean values and the noise given by this model, which decay exponentially from A to P, seem to match current experimental data (see Figs. 2.1, 2.2), with some caveats regarding the effective diffusion constant [1]. In both figures, the first 20% of the embryo is assumed to be part of a diffuse source of unknown local concentration and is therefore not considered part of the overall Green function fit. The main portion of Fig.2.1 has a line with predicted values, and two more with predicted uncertainties from both intrinsic and experimental noise.

The inset shows probability distributions with only intrinsic (non-experimental) noise for nuclei at 47% and 49% embryo length. In spite of significant overlap in the probability distributions of Bicoid concentrations, the embryo is generally capable of distinguishing on which side of the 48% embryo length boundary they fall. While the mechanism of such precision is not explored here, it is useful to know that the minimum reasonable noise, that of a Poisson distribution, can be considered correct and exact given the basic assumptions mentioned previously, and also fits with the experimentally observed noise.

These statistical fluctuations, given the Poisson form of the solution, are easy to calculate: $\sigma = \sqrt{gG_x/k} = \sqrt{ge^{\ln(z)x/\Delta x}(1-z)/k}$. We see that, since

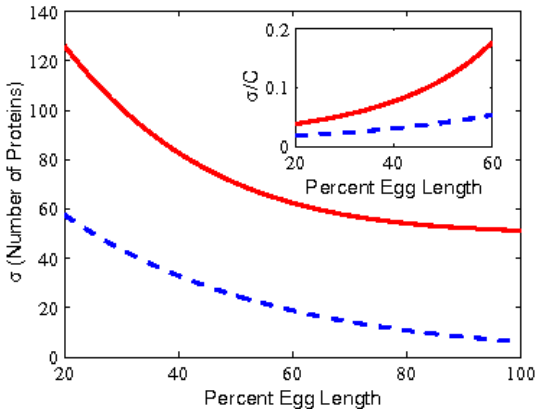

Figure 2.2: (Color online) Calculated noise from Fig. 2.1; dotted blue line shows intrinsic noise only, while solid red line shows both intrinsic and predicted experimental noise. Inset shows the predicted total experimental standard deviation divided by the mean, with dotted blue and solid red lines having the same meaning. Both solid lines follow roughly the trends as in [1], though without errors from nuclear identification they are somewhat smaller than the real experimental uncertainties.

$\ln z < 0$, the size of the fluctuation decays from A to P (clearly shown in Fig.2.2). Adding expected experimental noise from photon counting and focal plane alignment gives a larger (and no longer purely Poisson) noise. The inset shows fractional uncertainty, σ/C, where C is the number of Bicoid molecules, and the trend of increasing total (experimental plus intrinsic) fractional uncertainty from A to P agrees with experiment.

It should be noted that, even if later and even more detailed experiments should find that other distributions prove more realistic, the examined model should give valuable insight into the actual mechanisms in *D. melanogaster*: non-Poisson generation, non-monomer decay, or some other important process not previously mentioned would be vital in forming the shape of the distribution.

2.5 Discussion

While the precise mathematics have involved a one-dimensional problem with a source exactly at one end, it would not be difficult to prove the validity of the same kind of solution with a different geometry. Another boundary condition, a moved or spread-out source, and an additional dimension or two

should make it less easy to find the solution for G_x by hand, but the problem is not difficult with a computer. In any case, the validity of the general solution, with a Poisson distribution at every point in space, can be applied in any situation for which there are particles which diffuse, decay, and have one or multiple Markovian (Poisson-type) sources.

It important that, even though diffusion relates the concentration at one point in space with a concentration at another, it does not cause spatial correlations in this system. This is an important result because, while experimenters and theorists have always assumed Poisson-type intrinsic noise was the minimum possible, additional intrinsic noise and correlations have not previously been ruled out [13]. In this system, they do not exist because each protein's existence and location are independent of every other protein's existence and location. Spatial correlations may exist in cases where protein generation is non-Poisson, protein decay is non-monomer, or spatial transport does not have the traditional $\nabla^2 C$ form. Of these cases, this chapter's methods should be most easily generalized to non-Poisson protein generation.

Now we turn to the discussion of spatial landscape, a different way to view the probability distributions involved. We use generalized potential landscape $U = -\ln P$ to relate with the steady state probabilistic functional obtained by the exact solution of the spatial dependent master equation above. In Figure 2.3, we show the landscape in concentration and space. We can see from the bottom panel that the shape of the landscape at each spatial point is like a funnel with the bottom of lowest potential corresponding to the peak of the probabilistic distribution at that location. This is also clear from the two-dimensional representations of the potential versus protein number shown above the main graph at $20\%, 50\%$, and 80% egg length. The widths of the funnels are measured by the variances in potential at each spatial locations. A funneled landscape implies that the network is stable and robust. In this way, it can perform its biological function effectively and reliably. As we can see the funneled landscape becomes narrower from anterior to posterior. This implies varying stability and robustness distributed along spatial locations.

In summary, we used a relatively common model of protein production, diffusion, and degradation, to solve exactly and analytically for the stochastic distribution of the Bicoid protein in *Drosophila melanogaster*. The probabilistic solution is a Poisson distribution at each point in space, with the mean of the Poisson distribution decaying exponentially away from the source, and matches current experimental data well. The intrinsic fluctuations, noise due to a finite number of molecules in the system and which do not exist in the bulk, decrease away from the source at a slower rate than the mean. We also discussed how to uncover the underlying spatial landscape from the probabilistic

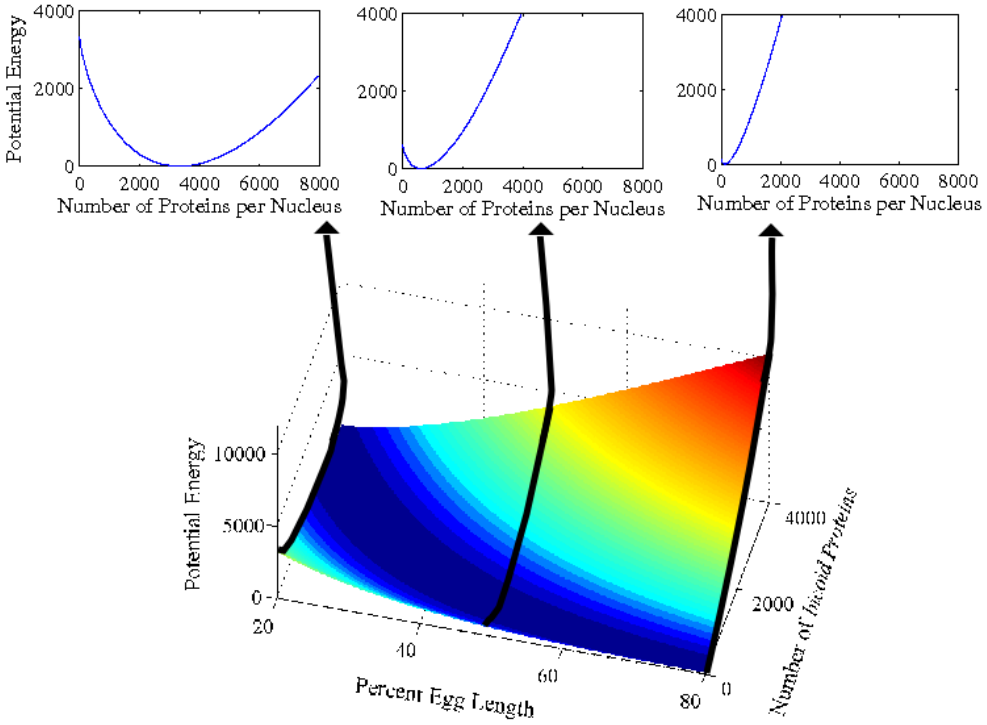

Figure 2.3: Potential versus number of proteins over space. The main graph shows the complete figure, in which each point in space (percent egg length) has its own potential energy function. Three of these are shown explicitly above the main graph, at 20, 50, and 80 percent egg length.

distribution. The landscape provides a global and physical foundation of quantitatively addressing the critical issues of stability, robustness and function of the spatial dependent cellular networks. The methodology used here can be easily generalized to more dimensions and different boundary conditions, and can be applied to any stochastic system with similar creation, diffusion, and decay processes.

Chapter 3

Toggle Switch

3.1 Introduction

Though we have begun to understand many things about genetic networks, there are many more unknowns[1]. In many cases, we still lack knowledge of how specific genes function in the presence of regulating genes and proteins, and how interactions of many genes can produce sometimes non-intuitive results. It is accepted that gene switches turning on and off control certain proteins' production, and that these protein products in turn act on genetic switches. The two processes often create a complicated network with many-body interactions and feedback loops. This complication makes the system difficult to study, but also sometimes provides surprising and useful behavior. Additionally, intrinsic noise in the systems can create both difficulties and new behavior in gene network patterns.

Complete and exact solutions to these problems are severely limited by system size in both number of types of proteins and number of each type of protein. To gain an understanding of a complicated gene network within the limits of current computing power, an efficient and accurate approximation scheme is necessary.

One such scheme is the Hartree mean field approximation, which significantly reduces the system's effective dimensionality. This scheme been applied to the toggle switch, in which two genes mutually repress each other, to find the steady state probabilities [32, 33].

This chapter has three aims. First, we explore the Hartree approximation and the moment equations that can be derived using it. Second, we examine

[1]The data and ideas from this chapter, and much of the language in this chapter, were originally co-authored with Keun-Young Kim and Jin Wang. Reproduced in part with permission from [31]. Copyright 2007 American Chemical Society.

the dynamics of the single molecule toggle switch and explore the steady-state properties as the long time limits of the dynamic evolution equations. Third, we study the statistical fluctuations, noise evolution and time scale to equilibrium using our developed time-dependent formalisms for the single molecule toggle switch [34].

3.2 Method and Materials

A discussion of the biological background, assumptions, and methods used in this and future chapters follows.

A "repressor" is a protein that binds to a gene to decrease the rate of transcription. In the toggle switch system, there are two genes, each of which produces via transcription a protein which acts as a repressor on the other gene. "Activators," which have the opposite effect, exist in many systems but are not relevant to the toggle switch, repressilator, or self-repressor, and so will be ignored here.

The situation is further simplified by the fact that only a finite number of proteins can bind to a given gene. In the toggle switch, both of the repressions are dimer (by which we mean that two proteins are involved; in the sense of the total number of molecules involved, the reaction is actually tri-molecular), and no additional repression is possible. Therefore, each of the two genes can be represented as having an "on" state, with no proteins bound to it, and an "off" state, with two proteins from the opposite gene bound to it. Binding is assumed to be highly cooperative, so intermediate states with single proteins bound to genes are assumed to be very short-lived and are ignored. The described reactions are given by

$$\mathcal{O}_A^{\text{on}} + 2B \underset{f_A}{\overset{h_A}{\rightleftharpoons}} \mathcal{O}_A^{\text{off}} \tag{3.1}$$

$$\mathcal{O}_B^{\text{on}} + 2A \underset{f_B}{\overset{h_B}{\rightleftharpoons}} \mathcal{O}_B^{\text{off}} \tag{3.2}$$

in which A and B represent the proteins that genes A and B, respectively, produce. $\mathcal{O}_A^{\text{on}}$ is equivalent to, in other common notation, DNA_A, and $\mathcal{O}_A^{\text{off}}$ is $\text{DNA}_A{\cdot}2B$. Therefore, $\mathcal{O}_A^{\text{on}}$ is gene A without repressors bound to its regulation area, or in the "on" state; $\mathcal{O}_A^{\text{off}}$ is gene A with two B proteins bound to it, or in the "off" state. Factors of 2 mean that the repression is a dimer reaction. The constant h represents the rate of binding, and f is the rate of unbinding.

In this chapter and the next, we also ignore the role of mRNA in the system, instead combining the transcription and translation processes into a

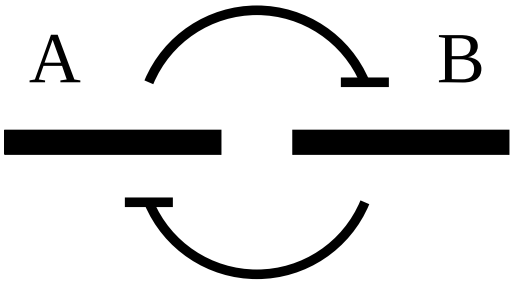

Figure 3.1: Illustration of the toggle switch. The flat-headed arrows represent repression via protein binding.

single stochastic process. Protein production can then be represented by

$$\mathcal{O}_A^{\text{on}} + \left(\begin{array}{c} \text{amino acids/} \\ \text{protein source} \end{array} \right) \xrightarrow{g_{A1}} \mathcal{O}_A^{\text{on}} + b_A A \tag{3.3}$$

$$\mathcal{O}_A^{\text{off}} + \left(\begin{array}{c} \text{amino acids/} \\ \text{protein source} \end{array} \right) \xrightarrow{g_{A0}} \mathcal{O}_A^{\text{off}} + b_A A. \tag{3.4}$$

where g_{1A} is the rate of generation when the gene is "on," g_{0A} is the rate of generation when the gene is "off," and b_A is the number of A proteins produced in a single "burst." The amino acid protein source is a general collection of basic materials for producing proteins which is assumed to be sufficient for production and is not explicitly modelled. Also, the system is symmetric, so the two equations with $A \to B$ are included as well. In all the studies done in this paper, $b_A = b_B = 1$. Larger (and sometimes variable) burst sizes can exist and are not equivalent to a simple rescaling of g.

Lastly, both kinds of proteins can degrade:

$$A \xrightarrow{k_A} \left(\begin{array}{c} \text{degradation byproducts/} \\ \text{protein sink} \end{array} \right) \tag{3.5}$$

$$B \xrightarrow{k_B} \left(\begin{array}{c} \text{degradation byproducts/} \\ \text{protein sink} \end{array} \right) \tag{3.6}$$

in which the protein sink, often represented in the literature by $\emptyset$, is a collection of byproducts of protein degradation, which are again not explicitly modeled. k_A is the decay rate of the A protein, which is often set to 1 to define a useful time scale, especially when the system has $k_A = k_B$. In all, the system is traditionally represented by Fig. 3.1, and is a useful model of several systems, including the bacteriophage λ[34, 35]. It is, in the chemical sense, an open driven system, as described in [36].

From these reactions, one can write down the master equation:

$$\begin{aligned}
\frac{dP_{1,1}(n_A,n_B)}{dt} = \; & -h_A n_B(n_B-1)P_{1,1}(n_A,n_B) \\
& -h_B(n_A(n_A-1)P_{1,1}(n_A,n_B) \\
& +f_A P_{0,1}(n_A,n_B) + f_B P_{1,0}(n_A,n_B) \\
& +(k_A(n_A+1)P_{1,1}(n_A+1,n_B) + k_B(n_B+1)P_{1,1}(n_A,n_B+1) \\
& -k_A(n_A)P_{1,1}(n_A,n_B) - k_B(n_B)P_{1,1}(n_a,n_B)) \\
& +g_{A1}(P_{1,1}(n_A-1,n_B) - P_{1,1}(n_A,n_B)) \\
& +g_{B1}(P_{1,1}(n_A,n_B-1) - P_{1,1}(n_A,n_B))
\end{aligned} \tag{3.7}$$

$$\begin{aligned}
\frac{dP_{0,1}(n_A,n_B)}{dt} = \; & h_A n_B(n_B-1)P_{1,1}(n_A,n_B) \\
& -h_B(n_A(n_A-1)P_{0,1}(n_A,n_B) \\
& + - f_A P_{0,1}(n_A,n_B) + f_B P_{0,0}(n_A,n_B) \\
& +(k_A(n_A+1)P_{0,1}(n_A+1,n_B) + k_B(n_B+1)P_{0,1}(n_A,n_B+1) \\
& -k_A(n_A)P_{0,1}(n_A,n_B) - k_B(n_B)P_{0,1}(n_a,n_B)) \\
& +g_{A0}(P_{0,1}(n_A-1,n_B) - P_{0,1}(n_A,n_B)) \\
& +g_{B1}(P_{0,1}(n_A,n_B-1) - P_{0,1}(n_A,n_B))
\end{aligned} \tag{3.8}$$

$$\begin{aligned}
\frac{dP_{1,0}(n_A,n_B)}{dt} = \; & -h_A n_B(n_B-1)P_{1,0}(n_A,n_B) \\
& +h_B(n_A(n_A-1)P_{1,1}(n_A,n_B) \\
& +f_A P_{0,0}(n_A,n_B) - f_B P_{1,0}(n_A,n_B) \\
& +(k_A(n_A+1)P_{1,0}(n_A+1,n_B) + k_B(n_B+1)P_{1,0}(n_A,n_B+1) \\
& -k_A(n_A)P_{1,0}(n_A,n_B) - k_B(n_B)P_{1,0}(n_a,n_B)) \\
& +g_{A1}(P_{1,0}(n_A-1,n_B) - P_{1,0}(n_A,n_B)) \\
& +g_{B0}(P_{1,0}(n_A,n_B-1) - P_{1,0}(n_A,n_B))
\end{aligned} \tag{3.9}$$

$$\begin{aligned}
\frac{dP_{0,0}(n_A,n_B)}{dt} = \; & h_A n_B(n_B-1)P_{1,0}(n_A,n_B) \\
& +h_B(n_A(n_A-1)P_{0,1}(n_A,n_B) \\
& -f_A P_{0,0}(n_A,n_B) + f_B P_{0,0}(n_A,n_B) \\
& +(k_A(n_A+1)P_{0,0}(n_A+1,n_B) + k_B(n_B+1)P_{0,0}(n_A,n_B+1) \\
& -k_A(n_A)P_{0,0}(n_A,n_B) - k_B(n_B)P_{0,0}(n_a,n_B)) \\
& +g_{A0}(P_{0,0}(n_A-1,n_B) - P_{0,0}(n_A,n_B)) \\
& +g_{B0}(P_{0,0}(n_A,n_B-1) - P_{0,0}(n_A,n_B))
\end{aligned} \tag{3.10}$$

Some choices of notation would indicate that $\frac{h}{2}$ should be used instead of h in these equations, but as this is a simple redefinition we simply note the current choice of notation.

Such a set of equations, already difficult to deal with, are made more so by the fact that each one is doubly infinite; each $P_{1,1}(n_A,n_B)$, for instance, is an equation for a probability of n_A A proteins and n_B B proteins; both n_A and n_B range from 0 to ∞. One can reasonably cut the range off at some maximum values N_A and N_B, but even then the number of degrees of freedom in the

system scales as $4 \cdot N_A \cdot N_B$. However, it is possible to find a smaller, more manageable set of equations, whose degrees of freedom scale as $4 \cdot (N_A + N + B)$. A Hartree-type approximation, inspired by the same approximations which give electron wavefunctions in multi-electron atoms, considers the probability distribution for each type of protein separate from that of the other. Each type of protein has a mean-field type of effect on the other. It gives

$$\begin{aligned} \frac{dP_{A1}(n_A)}{dt} = \; & -h_A n_B (n_B - 1) P_{A1}(n_A) + f_A P_{A0}(n_A) \\ & + k_A (n_A + 1) P_{A1}(n_A + 1) - k_A n_A P_{A1}(n_A) \\ & + g_{A1} P_{A1}(n_A - 1) - g_{A1} P_{A1}(n_A) \end{aligned} \tag{3.11}$$

$$\begin{aligned} \frac{dP_{A0}(n_A)}{dt} = \; & h_A n_B (n_B - 1) P_{A1}(n_A) - f_A P_{A0}(n_A) \\ & + k_A (n_A + 1) P_{A0}(n_A + 1) - k_A n_A P_{A0}(n_A) \\ & + g_{A0} P_{A1}(n_A - 1) - g_{A0} P_{A1}(n_A) \end{aligned} \tag{3.12}$$

with two additional equations for which $A \leftrightarrow B$. In this notation, $P_{A1}(n_A)$ is the probability of gene A being in the "on" state and n_A A proteins existing, and $P_{A0}(n_A)$ is the probability of gene A being in the "off" state with n_A A proteins.

These equations, while much simpler than the master equation without approximation, are still difficult to use. Therefore, instead of attempting to solve them exactly, we use moment equations, given by

$$\sum_{n_A} (n_A)^m \frac{dP_{A1}(n_A)}{dt} = \sum_{n_A} (n_A)^m \begin{pmatrix} -h_A n_B (n_B - 1) P_{A1}(n_A) \\ + f_A P_{A0}(n_A) \\ + k_A (n_A + 1) P_{A1}(n_A + 1) \\ - k_A n_A P_{A1}(n_A) \\ + g_{A1} P_{A1}(n_A - 1) \\ - g_{A1} P_{A1}(n_A) \end{pmatrix} \tag{3.13}$$

$$\sum_{n_A} (n_A)^m \frac{dP_{A0}(n_A)}{dt} = \sum_{n_A} (n_A)^m \begin{pmatrix} h_A n_B (n_B - 1) P_{A1}(n_A) \\ - f_A P_{A0}(n_A) \\ + k_A (n_A + 1) P_{A0}(n_A + 1) \\ - k_A n_A P_{A0}(n_A) \\ + g_{A0} P_{A0}(n_A - 1) \\ - g_{A0} P_{A0}(n_A) \end{pmatrix} \tag{3.14}$$

which, with some rearrangement and shifting of terms in the sum, become

$$\begin{aligned}\frac{d < n_{A1}^m >}{dt} = & -h_A < n_B(n_B-1) >< n_{A1}^m > + f_A < n_{A0}^m > \\ & + k_A(< n_{A1}(n_{A1}-1)^m - n_{A1}^{m+1} >) \\ & + g_{A1}(< (n_{A1}+1)^m - n_{A1}^m >)\end{aligned} \tag{3.15}$$

$$\begin{aligned}\frac{d < n_{A0}^m >}{dt} = & \; h_A < n_B(n_B-1) >< n_{A1}^m > - f_A < n_{A0}^m > \\ & + k_A(< n_{A0}(n_{A0}-1)^m - n_{A0}^{m+1} >) \\ & + g_{A0}(< (n_{A0}+1)^m - n_{A0}^m >)\end{aligned} \tag{3.16}$$

where $< n_{A1}^m >= \sum_{n_A}(n_A)^m P_{A1}(n_A)$ is essentially the probability that gene A is "on" multiplied by the mth moment of the number of A proteins in the system. $< n_B(n_B-1) >$ is, to be more precise, $< n_{B0}^2 - n_{B0} > + < n_{B1}^2 - n_{B1} >$, and again there are two identical equations with $A \leftrightarrow B$. Note that these moments can be added linearly, e.g. $< n_{A1}^x + n_{A1}^y >=< n_{A1}^x > + < n_{A1}^y >$, and that because in general $P_{A1}(n_A) \neq P_{A0}(n_A)$, both the probabilities of being in "on" and "off" states and the moments for the "on" and "off" states should not necessarily have any simple relation to each other.

In order to single out the most easily understood parameters, and in order to compare our results more easily to other results, it can be convenient to define the following:

$C_{A1} =< n_{A1}^0 >$,	the probability of gene A being in the "on" state.
$C_{A0} =< n_{A0}^0 >$,	the probability of gene A being in the "off" state.
$X_{A1} = \frac{< n_{A1}^1 >}{C_{A1}}$,	the average number of A proteins if the gene is in the "on" state.
$X_{A0} = \frac{< n_{A0}^1 >}{C_{A0}}$,	the average number of A proteins if the gene is in the "off" state.
$\omega = \frac{f}{k}$,	a measure of the relative speed of protein unbinding from the gene.$\omega >> 1$ is called the "adiabatic limit."
$X_{eq} = \frac{f}{h}$,	the ratio of the rates of protein unbinding from and binding to the gene. It is called the equilibrium constant.
$X_{ad} = \frac{g_1+g_0}{2k}$,	the rate of protein synthesis relative to protein self-degradation, called the adiabatic parameter. It generally encourages quick binding.

Of course, the same definitions can be used for the B gene and proteins as well.

In certain situations, moment equations can couple infinitely; e.g., the first moment can depend on the second, which depends on the third, and so on. Such a problem makes it necessary to make assumptions about higher-

order moments, which in effect becomes an ansatz for the system. In fact, this system has been studied using a Poisson ansatz[32]. However, because the toggle switch has no self-interaction, there is no need for an ansatz, and the moment equations can be solved without further approximations using a simple computer program. We did use the Poisson ansatz for a small number of calculations, but only for the sake of comparison between our own Poisson data and others or for comparing our normal moment equations to Poisson ansatz results. For the calculations, we wrote C programs for solving the moment equations, on a Dell Linux desktop which was more than adequate for the calculations.

3.3 Results and Discussions

The steady state equations based on the Poisson ansatz were derived previously [32] and the steady state moment equations were studied [33]. We solved the time-dependent dynamic equations directly and take the long time limit for the steady state solutions. Fig.3.2, corresponding to a figure given previously by [33], shows the probability that genes are in the active state, as a function of the adiabatic parameter $X_{ad} = \frac{g_1+g_0}{2k}$, where $g_{A1} = g_{B1} = g_1$ and so forth. Exact solutions of the moment equations are compared with the Poisson Ansatz solutions for a single molecule symmetric switch. We see that, in the toggle switch, there is a transition from mono-stability, when the synthesis rate of proteins is low, to bistability, when the synthesis rate is high. This result is consistent with previous studies [32, 33]. It can be explained in a relatively simple way. For low-g systems, there is very little repression, because even at maximum production the $hn(n-1)$ term is small. Both genes are therefore almost always in the "on" state, so the system naturally has a single steady state, with both genes "on." At higher synthesis rates, it is possible for either gene to synthesize enough proteins to consistently repress the other, leaving the repressing gene "on" and the repressed gene "off." Since two choices exist for which gene is "on," the system should be expected to have two stable states.

In the previous section, we discussed the Poisson ansatz. Qualitatively, we expect that the Poisson ansatz should be a good approximation for each protein in "on" and "off" states separately in the limit $\omega = \frac{f}{k} << 1$, where the "birth-death term" is dominant, because in this regime each gene state should be able to produce proteins almost independently of mixing effects (genes switching from on to off or vice versa). Cases in which genes are permanently in the "on" or "off" state have perfect Poisson distributions; small effects from mixing should produce slight perturbations into the Poisson distributions.

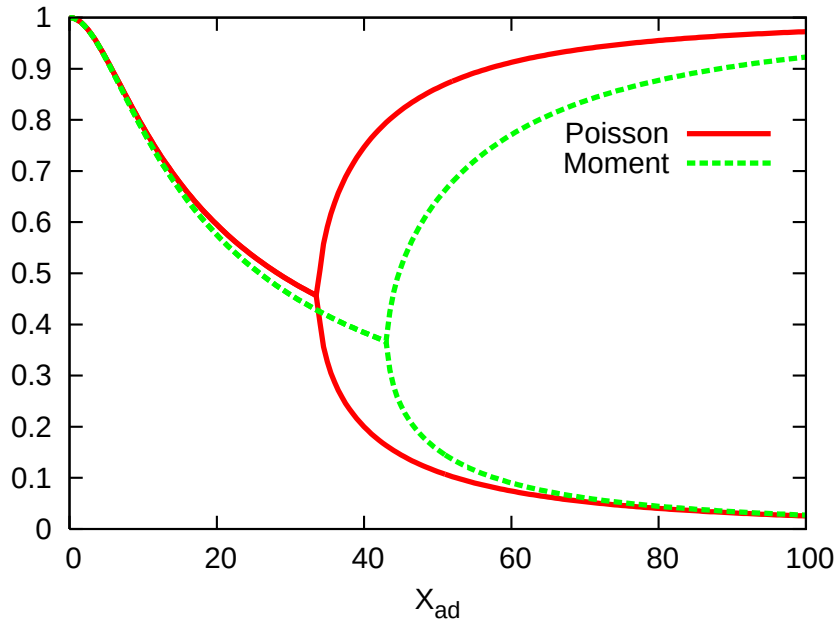

Figure 3.2: Probability C that genes are in the active state as a function of $X_{ad} = (g_1 + g_0)/2k$ for a symmetric switch. Exact moment equation solutions are compared with Poisson ansatz solutions, for a single symmetric switch, $X_{eq} = f/h = 1000$, and $\omega = f/k = 0.5$.

$f_A(=f_B)$	C_{A1}	C_{B1}	X_{A1}	X_{A0}	X_{B1}	X_{B0}	F_{A1}	F_{A0}	F_{B1}	F_{B0}
500	0.962	0.013	193	192	6.26	6.26	1.01	1.01	1.00	1.00
	0.962	0.013	193	192	6.26	6.26				
50	0.962	0.013	193	189	6.28	6.26	1.12	1.12	1.01	1.00
	0.962	0.013	193	189	6.29	6.26				
5	0.962	0.013	194	163	6.50	6.25	1.80	5.79	1.06	1.05
	0.962	0.013	194	163	6.51	6.26				
0.5	0.960	0.013	197	72.5	8.61	6.20	1.98	44.6	1.95	1.45
	0.962	0.013	198	72.5	8.63	6.21				
0.05	0.940	0.013	199	19.0	25.8	5.99	1.26	42.8	11.3	3.68
	0.958	0.013	200	19.0	25.5	5.98				
0.005	0.884	0.014	200	10.9	75.3	5.35	1.05	9.14	10.8	3.62
	0.905	0.014	200	10.9	74.8	5.35				
0.0005	0.859	0.014	200	10.1	98.6	5.05	1.01	1.89	2.47	1.42
	0.863	0.014	200	10.1	96.8	5.05				

Table 3.1: Asymmetric toggle switch with high synthesis rate: $k_A = k_B = 1$,$g_{A1} = 200$, $g_{B1} = 100$, $g_{A0} = 10$, $g_{B0} = 5$, $h_A = \frac{f_A}{500}$, $h_B = \frac{f_B}{250}$ $F(X_{A1})$ means Fano factor of (X_{A1}) etc. The first line of each f_A is based on moments equation, and the second line is based on Poisson ansatz.

This approximation should be best in the most probable gene state because the probabilities being introduced into the state via mixing should be small compared to the overall probabilities. Since we set $k_A = k_B = 1$, $\omega = f_A = f_B = f$ determines adiabaticity. Small f implies that the unbinding rate of the regulatory protein to DNA is slow compared with the degradation rate of protein synthesis.

Table 3.1 is the long time limit steady state results of a toggle switch. The first line for each f represents the results of the moment equation, using Eq(3.15), and the second line for each f shows the results from the Poisson ansatz. The last four columns refer to F, the Fano factor, which is defined as $\frac{\text{variance}}{\text{mean}}$ and would be equal to 1 if the distributions were exactly Poisson. Given the chosen initial conditions, with gene A activated and gene B repressed, the long time limit of the system will be that gene A is activated and gene B is repressed. It is the dominant distribution in each case that we expect to see the best agreement with the Poisson assumption. Therefore, in this parameter regime, we expect C_{A1} and C_{B0} to be large, X_{A1} and X_{B0} to be close to g_{A1} and g_{B0} respectively, and F_{A1} and F_{B0} to be close to 1.

Indeed, in the extreme non-adiabatic limit ($\omega = f = 0.0005$), this appears

to be the case. F_{A1} and F_{B0} are very close to 1, and the values of X_{A1} and X_{B0} agree with both the values of g_{A1} and g_{B0} and the values of X_{A1} and X_{B0} obtained from the Poisson ansatz. This strongly supports our interpretation of a basically Poisson distribution with a very small perturbative addition. F_{A0} and F_{B1} are larger but not extremely large, which suggests that the Poisson ansatz is somewhat but not entirely unrealistic in these two individual states; the distribution is slightly more spread out, likely with a fat tail on the high-n end. In a less mathematical sense, this should simply mean that in these states we should expect large differences between the mean value and any given single-molecule experiment result, even though the average value should still be the same. To the extent that the ansatz is not a good one, however, the overall effect on the system appears to be small; moment equation values for the C and X values for both proteins in both genetic states still agree with those acquired from the Poisson ansatz. Note, though, that the overall Fano factor for the combined probability distribution, which we get by adding the "on" and "off" states together, is not necessarily even close to 1 for either gene. In fact, it should be much larger, because the system is close to two Poisson distributions with different means added linearly.

As the adiabaticity increases with increasing $f(=\omega)$, agreement with the Poisson assumption grows worse for both proteins in both genetic states. At these values of ω, there is a large spread in the probability distributions, and they may have multiple peaks. Physically, this means that even if the "on" or "off" state could be isolated experimentally, the mean values calculated may not be representative of what we expect to measure in any individual single-molecule experiment. The increased spread in these distributions generally suggests that the means of the distributions are very poorly suited to describing the possible behaviors of the system, and that stochastic treatment is especially necessary for taking care of the fluctuations at these values of ω.

However, at still larger values of ω, above $\omega \sim 10$, the system becomes Poisson-like again. Also, these distributions have more similar means ($X_{A1} \approx X_{A0}$, $X_{B1} \approx X_{B0}$). Such behavior is explained by the fact that high binding and unbinding rates mix the "on" and "off" gene states so much that only a single scale can emerge from them, and gene A's protein production (for instance) is essentially a single Poisson process with a generation rate of $C_{A1}g_{A1} + C_{A0}g_{A0}$.

Using the same parameters as in Table 3.1 but decreasing the protein synthesis rates (g_A and g_B) gives Table 3.2. Smaller generation rates mean that genes which are already almost completely "on" should produce fewer proteins, which therefore should repress their target genes considerably less. This certainly occurs for the B gene, for which the fractional increase in C_{B1} from

$f_A(=f_B)$	C_{A1}	C_{B1}	X_{A1}	X_{A0}	X_{B1}	X_{B0}	F_{A1}	F_{A0}	F_{B1}	F_{B0}
500	0.968	0.249	38.8	38.7	5.75	5.74	1.00	1.00	1.01	1.01
	0.968	0.249	38.8	38.7	5.75	5.74				
50	0.968	0.249	38.8	38.1	5.81	5.72	1.02	1.03	1.06	1.06
	0.968	0.249	38.8	38.1	5.81	5.72				
5	0.965	0.250	38.9	32.7	6.42	5.52	1.15	1.95	1.52	1.54
	0.968	0.249	39.0	32.8	6.41	5.51				
0.5	0.947	0.251	39.3	14.4	10.5	4.18	1.26	9.71	3.15	4.02
	0.961	0.249	39.5	14.5	10.5	4.16				
0.05	0.916	0.255	39.8	3.80	17.7	1.79	1.08	9.35	2.00	4.67
	0.925	0.254	39.9	3.80	17.7	1.79				
0.005	0.908	0.256	40.0	2.19	19.7	1.09	1.01	2.63	1.13	1.80
	0.909	0.256	40.0	2.19	19.7	1.09				
0.0005	0.907	0.256	40.0	2.02	20.0	1.01	1.00	1.18	1.01	1.09
	0.907	0.256	40.0	2.02	20.0	1.01				

Table 3.2: Asymmetric toggle switch with low synthesis rate: $k_A = k_B = 1$,$g_{A1} = 40$, $g_{B1} = 20$, $g_{A0} = 2$, $g_{B0} = 1$, $h_A = \frac{f_A}{500}$, $h_B = \frac{f_B}{250}$. The first line of each f_A is based on moments equation, and the second line is based on Poisson ansatz.

Table 3.1 to Table 3.2 is enormous. However, with the increased B gene activity from Table 3.1 to Table 3.2 can come more B proteins, in spite of a decreased overall generation rate. Therefore, the effect of decreased generation rates on protein B's repression target, C_{A1}, is small and not always positive. In all cases, F_{B0} significantly increases because there is a larger non-Poisson probability distribution with which the B off-state mixes.

This does not yet explain the general decrease in Fano factors from Table 3.1 to Table 3.2 due to decreased synthesis, which occurs in the A protein and possibly in the B protein as well (though at best the effect is somewhat masked in the B proteins by the previously explained increase). The reasons for this may be similar to those mentioned for the bifurcation in Fig. 3.2; smaller synthesis rates should mean less likelihood of binding, and therefore less switching between "on" and "off" states. Less switching should imply less perturbation to essentially Poisson-like distributions, as described before.

From the time-dependent solution, we can estimate the time to reach equilibrium. Fig. 3.3 shows the time-evolution of the protein numbers for a toggle switch. The horizontal axis is time, and the time scale is defined by $k = 1$. Such a parameter choice makes the other time scales easier to put into context;

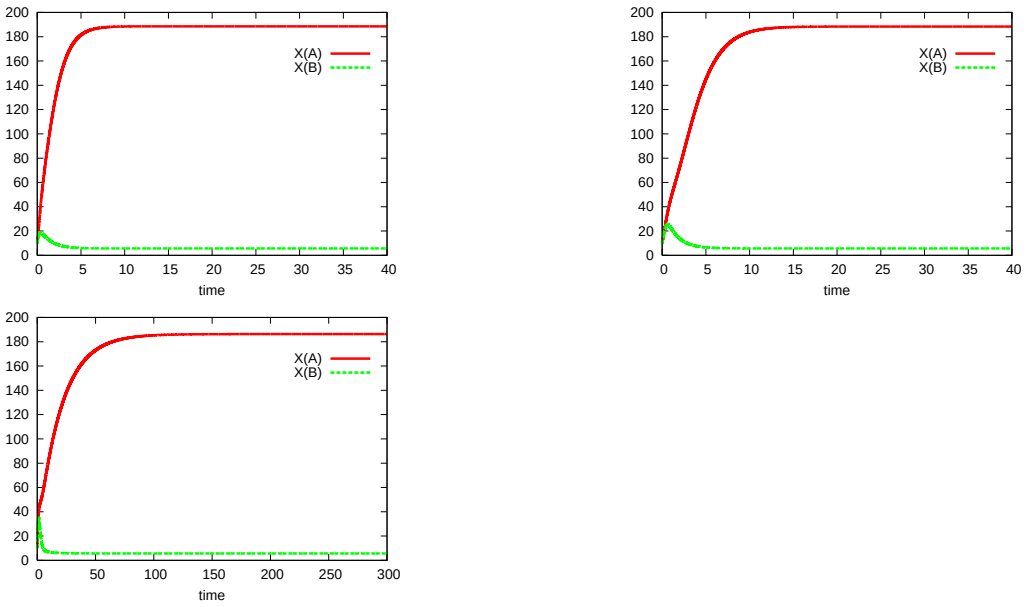

Figure 3.3: Time evolution of protein number X (toggle switch) for different unbinding rate of protein to DNA at given protein synthesis rate g, protein degradation rate k and binding rate of protein to DNA h (time is in units of the inverse of rate coefficients): $k_A = k_B = 1$,$g_{A1} = 200$, $g_{B1} = 100$, $g_{A0} = 10$, $g_{B0} = 5$, $h_A = \frac{f_A}{500}$, $h_B = \frac{f_B}{250}$. (a) $f_A = 5$ (b) $f_A = 0.5$ (c) $f_A = 0.05$.

smaller binding and unbinding rates mean that any individual protein is unlikely to bind before it decays, while larger binding and unbinding rates would mean each protein would likely be bound many times before it decays. We observe that it takes a longer time to reach the steady state in the small binding and unbinding rate limit, which would be expected from non-stochastic calculations as well. The binding and unbinding in this case is essentially a rate-limiting step because it is so slow compared to decay and synthesis, the only other processes involved in bringing the system to the steady state.

Fig. 3.4 is the time evolution of the Fano factors. This shows that the system is often more noisy with larger statistical fluctuations during the course to a steady state than at the steady state, which agrees with results from [6]. After peaking, the statistical fluctuations tend to decay with time to reach the steady state value. Smaller Fano factors might imply more stability, which would explain why the Fano factors would tend to jump immediately after the system is changed and gradually decrease as it moved towards stability. Furthermore, the Fano factors in time show the same trends as shown in Tables 3.1 and 3.2: both proteins in both genetic states have smaller Fano factors at $\omega = 5$ (Fig. 3.4(a)), and the dominant "on" state Fano factor (A,on) peaks

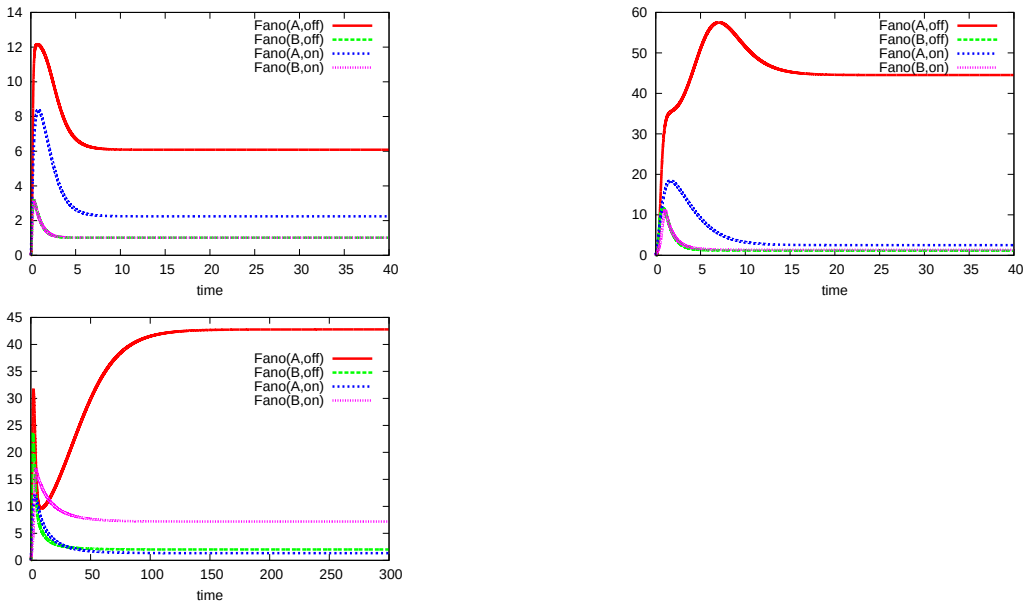

Figure 3.4: Time evolution of the Fano factor(toggle switch) F for different unbinding rate of protein to DNA at given protein synthesis rate g, protein degradation rate k and binding rate of protein to DNA h (time is in units of the inverse of rate coefficients): $k_A = k_B = 1$,$g_{A1} = 200$, $g_{B1} = 100$, $g_{A0} = 10$, $g_{B0} = 5$, $h_A = \frac{f_A}{500}$, $h_B = \frac{f_B}{250}$. (a) $f_A = 5$ (b) $f_A = 0.5$ (c) $f_A = 0.05$.

at $\omega \approx 0.5$ while the dominant "off" state Fano factor (B,off) peaks closer to $\omega \approx 0.05$.

Fig. 3.5 is a different approach to Fano factors, this time exploring the steady state values in the symmetric toggle switch ($f_A = f_B = f$, etc.) for a large range of ω, $X_{eq} = h/f$, and $X_{ad} = \frac{g_1+g_0}{2}$. For the sake of avoiding very small numbers and precision problems in division, we chose $g_0 = \frac{g_1}{20}$ instead of $g_0 = 0$, which would otherwise be more convenient for overall comparison with [33]. In (a) and (b) we give the Fano factor only for the off-state: (a) shows X_{eq} versus ω with X_{ad} held constant at 50, and (b) shows X_{ad} versus ω with X_{eq} held constant at 1000. As in Table 3.1, around $\omega = 1$ and somewhat lower the Fano factor grows significantly, suggesting significantly non-Poisson behavior. At high ω, however, it is approximately 1 because of the high rate of mixing between on- and off-states. Also, at very low ω, the system starts having some properties of adiabatiticity again; each genetic state ("on" or "off") can now behave almost independently of the other, as Poisson distributions with their respective generation rates, but with very slight perturbations from mixing. Note that low X_{ad} generally reduces the Fano factor, since fewer proteins imply less binding (and hence both that the "on" state dominates

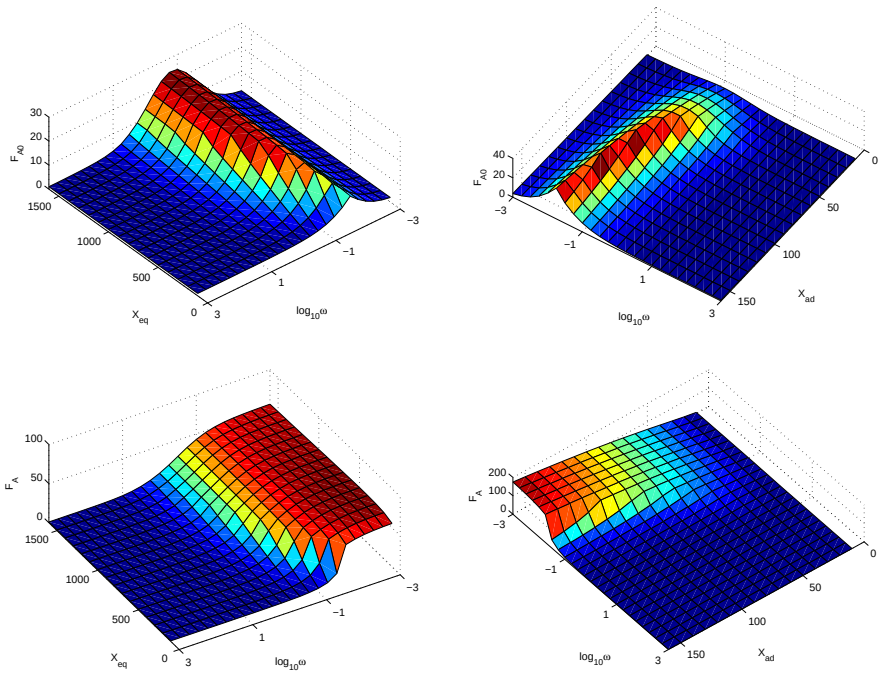

Figure 3.5: Fano factors in a symmetric toggle switch ($h_A = h_B = h$, etc). (a) and (b) show the off-state: (a) shows X_{eq} versus ω and (b) shows X_{ad} versus ω. (c) and (d) show the total Fano factor: (c) shows X_{eq} versus ω and (d) shows X_{ad} versus ω.

and that there is less switching), which means smaller perturbations to each individual state's Poisson behavior. Very low X_{eq}, however, yields greatly increased binding, which should mean that the state is almost always in the "off" state. Therefore, as X_{eq} decreases the "off" state tends towards Poisson-like behavior with Fano factor near 1. However, even with a very small X_{eq}, some mixing is inevitable because unbinding is a constant process. (c) and (d) show the total Fano factor, obtained by combining the on- and off-states: (c) shows X_{eq} versus ω and (d) shows X_{ad} versus ω. Again, at high ω rapid switching gives essentially a single Poisson distribution, and therefore a Fano factor of approximately 1. The factor again increases near $\omega = 1$, but does not decrease at low ω. This is because, while both on and off states are close to Poisson distributions individually, when added together they become a two-peaked system which has significant non-Poisson behaviors. Low X_{ad} again tends to decrease the Fano factor, since less binding means the "on" state dominates, and it is Poisson-like for reasons explained previously. Decreasing X_{eq} again means that the "off" state dominates more and therefore decreases the Fano factor.

In Fig. 3.6 (a) and (b), a characteristic time value given by $\int t|C(t) - C(\infty)|dt$ is plotted with ω versus X_{eq} and X_{ad} respectively, also in the symmetric toggle switch. There are two notable behaviors in the graphs: the overall trend and the spiked line running through the middle. The overall trend is that slow binding and unbinding lead to longer times to equilibrium, which is unsurprising given that even non-stochastic equations would respond in the same way due to the rate limiting effect of slow switching. The extremely high peaks in lines through the two graphs correspond well to the boundaries suggested by the crude phase diagrams in Fig. 3.6 (c) and (d), designed to distinguish unimodal probability distributions from multimodal distributions based on the behavior of moment equations. Though the methods we used were poorly suited to reconstructing actual probability distributions, the relevant transition line in the ω versus X_{ad} graph also seems to agree with the line between bistable and other kinds of systems using very similar parameters in [37]. We therefore suggest that these extremely unusual values are due to a second relaxation time scale, as demonstrated by Fig. 3.7, for which we propose an interpretation.

Our suggested explanation is based on Fig. 3.8 and the known phenomenon of transition between probability peaks in the bistable case. We assume the system begins with a probability distribution centered at point A. The shorter time scale, labelled (1), would be the time necessary to enter the probability peak centered at B, one of the two stable points of the bistable system. The longer time scale, (2), would be the time necessary for transition of half of

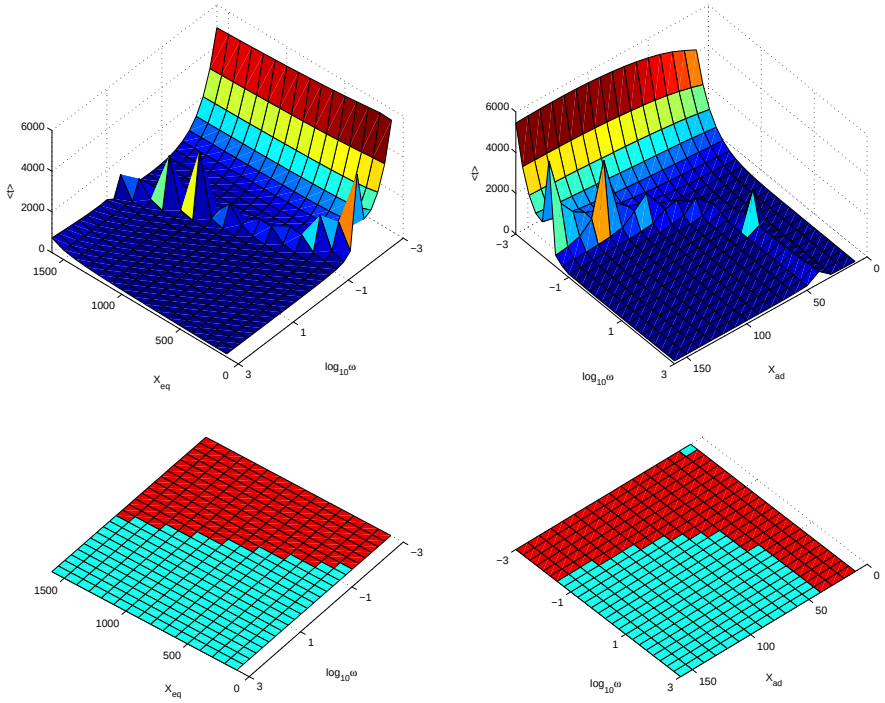

Figure 3.6: (a) and (b) give a measure of the amount of time the system takes to settle into its final state, with ω versus X_eq and X_ad respectively. (c) and (d) are crude phase diagrams based on the structure of the solutions and the possibility of obtaining a second (generally identical but mirror-imaged) solution.

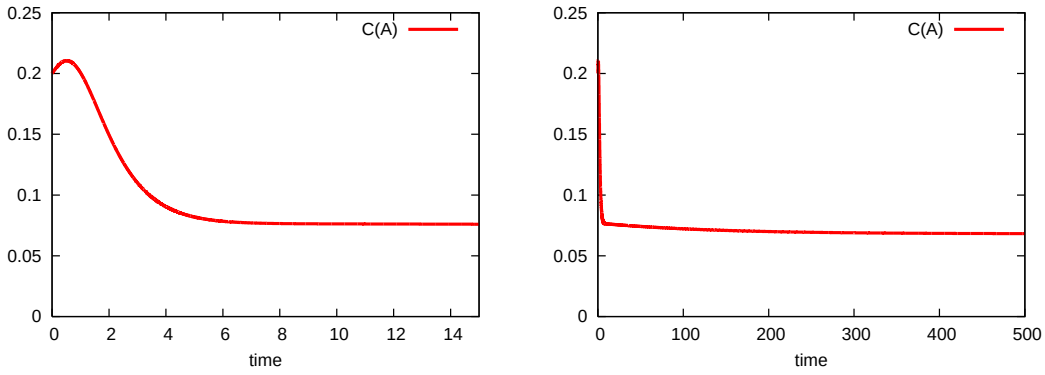

Figure 3.7: The same system ($\omega = 0.044, X_{ad} = 100, X_{eq} = 10^3$), near the phase transition demonstrated in Fig. 3.6, shown with two different time scales. If the apparent long-time limit from (a) were assumed to be the steady state, this would be incorrect, because there is a second and slower settling process shown in (b).

the probability in that peak to the other, at point C. This effect should be observable when the peaks are close to each other, but when the two states B and C are far enough away it should die off (time scale 2 approaches infinity and is not observable). When the two states have completely separated, the Hartree approximation's limitations should make it unobservable. This whould explain the bistable-monostable line; the bistable-tristable line, considerably wider than the bistable-monostable line, may actually be a slightly distorted view of the entire tristable area, with the second time scale representing the transition time between peaks at g_{0A}, g_{0B} and either g_{0A}, g_{1B} or g_{1A}, g_{0B}.

Alternatively, one may interpret the idea through the language of trajectories. Any individual system would follow the probability distributions shown in Fig. 3.8; it would start at A, quickly work its way to B, and possibly eventually work its way to C. In order to pass from B to C, it would first have to enter the lower-probability states between the two peaks, which essentially should serve as a bottleneck (larger distances between B and C would serve as long bottlenecks, and lower probabilities between B and C would give thin bottlenecks). When B and C are close to each other, the bottleneck should be both short and wide, giving a smaller time scale; when they are far apart, the bottleneck should be longer and thinner, giving a larger time scale. As B and C separate more, they should eventually approach infinity and become unobservable. While this does not truly account for all possible dynamics of the system (which should include a flux within probability peaks, even at steady state), we suggest it may be a reasonable approximate explanation.

The single molecule toggle switch shows bistability, and we can discuss the

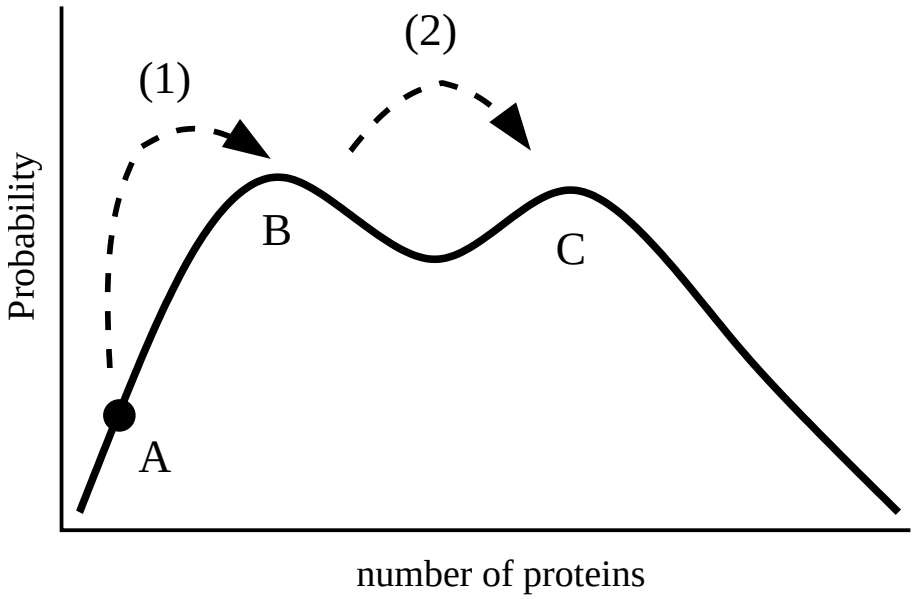

Figure 3.8: One-dimensional representation of a bistable system, with stable points B and C. If the system begins at A, the time necessary to go to peak B, labelled (1), should be small; the time necessary to go from B to C, (2), should be larger.

transition between two attractors. Gardner *et al* constructed a synthetic toggle switch and showed bi-stability [34], where they used chemical or thermal inducers. We can take into account this experimental inducer effect by considering the reaction probabilities as a function of time. The experimental set-up resembles step functions and we can implement it by using two hyperbolic tangent functions. Fig. 3.9 is one example of the transition between two fixed points. The first gray shading indicates the inducer effect on gene A (increased unbinding of protein B from gene A), and the second indicates the inducing on gene B (increased unbinding of protein A). We set the initial conditions such that, by $t = 100$, the system is at the one fixed point where protein B is abundant. If we induce production of protein A at $t = 100$, a transition to the other fixed point (A dominant) occurs. When we change the binding back to its previous value at $t = 150$, the state shifts position but still keeps A dominant. This means that there are (at least) two stable states which can exist in the same physical system. The same thing happens when we apply a second temporary induction, increasing unbinding of protein A from gene B and therefore increasing protein B production, from $t = 250$ to $t = 300$. The graph of interval ($100 \leq t \leq 400$) in Fig.3.9 qualitatively explains the earlier experimental findings (Figure 5 in [34]). This shows clearly the intrinsic

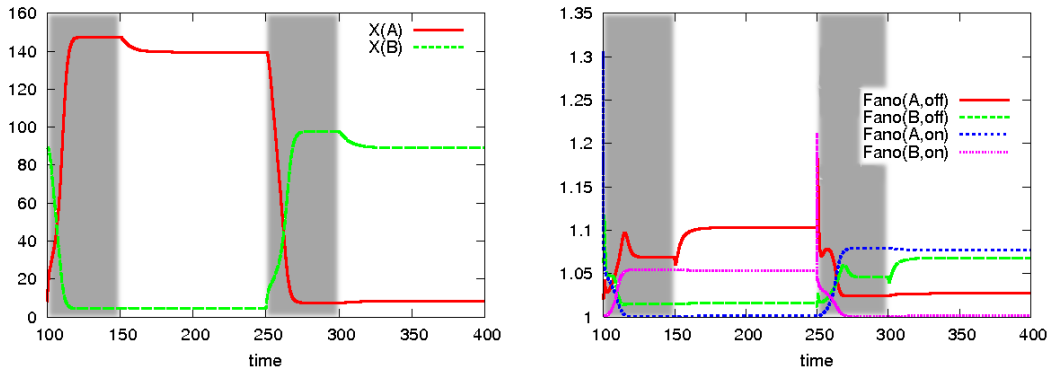

Figure 3.9: The transition between two fixed points (the number of proteins X versus time, and the Fano factor versus time) at given unbinding rate of protein to DNA, protein synthesis rate g, protein degradation rate k and binding rate of protein to DNA h (time is in units of the inverse of rate coefficients): $k_A = k_B = 1$, $g_{A0} = b_{B0} = 0$, $g_{A1} = 150$, $g_{B1} = 100$, $f_A = 0.2$, or $= 0.8(100 \leq t \leq 150)$, $f_B = 0.2$, or $= 0.8(250 \leq t \leq 300)$, $h_A = \frac{0.2}{500}$, $h_B = \frac{0.2}{1000}$.

bi-stability of the toggle switch.

3.4 Conclusions

We studied the toggle switch gene regulatory network using the master equation. The Hartree approximation of the master equation makes it unnecessary to solve the linear coupled differential equations with a huge number (exponential) of state variables to nonlinear coupled differential equations, replacing them with a small number (multiples) of parameters. Without self-interactions, the moment equations may be used for more accurate solutions.

By exploring the intrinsic statistical fluctuations of the toggle switch due to the finite of molecules in the cell, we provide a bridge to connect the theoretical investigations and single molecule measurements [38, 39]. Explicit time dependent Fano factors describe noise evolution and show a noisy state when the system is not in equilibrium.

Our studies of Fano factors showed two regimes in which the system was well-approximated by Poisson distributions. In the adiabatic limit, when binding and unbinding are slow compared to decay, "on" and "off" states each individually had Poisson- like distributions with different means, with small perturbations due to mixing. In the opposite limit, switching occurs so quickly that the two states are essentially interchangeable and Poisson.

We identify interesting effects on system time scales at the transition be-

tween bistability and other kinds of stability. Specifically, at this transition a second relaxation time scale becomes important and considerably longer than other time scales for similar systems. While the shorter time scale can be associated with climbing to a nearby probability peak, the longer time scale may be more related to the time necessary to transition from one probability peak to another. This would mean that, near phase transitions, it may be difficult to determine the exact value of the steady state observables such as protein number. Further work will be necessary to examine the associated ideas more closely.

We demonstrate time evolution dynamics in the toggle switch, and show the effects of inducing switching between its bistable states. Probabilistic switching between the two states is also of interest. One way to attack this problem is to use an "effective potential." [32, 40]. However, this is justified in some approximation limit and is not easily calculated in multi-variable systems..

Chapter 4

Repressilator

4.1 Introduction

Biological oscillation and its mechanisms have recently become a subject of intense study by a number of experimental and theoretical groups[1]. Its study is still in the early stages, and while the phenomenon itself is believed to be of great importance, knowledge of mechanisms by which it can occur is quite incomplete.

Oscillation is used in a number of biological systems, most notably in the circadian rhythms responsible for keeping organisms' biochemical processes in line with the day-night schedule of the Earth. In addition to 24-hour clocks, researchers have discovered genetic oscillators being used to determine timing of ovulation, and other important systems of this type are likely given that the tools for detailed study of the general phenomenon are still new. Malfunction of these oscillators is implicated in common conditions such as insomnia or jetlag [42], and more serious conditions such as bipolar disorder [43, 44]. Understanding biological oscillation and conditions which may interfere with it, then, are of great importance.

Repressilators, three-gene systems in which the genes cyclically repress each other, are interesting oscillating networks that are potentially important to synthetic biology[2, 4, 45, 46]. While undeniably artificial, they are easy to understand. Fig. 4.1 shows the system, one of the first synthetic networks shown to be capable of reliable oscillation behavior.

There are two aims of this chapter. The first is to show that the repressilator can exhibit monostable, spiral and limit-cycle oscillating behavior. The

[1]The data and many ideas from this chapter, and some of the language in this chapter, were originally co-authored with Keun-Young Kim and Jin Wang. Reproduced in part from [41]. Copyright 2007 American Institute of Physics.

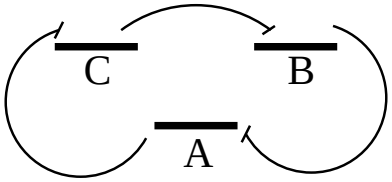

Figure 4.1: Network-style depiction of a repressilator, with three genes (A, B, C) cyclically repressing each other.

second is to quantitatively characterize the intrinsic noise of the system, as well as the correlations with order, amplitude and period of the repressilator oscillations.

4.2 Methods

We now use the time-dependent Hartree approximation scheme demonstrated in the previous chapter to reduce the dimensionality and solve the associated master equations to follow the evolution dynamics of the repressilator.

Our model is based on the following biochemical reaction picture for this gene network[4, 45]. Proteins of species A are synthesized from gene A and degraded at certain rates (g and k respectively). The synthesis rate g_A depends on the state of the gene A. Proteins of species A can bind to gene C and repress its synthesis of C species proteins. C proteins can bind to and repress gene B. B proteins, in turn, can bind to and repress gene A. This forms a network cycle (See Fig.4.1).

The corresponding Hartree-approximated master equation for the repressilator is then given by

$$\begin{aligned}\frac{dP_{A1}}{dt} = & -\tfrac{h}{2}n_B(n_B-1)P_{A1}(n_A) + fP_{A0}(n_A) \\ & +k(n_A+1)P_{A1}(n_A+1) - k(n_A)P_{A1}(n_A) + gP_{A1}(n_A-1) - gP_{A1}(n_A),\end{aligned} \tag{4.1}$$

$$\begin{aligned}\frac{dP_{A0}}{dt} = & \tfrac{h}{2}n_B(n_B-1)P_{A1}(n_A) - fP_{A0}(n_A) \\ & +k(n_A+1)P_{A0}(n_A+1) - k(n_A)P_{A0}(n_A),\end{aligned} \tag{4.2}$$

with Equations 4.1 and 4.2 permuted cyclically ($A \to C$ and $B \to A$, and $A \to B$ and $B \to C$).

With dimer protein repressors, the Poisson approximations reduce these six sets of infinite equations to nine equations:

$$\dot{C_{A1}} = -h_A C_{A1}\{C_{B1}X_{B1}^2 + (1-C_{B1})X_{B0}^2\} + f_A(1-C_{A1}) \tag{4.3}$$

$$\dot{X_{A1}} = -f_A \frac{1 - C_{A1}}{C_{A1}}(X_{A1} - X_{A0}) + g_{A1} - k_A X_{A1} \tag{4.4}$$

$$\dot{X_{A0}} = f_A(X_{A1} - X_{A0}) + g_{A0} - k_A X_{A0} - \frac{X_{A1} - X_{A0}}{1 - C_{A1}} \dot{C_{A1}} \tag{4.5}$$

$$\text{three equations } (A \to C, B \to A) \tag{4.6}$$

$$\text{three equations } (A \to B, B \to C) \tag{4.7}$$

where we eliminated three variables by the probability conservation ($C_{\alpha 1} + C_{\alpha 0} = 1$), and recollected terms.

We can also solve the corresponding moment equations exactly instead of using the Poisson approximation.

$$\frac{d}{dt} C_{A1} = \begin{array}{l} -h_A C_{A1}(C_{B1}(< n_{B1}^2 > - < n_B 1 >) \\ +C_{B0}(< n_{B0}^2 > - < n_B 0 >)) + f_A C_{A0} \end{array} , \tag{4.8}$$

$$\frac{d}{dt}(C_{A1} < n_{A1} >) = \begin{array}{l} g_{A1} C_{A1} - k_A C_{A1} < n_{A1} > \\ -h_A C_{A1} < n_{A1} > (C_{B1}(< n_{B1}^2 > - < n_{B1} >) \\ +C_{B0}(< n_{B0}^2 > - < n_{B0} >)) + f_A C_{A0} < n_{A0} > \end{array} , \tag{4.9}$$

$$\frac{d}{dt}(C_{A0} < n_{A0} >) = \begin{array}{l} g_{A0} C_{A0} - k_A C_{A0} < n_{A0} > \\ +h_A C_{A1} < n_{A1} > (C_{B1}(< n_{B1}^2 > - < n_{B1} >) \\ +C_{B0}(< n_{B0}^2 > - < n_{B0} >)) - f_A C_{A0} < n_{A0} > \end{array} , \tag{4.10}$$

$$\frac{d}{dt}(C_{A1} < n_{A1}^2 >) = \begin{array}{l} g_{A1} C_{A1}(2 < n_{A1} > +1) \\ +k_A C_{A1}(-2 < n_{A1}^2 > + < n_{A1} >) \\ -h_A C_{A1} < n_{A1}^2 > (C_{B1}(< n_{B1}^2 > - < n_{B1} >) \\ +C_{B0}(< n_{B0}^2 > - < n_{B0} >)) + f_A C_{A0} < n_{A0}^2 > \end{array} , \tag{4.11}$$

$$\frac{d}{dt}(C_{A0} < n_{A0}^2 >) = \begin{array}{l} g_{A0} C_{A0}(2 < n_{A0} > +1) \\ +k_A C_{A0}(-2 < n_{A0}^2 > + < n_{A0} >) \\ +h_A C_{A1} < n_{A1}^2 > (C_{B1}(< n_{B1}^2 > - < n_{B1} >) \\ +C_{B0}(< n_{B0}^2 > - < n_{B0} >)) - f_A C_{A0} < n_{A0}^2 > \end{array} \tag{4.12}$$

$$\text{five equations } (A \to C, B \to A) \tag{4.13}$$

$$\text{five equations } (A \to B, B \to C) \tag{4.14}$$

The moment equations, Eq.(4.8)-Eq.(4.9), reduce to Eq(4.3)-Eq.(4.5), if we assume Poisson relationships between the mean and the standard deviation of numbers of proteins.

4.3 Results and Discussion

We show that the repressilator exhibits mono-stable, spiral, and limit cycle oscillating behavior, and we characterize quantitatively the statistical fluctuations of the network.

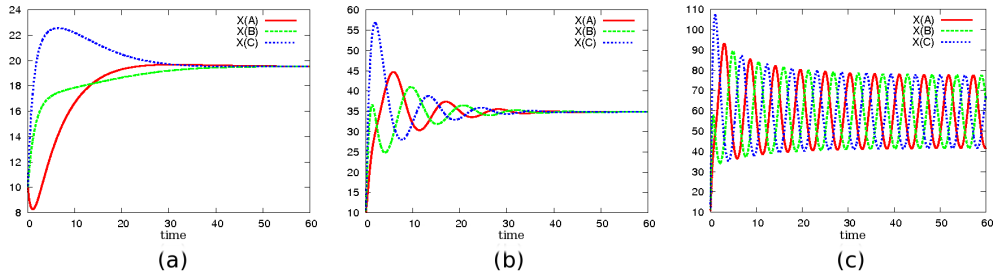

Figure 4.2: Number of proteins X versus time, using the Poisson approximation: $k = 1$, $g_0 = 0$, $h = \frac{f}{500}$, (a) stable node ($g_1 = 30$, $f = 0.1$). (b) stable spiral.($g_1 = 100$, $f = 0.2$). (c) limit cycle.($g_1 = 300$, $f = 0.5$). $< X_A >= C_{A1}X_{A1} + C_{A0}X_{A0}$, etc.

Fig.4.2(a)-(c) show the time dependent mean number of proteins and the probability of the three types of proteins produced by the corresponding three genes with Poisson Ansatz. We can see that when the protein synthesis rate and the unbinding rate of proteins from the gene are low relative to the protein self-degradation rate, the system is mono-stable as shown in Fig. 4.2 (a). With increased protein synthesis rate and unbinding rate, the system approaches its mono-stable state in a spiral fashion, shown in Fig. 4.2 (b). A further increase in the protein synthesis rate and the unbinding rate of proteins from the gene results in limit cycle behavior, with each type of protein number oscillating with a phase difference of 120 degrees to the others, as shown in Fig. 4.2 (c). This is the repressilator behavior observed in experiments [4].

So far, we have used only the Poisson approximation. However, Poisson distributions can only be truly accurate in systems which are in stable equilibrium states. Since the repressilator is a dynamically fluctuating system, we do not expect the Poisson approximation to be exact. Furthermore, statistical fluctuations beyond those expected from the Poisson approximation can be significant and need to be quantitatively addressed. Fig. 4.3, Fig. 4.4, and Fig. 4.5 were made using moment equations with the same parameters as Fig. 4.2, and show different period and amplitude from Fig. 4.2. We use the Fano factor to describe quantitatively the statistical fluctuations beyond the Poisson distribution.

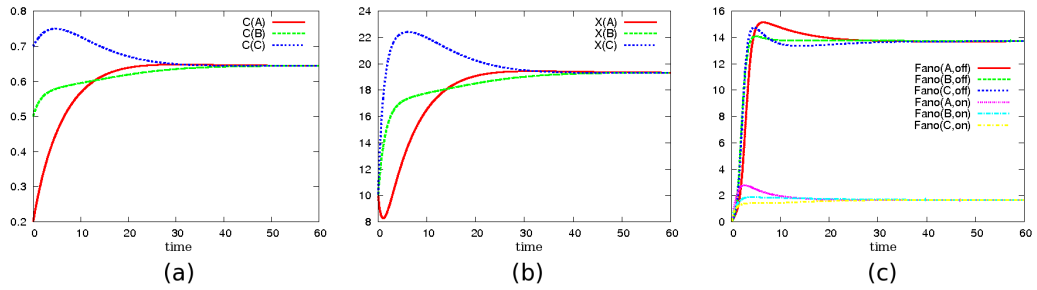

Figure 4.3: Time evolution of probability C, protein number X and Fano factor F with small protein synthesis rate g and unbinding rate f relative to self degradation rate k.

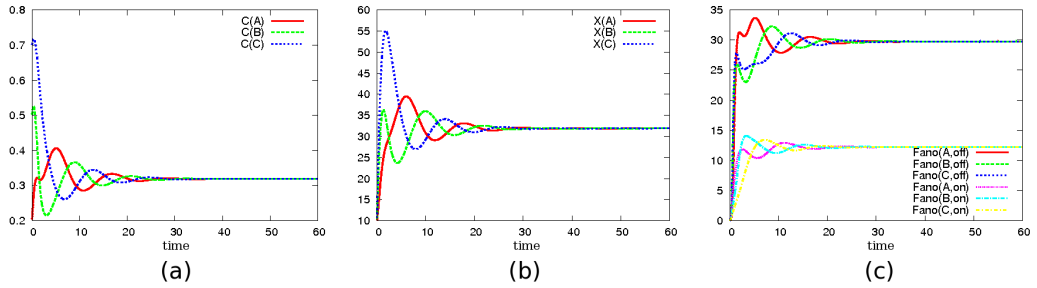

Figure 4.4: Time evolution of probability C, protein number X and Fano factor F with medium protein synthesis rate g and unbinding rate f relative to self degradation k rate.

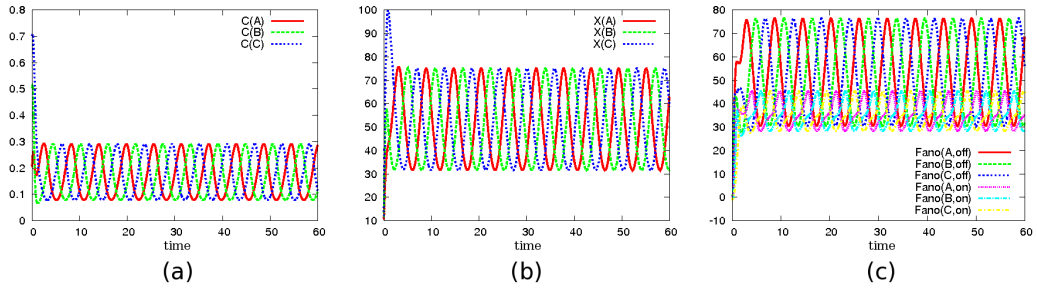

Figure 4.5: Time evolution of probability C, protein number X and Fano factor F with large protein synthesis rate g and unbinding rate f relative to self degradation rate k

We can see that the when the protein synthesis rate and the unbinding rate of proteins from the genes are low, the system is again mono-stable (Fig. 4.3). This is qualitatively similar to, but quantitatively different from, the results of the Poisson approximation. The Fano factor is approximately 2 (close to Poisson) in the "on" genetic state or 14 in the "off" state, as shown in Fig. 4.3(c).

As rates of protein synthesis and unbinding from the gene increase relative to the protein degradation rate, the system again transforms into a spiral approaching a mono-stable state, shown in Fig. 4.4. Again, however, while qualitatively similar to the Poisson approximation results, this is quantitatively different. The Fano factor, Fig. 4.4(c), is around 10 ("on") to 30 ("off"), meaning much larger statistical fluctuations than in the Poisson case.

When we further increase the rates of protein synthesis and unbinding from the gene, the system again becomes oscillatory, as shown in Fig. 4.5. There is a significant quantitative difference from Poisson approximation results, though again the behavior is qualitatively similar. Furthermore, the Fano factor, Fig.4.5 (c), oscillates with an amplitude on the order of tens with average around 40 (in the "on" genetic state) or 60 (in the "off" state). This means that it has much larger fluctuations and is very different from the Poisson distribution. The statistical distribution of the fluctuations in protein concentrations, as characterized by the higher-order moments, are therefore significant. This implies that the inherent distribution of protein concentrations must decay much more slowly than Poisson exponential; it has a long, or fat, tail.

The long tail of the distribution implies that, while large statistical fluctuations may happen rarely, they make a significant difference to the system. Such a phenomenon, called intermittency [47], can be seen as analogous to earthquakes, in which small frequent events cause little damage, but rare large earthquakes can cause a great deal of damage.

In Fig. 4.6(a), we plot the average protein concentrations versus the ratio of the protein unbinding rate f to the self-degradation rate k, $\omega = f/k$, and the ratio of protein synthesis rate g to self degradation rate k, $X_{ad} = g/k$. In this graph, we keep the ratio of the protein unbinding rate f to the binding rate h, $X_{eq} = f/h$, constant. We find that as ω and X_{ad} increase, the average protein number increases. When unbinding is significant (ω large), the genes are less repressed; this promotes protein production. Higher synthesis rates also enhance protein production, making the average protein concentration higher.

In Fig. 4.6(b), we plot the average protein concentrations versus the ratio of protein unbinding (to the gene) rate to self degradation rate $\omega = f/k$ and

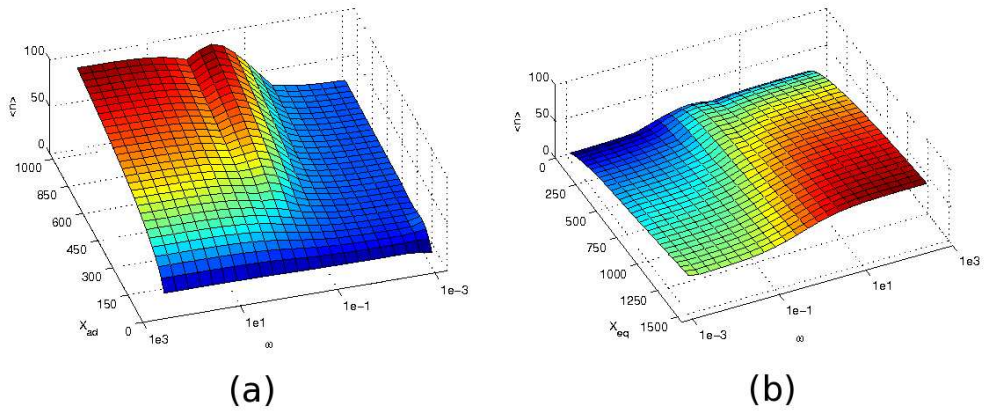

Figure 4.6: a: Average protein numbers versus ratio of protein unbinding rate f to self degradation rate k, ω, and ratio of protein synthesis rate g to self-degradation rate k, X_{ad}. b: Average protein numbers versus ratio of protein unbinding rate f to self-degradation rate k, ω, and ratio of unbinding rate to binding rate h, X_{eq}.

the ratio of unbinding to binding $X_{eq} = f/h$, keeping the ratio of protein synthesis rate to self degradation rate, $X_{ad} = g/k$, as constant. We find that as ω and X_{eq} increase, the average protein number increases. When unbinding is significant (ω large), the repression of the gene is less; this promotes the protein production. When unbinding is more significant than the binding, the repression through binding is less effective. This also enhances the protein production. Therefore the average protein concentration is also higher.

In Fig. 4.7, we can see the interrelationships among order, statistical fluctuations, amplitude and period of oscillations, versus ω and X_{ad}, keeping X_{eq} constant.

In Fig. 4.7(a) and (b), we can explore the phase diagram and the corresponding statistical fluctuations in the space of unbinding ω and protein synthesis X_{ad}. In the low ω and X_{ad} region, the system tends to be monostable (region III, deep blue). In the low unbinding ω and medium to high protein synthesis X_{ad} region (region II, light green), the system tends towards spirals. We found that the corresponding averaged Fano factors of statistical fluctuations of the protein numbers are high. In this region, the unbinding rate is small, and the proteins suppressing the gene do not stop that suppression quickly. Further, as suggested in the chapter on the toggle switch, at low ω the protein distribution is more spread out and may even be bimodal, with the mode at higher protein number corresponding to the unrepressed gene. Higher

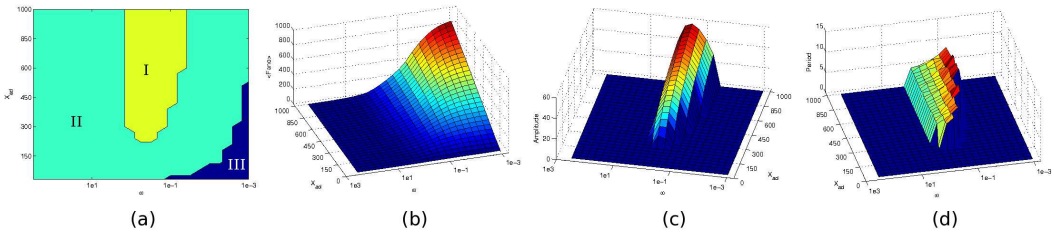

Figure 4.7: The relationships among order and stability, fluctuations, amplitude and period of oscillations of the repressilators versus ω and X_{ad}. a: Phase diagram of oscillation (I, yellow), spiral (II, light green), and stable (III, deep blue) dynamic behavior versus ω and X_{ad}. b: Average Fano factors versus ω and X_{ad}. c: Amplitude of repressilator oscillations versus ω and X_{ad}. d: Period of repressilator oscillations versus ω and X_{ad}.

protein numbers in the "on" state are more likely to cause repression sooner. Therefore, the corresponding protein production rate is small, any change in protein numbers can cause significant fluctuations due to the limited number of the proteins available (and the inherent noise one finds in a bimodal system is added in cases of low ω). Enhancing the synthesis rate X_{ad} of the proteins can suppress more of the genes and therefore lead to more bimodality or less protein production. This can also enhance the fluctuation (As X_{ad} increases, the average fano factors increase). The network experiences large fluctuations in this region. Monostable behavior emerges with less fluctuations than spirals. Although the unbinding ω is small favoring large fluctuations, the corresponding protein synthesis X_{ad} is small favoring less suppression of production. With small unbinding ω and medium to large X_{ad}, spirals emerge from large fluctuations. When unbinding ω increases, the average fano factors decrease. More significant unbinding will suppress genes less and produce more proteins. The intrinsic fluctuations is smaller when the number of the proteins is larger. The oscillatory repressilators emerge in this region with medium ω and medium to large X_{ad}.

In Fig. 4.7(c) and Fig. 4.7(d), we can explore the amplitude and period of the repressilators in the space of unbinding ω and protein synthesis X_{ad}. We find that the amplitude increases as the unbinding ω increases. This is because unbinding leads to more protein production and less fluctuations. Also, with less fluctuations, the amplitude of the oscillations can be larger without destroying the coherence. The system can sustain a larger amplitude of oscillations with less statistical fluctuations. When protein synthesis X_{ad} increases, the amplitude of oscillation increases because the number of proteins we would

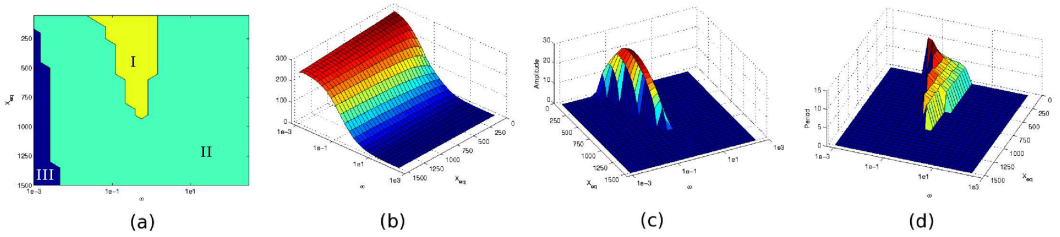

Figure 4.8: The relationships among order and stability, fluctuations, amplitude, and period of oscillations of repressilators versus the ratio of protein unbinding rate f to self-degradation rate k, ω, and the ratio of protein unbinding rate f to binding rate h, X_{eq}. a: Phase diagram of oscillation (I, yellow), spiral (II, light green), and stable (III, deep blue) dynamic behavior versus ω and X_{eq} . b: Average fluctuation Fano factor versus ω and X_{eq}. c: Amplitude of repressilator oscillations versus ω and X_{eq}. d: Period of repressilator oscillations versus ω and X_{eq}.

normally expect in the system is larger, and with no other considerations an oscillation over some percentage of a larger number is a larger oscillation.

In Fig. 4.7(d), we see that the period of repressilators decreases as unbinding ω increases. This is also because the less noise allows faster oscillation with less probability of destroying coherence, and also because faster binding and unbinding simply make one of the more important steps in the system less time-intensive. As the protein synthesis X_{ad} increases, the period slightly decreases. With increased protein production, there are smaller relative fluctuations and the period of oscillations can "afford" to be faster.

In Fig. 4.8, we can see the interrelationships among order, statistical fluctuations, amplitude, and period of oscillations, in the space of the ratio of protein unbinding to self-degradation, ω, and the ratio of unbinding to binding, X_{eq}, keeping the ratio of protein synthesis to self degradation, X_{ad}, constant.

In Fig. 4.8(a) and (b), we explore the phase diagram and the corresponding statistical fluctuations in the space of unbinding relative to self-degradation ω and unbinding relative to binding X_{eq}. In the very low ω region, the system tends to be mono-stable (region III, deep blue). In the small ω and low to high X_{eq} region (region II, light green), the system tends towards spirals. We find that the average fano factors of the protein numbers are high. In this region, the unbinding rate is small, so the system is likely represented by two more or less distinct states as mentioned in the chapter on the toggle switch. Smaller unbinding relative to binding X_{eq} can suppress the genes more and therefore lead to less protein production, enhancing fluctuations

(as X_{eq} decreases, average Fano factors increase). The network experiences large fluctuations in this region, and both mono-stable and spiral behavior emerge. When ω and X_{eq} increase, the average fano factors decrease. More significant unbinding will cause more mixing between the "on" and "off" states, decreasing the noise of the combined state. Again, the oscillatory repressilator emerges in the region where statistical fluctuations are comparatively smaller than the nearby mono-stable and spiral regions.

In Fig. 4.8(c) and (d), we examine the amplitude and period of the system versus unbinding relative to protein degradation ω, and unbinding relative to binding X_{eq}. We see that the amplitude increases as unbinding increases. This is because unbinding leads to more protein production and less fluctuation. Also, with less fluctuations, the amplitude of the oscillations can be larger without destroying coherence. However, in both Fig. 4.7 and Fig. 4.8 we note that in regions with very small fluctuation (Fano factor ~ 1), there is no oscillation; simple network topology and a lack of fluctuations are clearly insufficient for oscillation.

In Fig. 4.8(d), we find that the period of repressilators decreases as unbinding ω increases. This is also because smaller fluctuations allow more rapid oscillations without destroying the coherence through errors. As the unbinding relative to binding rate X_{eq} increases, the period decreases. As more unbinding leads to more proteins produced, there are less fluctuations and period of oscillations can "afford" to be a little faster.

4.4 Conclusions

We studied the repressilator gene network using the master equation formalism with the Hartree approximation demonstrated in the previous chapter. This system shows three kinds of distinct and important behavior: mono-stability, spirals, and limit cycle oscillation. Explicit time-dependent Fano factors describe noise evolution, and show large statistical fluctuations out of equilibrium, implying that the protein distributions are very far from Poisson. This is very relevant to the experimental studies of single molecule gene regulation dynamics [38, 39] as well as experimental studies on synthetic networks [2, 4]. We explored the phase space and the interrelationships among fluctuations, order, amplitude and period of oscillations of the repressilators. We found that repressilators follow ordered limit cycle orbits and are more likely to appear in regions of low but not extremely low flucuation. The amplitude of the repressilators increases as the suppression of the genes decreases and production of proteins increases. Oscillation periods decrease as the suppression of the genes decreases and protein production proteins increases.

Chapter 5

Self-Repressor

5.1 Introduction

Repressilators are not the only genetic oscillators[1]. A number of mechanisms can be used by organisms to give a reasonably coherent oscillation, including self-repressors with maturation times or other explicit, deterministic time delays (as opposed to chemical equation-type intermediate steps)[48–51], combined repression and activation loops[52]; highly non-linear protein degradation[53]; very large numbers (hundreds) of intermediate steps[54]; and self-repressors whose production and gene repression involve diffusion through the nuclear membrane[55, 56]. A self-activator with one stable and one semi-stable state, and stochastic switching between the two, can cause what may be called incoherent oscillation[57]. Additionally, certain kinds of self-repression can cause behavior which is not coherent oscillation when one considers the deterministic average protein and mRNA concentrations, but which still appear quite oscillatory and reasonably coherent when one considers any given system's stochastic trajectory through protein and mRNA concentration space over time[58].

This last example is of particular interest because in part of its simplicity; it is an uncomplicated system, easily modelled using straightforward Markov chain Monte Carlo methods. (Deterministic time delays of the type $t - \tau$, the main method used to get a simple self-repressor to oscillate, require non-Markov simulation can be difficult to justify or interpret.) There is, however, another reason for interest in the work. The binding mechanism suggested by the mathematics presented by McKane and Newman in [58] is in fact a combination of multiple n-mer bindings of regulatory proteins to the gene,

[1]The data and ideas from this chapter, and much of the language in this chapter, have been submitted for publication, co-authored with Haidong Feng and Jin Wang.

where the primary n here depends on the parameter chosen and the average number of proteins present in the system. While this in itself may be non-physical, it inherently suggests that the number of bound proteins may be able to change the system from non-oscillatory to oscillatory with noise to (possibly) oscillatory even without noise.

We note that the simplest possible version of this last suggestion, a deterministic model with mRNA and with n regulatory proteins binding directly and instantaneously to the gene, cannot be oscillatory (though the stochastic version of the system can be at least to some extent, and we explore this briefly).

5.2 Methods

We performed both determinisic and stochastic calculations. The following were constant for all calculations: the degradation rate of mRNA $k_m = 1$ (hence, $\tau_m = 1$); the degradation rate of monomer protein $k_p = 1$ (hence, $\tau_p \geq 1$); the rates of gene binding and unbinding $\omega \gg 1$ (and so the system is very adiabatic); the ratio of the n-mer dissolution constant f to the formation constant h is $\frac{f}{h} = \left(\frac{g_1 g_p}{k_m k_p}\right)^n$; the rate of n-mer dissolution $f \gg 1$ in the simplest case, Fig. 5.1, and $f = 1$ in all other calculations; the number of complete n-mers which will cause the gene to be repressed by a factor of $1/e$, $X_{eq} = 10$; the protein synthesis rate from a single strand of mRNA $g_p = 3$; and the ratio of mRNA synthesis in the repressed versus the unrepressed gene state $g_0 = g_1/100$. It should be noted that g_p and X_{eq} can be rescaled in systems without internal noise; only when the actual number becomes important, as opposed to relative concentrations, do these quantities have any significant effect other than a rescaling. It should also be noted that the ratio τ_p/τ_m is in rough agreement with the average for this sort of gene in yeast [59].

Deterministic calculations (involving the average numbers of mRNA, proteins, and protein aggregates) used the simple set of equations

$$\frac{dm}{dt} = -k_m m + g_0 P_{\text{off}} + g_1 P_{\text{on}} = -k_m m + g_0 + \frac{g_1 - g_0}{1 - c/X_{eq}},$$

$$\frac{dp}{dt} = -k_p p + g_p m - h n p^n + f n c,$$

$$\frac{dc}{dt} = h p^n - f c,$$

where m is the number of mRNA molecules, p is the number of monomer proteins, and c is the number of n-mer proteins in the system at time t. Simple

modifications made additional intermediate steps possible (c becoming c_1, n becoming n_1, additional terms in the last equation for the $n_2c_1 \rightarrow c_2$ step, the extra equations for c_2 in terms of c_1 and c_3, etc.).

As an aside, we should mention that we assume all binding steps involved are fast compared to the other processes involved in the system, though binding events are relatively rare, and binding is highly cooperative. Slower binding, or binding that is less cooperative, can introduce very different elements to a system, as shown by [60].

The authors used time-series data to determine the oscillatory nature of the systems. A general, analytical solution to the linear stability analysis of the system would be infeasible, and even numerical solutions alone could have glossed over complex behavior less useful to an organism than regular oscillation. However, individual points in different regions identified by the time-series data have been checked numerically using linear stability analysis.

Stochastic calculations were straightforward, using the following equations:

$$R_{\text{mRNA synthesis}} = g_0 + \frac{g_1 - g_0}{1 - c/X_{eq}},$$

$$R_{\text{mRNA degradation}} = k_m m,$$

$$R_{\text{protein synthesis}} = g_p m,$$

$$R_{\text{monomer protein degradation}} = k_p p.$$

In the calculations for Fig. 5.1 (in which binding to the gene is coupled with n-mer formation), we used

$$c = \frac{h}{f} p(p-1)\dots(p-n+1).$$

For other stochastic calculations, in which the proteins bind to each other before binding to the gene, we used instead

$$R_{n-\text{mer formation}} = hp(p-1)\dots(p-n+1),$$

$$R_{n-\text{mer breaking}} = fc.$$

A time-step dt was then calculated using these rates, ensuring that events generally occurred one at a time by using $dt \sim 0.01/\sqrt{\sum R^2}$. (This is subtly different from the traditional Gillespie simulation in that multiple events were in theory allowed, if quite unlikely; however, the difference is very small and may reasonably be considered to be an advantage of our algorithm.) For practicality's sake, at very high levels of synthesis, multiple synthesis and

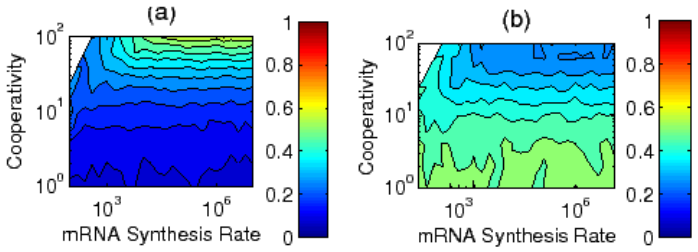

Figure 5.1: a: Stochastic calculation of coherence in a system which has mRNA and proteins cooperatively binding to the gene, with coherence given by $2\frac{\sum\Theta(d\phi)}{\sum|d\phi|} - 1$ where $d\phi$ is the difference in angle in mRNA-protein space. b: Stochastic calculation of the standard deviation of the period distribution divided by its mean. Both colormaps are on the same scale as Fig. 5.3.

decay events were allowed and occurred in the background of other events, but were kept to at most a 0.01% mean change in the number of molecules.

Each process was then treated as a Poisson process using the same dt, with the probability of a events of type b occurring being $\frac{e^{-R_b dt}(R_b dt)^a}{a!}$.

The programs used were written in C, and run using Fedora 10 Linux on a Dell desktop computer.

5.3 Results

We begin with the stochastic system described first, in which binding to the gene is coupled with n-mer formation. (This could be either because the binding to the gene is itself cooperative or because it is very fast once the n-mer is formed.) In the described regime, whose deterministic solutions yield at best decaying oscillation, we now explore the possibility of noise-induced oscillation. We note that the system in this case is between truly oscillatory and simply two-state with reasonably frequent switching. Simple two-state switching would lead to a "period distribution" (time between maxima) with a normalized standard deviation of that period $\frac{\sigma_t}{\bar{t}}$ of $\frac{1}{\sqrt{2}}$. Fig. 5.1(b) shows $\frac{\sigma_t}{\bar{t}}$, and Fig. 5.1(a) shows the coherence (generalized from the definition used by [46]). Both the coherence and $\frac{\sigma_t}{\bar{t}}$ strongly imply that increased cooperativity yields a steadier oscillatory behavior, but that in the regions explored there is no coherent oscillation (which would require a coherence close to 1 and a value of $\frac{\sigma_t}{\bar{t}}$ significantly less than 1).

So far, the only pieces of the puzzle considered have been multimer binding of proteins to genes, mRNA, and noise. We now add the final piece, the possi-

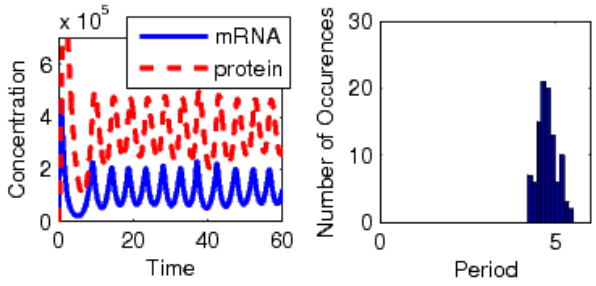

Figure 5.2: Left graph: oscillation due to an intermediate step with a cooperativity of 16 and noise. Right graph: period distribution for the same system.

bility of proteins binding together before they bind to the gene. The number of steps involved here keeps oscillation from occurring at a cooperativity of 1, but increased n causes oscillatory behavior when combined with the rest of the system.

Specifically, Fig. 5.2 shows a stochastic system with $n = 16$ and the distribution of periods, defined here as time between local maxima in protein number which are at least $0.4 \cdot \tau_{\text{mRNA}}$ (in order to remove less significant fluctuations from consideration). This system is clearly oscillatory, with a reasonably sharp period distribution.

It should be noted that 16 is at best a marginally reasonable value for cooperativity in a simple genetic system. However, it is clear now that simple intermediate steps can interact with cooperativity, and between them can produce oscillations in which neither is the primary factor in the behavior. Furthermore, examining the parameter space in more detail, we find telling behavior in both deterministic and stochastic cases. Fig. 5.3a shows the deterministic phase diagram of the system, made using time-series calculations. Region I is oscillatory, region II displays decaying oscillation, and region III is non-oscillatory. High cooperativity and synthesis rate are both clearly necessary for this; also, it should be noted that the presence of oscillation at all means that the additional intermediate step as compared with the system from Fig. 5.1 is necessary for oscillation.

In Fig. 5.3b, we see, in the same region, a graph of the coherence. This corresponds well with the deterministic calculations, as we would expect coherence in a truly oscillatory system. Fig. 5.3c shows $\frac{\sigma_t}{t}$, which again corresponds well, although not perfectly; some loss of specificity in period at very high cooperativity and synthesis implies that the oscillation may be imperfect or require that other parameters be very specific at some level. However, the

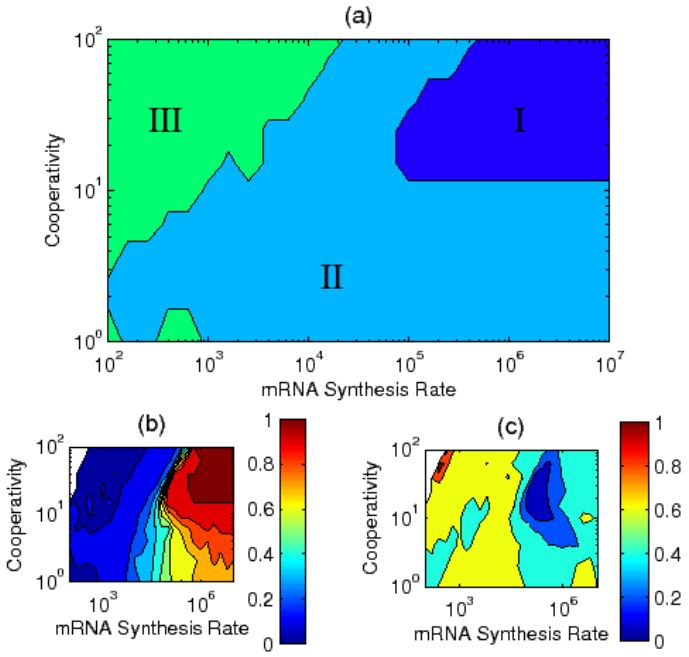

Figure 5.3: a: Deterministic calculation of oscillatory features in a system with mRNA, cooperative binding of proteins to each other, and a separate step binding to the gene; region I is oscillatory, region II has decaying oscillations, and region III is non-oscillatory. Due to difficulties distinguishing II from III, there is some overlap. b: Stochastic calculation of coherence, given by $2\frac{\sum\Theta(d\phi)}{\sum|d\phi|} - 1$ where $d\phi$ is the difference in angle in mRNA-protein space. c: Stochastic calculation of the standard deviation of the period distribution divided by its mean.

increase in $\frac{\sigma_t}{t}$ at these points is slight.

From these figures, it is apparent that coherence is possible with higher rates of synthesis, and that there is a tendency towards higher coherence with higher cooperativity. Additionally, we note that in the stochastic case there is no jump in coherence or $\frac{\sigma_t}{t}$ as we would find if the system exhibited the phase-transition behavior from Fig. 5.3a; in the stochastic case, the line between oscillatory and non-oscillatory is blurred. It is important to note that noise alone is not sufficient for reliable or even semi-reliable oscillation. The deterministically oscillating region clearly has more stable oscillation in the stochastic regime, and systems close to it are more capable of reliable behavior than those far away from it.

For comparison, we now choose a point in the region of deterministic de-

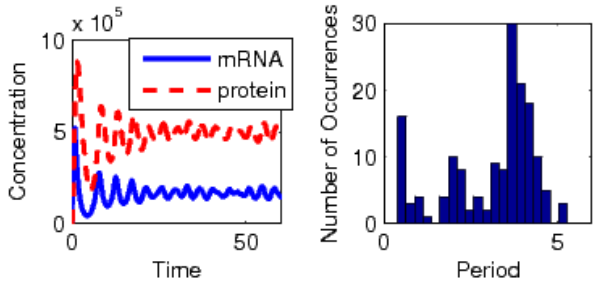

Figure 5.4: Left graph: behavior with small oscillatory tendencies due to an intermediate step with a cooperativity of 8 and noise. Right graph: period distribution for the same system.

caying oscillations, and make plots comparable to Fig. 5.2. In Fig. 5.4 we see a system whose behavior seems similar in many ways; the main difference in this case is a slightly widened period distribution. While this system is definitely in region II of Fig 5.3a (with a cooperativity of 8), and the period distribution may be too wide for a truly accurate clock, it is obvious that the line between coherent oscillation and incoherent oscillation-like behavior is blurred in this case by noise.

We see in these figures that the optimal cooperativity for this number of intermediate steps in binding is still, at best, at the very high end of cooperativity in reasonable biological systems. Therefore, we have attempted a less complete search of parameters considering additional steps (instead of monomer to n-mer, monomer to dimer to tetramer, etc.). While our searches were not exhaustive, we found at least one region in which such a system can oscillate at quite reasonable values for n, as low as 8 (octomer).

5.4 Conclusions

Our data show that oscillation occurs when cooperativity and intermediate steps are both present, when neither cooperativity nor intermediates would otherwise be sufficient. More, it implies that, with many intermediate steps, lower cooperativity can yield oscillation, while with higher cooperativity fewer intermediate steps are necessary. Oscillation behavior in deterministic calculations gives rise to more stable oscillation in the stochastic case.

Additionally, it should be noted that, even though we found no coherent oscillation for lower values of cooperativity, moderate coherence in an oscillation-like behavior may not necessarily be fatal or possibly even very detrimental

to a biological system which relies on it as a clock for less important functions. Many biological clocks receive input of some kind from outside sources, whether light and food for 24-hour clocks or some other form[61, 62]. These inputs can reset a biological clock, and one may argue that they might be able to do so more easily in a fundamentally inexact clock than, for instance, a single-mode one such as a repressilator. However, for time-sensitive vital functions, it would seem likely that higher coherence, which occurs when the deterministic equations also show oscillating behavior, is necessary.

In summary, we have discovered a relationship between generally realistic intermediate steps (as opposed to set-time time delays of the form $t-\tau$, which are realistic for a limited set of systems), cooperativity, and oscillation. Intermediate steps resulting from slow cooperative binding can cause oscillation at biologically relevant cooperativity. We have also explored the noise effect, in which the line between oscillatory and non-oscillatory (relatively clear in the deterministic case) is blurred in the stochastic case.

Chapter 6

Overall Conclusions and Discussion

We have shown some of the roles of intrinsic noise in simple genetic networks, and explored its effects on the behavior of those networks. In addition, we have explored oscillation behavior, and the interplay between it, cooperative binding, and noise.

Noise is an interesting contrast in these systems. For the most part, noise has a detrimental role; in order for a system to have reliable responses, noise must be accounted for and even overcome by a correct "choice" of network parameters. This is apparent in the case of Bicoid, in which the intrinsic noise interferes with the basic function of the protein (marking of position within the embryo for development purposes). It is also clear in the case of the repressilator, where increased noise interferes with coherence. However, in the case of the self-repressor it becomes clear that noise can serve a purpose. While noise-induced oscillation in deterministically non-oscillatory systems has been seen before, the interpretation we offer in the chapter on the self-repressor provides some explanation of the phenomenon. Noise makes the transition in phase space from oscillatory to non-oscillatory slow and blurred, and so regions which are non-oscillatory but close to oscillatory regions will have some coherence.

This might seem obvious in some sense. However, it has not been considered in the context before. Phase space contains a large number of parameters which in any realistic system cannot be changed. Specifically, the number of proteins in a complete n-mer which binds to a gene is unlikely to change without major changes to other parameters. It is therefore likely that the n in question is not generally considered part of "parameter space." The inclusion of cooperativity as a parameter and the understanding that noisy somewhat-oscillatory behavior can occur because of proximity to oscillatory behavior is

an important step forward in our understanding of these systems.

Nevertheless, it is also clear that noise-induced oscillation is less reliable or coherent than even noisy oscillation in regimes which would still oscillate deterministically. Again, this may be unsurprising, but it is important to keep in mind when one considers the utility of the system; a single noise-induced oscillator may serve as a clock for a minor biological function, but one which is especially important and time-sensitive must have an accurate clock system.

In addition to seeing noise's role, it is useful to understand the quantity of noise and the ways in which it can be affected. Having multiple interfering processes occurring at the same time scale, or non-adiabaticity, can be a factor in increasing noise and distorting probability functions which are otherwise relatively neat and clean (as shown in the chapter on toggle switches). Fano factors are a simple way to measure this sort of noise, since a Fano factor of 1 implies the smallest possible amount of noise this sort of system can normally have. The Fano factors we calculate for these systems bear out the idea of non-adiabaticity increasing noise.

At the same time, it is clear from our work on the repressilator and self-repressor that multiple processes having similar time scales can be useful for behavior which is more interesting than simple switching. In the first case, multiple genes have the same time scales for all their processes; each gene and its proteins therefore are essentially non-adiabatic with respect to each other. Further, though the original proposed repressilator used a slightly different method, the oscillatory regime we found was around $\omega = 1$, where the time scales of binding and unbinding are close to that of protein degradation. (In the original repressilator, mRNA degradation and protein degradation were on the same time scale instead [4].) In the second case, the self-repressor, the intermediate steps mentioned in the chapter must be on the same time scale as protein degradation to have a useful impact on behavior. Thus, in spite of creating additional and potentially problematic noise (as seen in the Fano factor figures in the repressilator chapter when additional non-adiabaticity is introduced), similar time scales can also be useful or necessary for function. For this reason, many systems which have usefully complex behavior may be inherently noisy. As life itself is a complex phenomenon which comes from these kinds of networks, we can easily argue that a more thorough understanding of these issues is potentially extremely important.

In addition, the observations we discussed bring up a few important questions. First, is the apparent correlation between noise and complex behavior more widespread than these few systems? Is noise a byproduct of non-adiabaticity as mentioned, or is it somehow necessary for complex behavior? If it is a byproduct, are there ways to reduce it, and what are the costs to

the organism associated with the noise? What could be the costs of reducing noise (energy cost of increased protein synthesis if the system needs a simple $1/\sqrt{\text{protein number}}$ noise reduction, or of protein synthesis and gene replication if another regulatory gene is necessary, or other similar costs)? When is it energetically favorable to reduce noise, and is there more to consider than only energy in the matter?

Answering such questions would lead to a great deal of understanding of life in general, especially questions involving evolution of complex organisms. That understanding would also doubtless be beneficial to the field of medicine because it would help to correct situations in which genetic and epigenetic factors are responsible for network malfunction. These questions, however, are quite large in scope, and could easily provide several lifetimes' worth of both theoretical and experimental scientific research.

Bibliography

[1] T Gregor, E.F. Wieschaus, A.P. McGregor, W. Bialek, and D.W. Tank. *Cell*, 130:141–152, 2007.

[2] L. You, R. S. Cox, III, R. Weiss, and F. H. Arnold. *Nature*, 428:868–871, 2004.

[3] H.H. McAdams and A. Arkin. *Proc. Natl. Acad. Sci.*, 94:814–819, 1997.

[4] Michael B. Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338, 2000.

[5] P.S. Swain, M.B. Elowitz, and E.D. Siggia. *Proc. Natl. Acad. Sci.*, 99: 12795–12800, 2002.

[6] M. Thattai and A. van Oudenaarden. *Proc. Natl. Acad. Sci.*, 98:8614–8619, 2001.

[7] J.M.G. Vilar, C.C. Guet, and S. Leibler. *J. Cell Biol.*, 161:471–476, 2003.

[8] J. Paulsson. *Nature*, 427:415–418, 2004.

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

[10] David Lepzelter and Jin Wang. Exact probabilistic solution of spatial-dependent stochastics and associated spatial potential landscape for the bicoid protein. *Physical Review E*, 77:041917, 2008.

[11] Wolfgang Driever and Christiane Nüsslein-Volhard. *Cell*, 54:95–104, 1988.

[12] Gary Struhl, Kevin Struhl, and Paul M. Macdonald. *Cell*, 57:1259–1273, 1989.

[13] Yu Feng Wu, Ekaterina Myasnikova, and John Reinitz. *Unpublished*, 2007.

[14] Filipe Tostevin, Pieter Rein ten Wolde, and Martin Howard. *PLOS Comp. Biol.*, 3:e78, 2007.

[15] C. H. Waddington. *Strategy of the gene. London: Allen and Unwin.*, page 290 p., 1957.

[16] R. A. Fisher. *The genetical theory of natural selection. Oxford:Clarendon*, page 251 p., 1930.

[17] S. Wright. *Proceedings of the Sixth International Congress on Genetics*, 1:356–366, 1932.

[18] H. Frauenfelder, S. G. Sligar, and P. G. Wolynes. *Science*, 254:1598–1603, 1991.

[19] Peter G. Wolynes, Jose N. Onuchic, and Dave Thirumalai. *Science*, 267: 1619–1620, 1995.

[20] Masaki Sasai and Peter G. Wolynes. *Proc. Natl. Acad. Sci.*, 100:2374–2379, 2003.

[21] N. G. Van Kampen. *Stochastic Processes in Physics and Chemistry, Elsevier Sience Publishers*, page 465 p., 1992.

[22] X.M. Zhu, L. Yin, L. Hood, and P. Ao. *Journal of Bioinformatics and Computational Biology*, 2:785–817, 2004.

[23] H. Qian and D. A. Bear. *Biophys. Chem.*, 114:213–220, 2005.

[24] J. Wang, B. Huang, X.F. Xia, and Z. R. Sun. *Biophys. J. Lett.*, 91: L54–L57, 2006.

[25] J. Wang, B. Huang, X.F. Xia, and Z. R. Sun. *PLOS Comp. Biol.*, 2:e147, 1385, 2006.

[26] K. Kim and J. Wang. *PLOS Comp. Biol.*, 3:e60, 2007.

[27] B. Han and J. Wang. *Biophys. J.*, 92:3755–3765, 2007.

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

[29] C.W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*. Springer-Verlag: Berlin, 1985.

[30] C.W. Gardiner and S. Chaturvedi. *J. Stat. Phys.*, 17:429–467, 1977.

[31] David Lepzelter, Keun Young Kim, and Jin Wang. Dynamics and intrinsic statistical fluctuations of a gene switch. *J. Phys. Chem. B*, 111:10239–10247, 2007.

[32] M. Sasai and P.G. Wolynes. Stochastic gene expression as a many-body problem. *Proc. Natl. Acad. Sci.*, 100:2374–2379, 2003.

[33] A.M. Walczak, M. Sasai, and P.G. Wolynes. Self consistent proteomic field theory of stochastic gene switches. *Biophys. J.*, 88:828–850, 2003.

[34] T.S. Gardner, C.R. Cantor, and J.J. Collins. *Nature*, 403:339–342, 2000.

[35] A. Arkin, J. Ross, and H.H. McAdams. *Genetics*, 149:1633–1649, 1998.

[36] H. Qian. *Ann. Rev. Phys.Chem.*, 58:113–142, 2007.

[37] T. Ushikubo, W. Inoue, M. Yoda, and M. Sasai. *Chemical Physics Letters*, 430:139–143, 2006.

[38] J. Yu, J. Xiao, X. Ren, K Lao, and X.S. Xie. *Science*, 311:1600–1603, 2006.

[39] L. Cai, N. Friedman, and X.S. Xie. *Nature*, 440:358–362, 2006.

[40] T.B. Kepler and T.C. Elston. Stochasticity in transcriptional regulation: Origins, consequences, and mathematical representations. *Biophys. J.*, 81:3116–3136, 2001.

[41] Keun Young Kim, David Lepzelter, and Jin Wang. Single molecule dynamics and statistical fluctuations of gene regulatory networks: A repressilator. *J. Chem. Phys.*, 126:034702, 2007.

[42] Takashi Ebisawa. Circadian rhythms in the cns and peripheral clock disorders: Human sleep disorders and clock genes. *Journal of Pharmacological Sciences*, 103(2):150–154, 2007.

[43] D.F. Kripke, D.J. Mullaney, M. Atkinson, and S. Wolf. Circadian rhythm disorders in manic-depressives. *Biol. Psychiatry*, 13(3):335–351, 1978.

[44] Colleen A. McClung. Circadian genes, rhythms and the biology of mood disorders. *Pharmacology and Therapeutics*, 114(2):222–232, 2007.

[45] Tomohiro Ushikubo, Wataru Inoue, and Masaki Sasai. Dynamics of repressilator: From noise to coherent oscillation. *Genome Informatics*, 14: 314–315, 2003.

[46] Mistumasa Yoda, Tomohiro Ushikubo, Wataru Inoue, and Masaki Sasai. Roles of noise in single and coupled multiple genetic oscillators. *Journal of Chemical Physics*, 126:115101, 2007.

[47] J. Wang and P. G. Wolynes. Intermittency of single molecule reaction dynamics in fluctuating environments. *Phys. Rev. Lett.*, 74:4317, 1995.

[48] Julian Lewis. Autoinhibition with transcriptional delay: A simple mechanism for the zebrafish somitogenesis oscillator. *Current Biology*, 13:1398–1408, 2003.

[49] William Mather, Matthew R. Bennett, Jeff Hasty, and Lev S. Tsimring. Delay-induced degrade-and-fire oscillations in small genetic circuits. *Physical Review Letters*, 102:068105, 2009.

[50] Jesse Stricker, Scott Cookson, Matthew R. Bennett, William H. Mather, Lev S. Tsimring, and Jeff Hasty. A fast, robust and tunable synthetic gene oscillator. *Nature*, 456:516–519, 2008.

[51] Pierre-Emmanuel Morant, Quentin Thommen, François Lemaire, Constant Vendermoëre, Benjamin Parent, and Marc Lefranc. Oscillations in the expression of a self-repressed gene induced by a slow transcriptional dynamics. *Physical Review Letters*, 102:068104, 2009.

[52] Tony Yu-Chen Tsai, Yoon Sup Choi, Wenzhe Ma, Joseph R. Pomerening, Chao Tang, and James E. Jr. Ferrell. Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science*, 231: 126–129, 2008.

[53] John J. Tyson, Christian I. Hong, Dennis Thron, and Bela Novak. A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. *Biophysical Journal*, 77:2411–2417, 1999.

[54] Luis G. Morelli and Frank Jülicher. Precision of genetic oscillators and clocks. *Physical Review Letters*, 98:228101, 2007.

[55] Albert Goldbeter. A mode for circadian oscillations in the *drosophila* period protein (PER). *Proceedings of the Royal Society B*, 261:319–324, 1995.

[56] Jin Wang, Li Xu, and Erkang Wang. Robustness, dissipations and coherence of the oscillation of circadian clock: potential landscape and flux perspectives. *PMC Biophysics*, 1, 2008.

[57] Daniel Schultz, Eshel Ben Jacob, José N. Onuchic, and Peter G. Wolynes. Molecular level stochastic model for competence cycles in *bacillus subtilis*. *PNAS*, 104:17582–17587, 2007.

[58] A. J. McKane, J. D. Nagy, T. J. Newman, and M. O. Stefanini. Amplified biochemical oscillations in cellular systems. *Journal of Statistical Physics*, 128:165–191, 2007.

[59] Vahid Shahrezaei and Peter S. Swain. Analytical distrobutions for stochastic gene expression. *PNAS*, 105(45):17256–17261, 2008.

[60] M. Santillán. On the use of the hill functions in mathematical models of gene regulatory networks. *Mathematical Modelling of Natural Phenomena*, 3(2):85–97, 2008.

[61] Jürgen Aschoff and Serge Daan. *The Entrainment of Circadian Systems*, volume 12 of *Handbook of Behavioral Neurobiology*, pages 7–43. Kluwer Academic/Plenum Publishers, 2001.

[62] Etienne Challet, Ivette Caldelas, Caroline Graff, and Paul Pévet. Synchronization of the molecular clockwork by light- and food-related cues in mammals. *Biological Chemistry*, 384:711–719, 2003.