Stochastic Modeling of Gene Regulatory Networks

Mistry, Hitesh

2007

MIMS EPrint: **2007.77**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

STOCHASTIC MODELING OF GENE REGULATORY NETWORKS

by

Hitesh Mistry

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(School of Mathematics)
in The University of Manchester
2007

Doctoral Committee:

ACKNOWLEDGEMENTS

I wish to express my deepest and sincere gratitude to Dr. M. Muldoon for his excellent supervision of this work. His encouragement and flair to motivate the progress of research have been invaluable.

Thanks are due to Professor J. Dold for providing me with the opportunity to undertake my PhD in Manchester. I am also indebted to Dr. J. P. Huke for many fruitful discussions on just about everything from Mathematics through to Biology and the mysteries of women!

I would like to thank my family, Hasumati, Balvanta, Anita and Rina Mistry, for their continued support throughout my time at University.

More thanks go out to my lifelong friends Kapil, Ryan, Hinesh, Umesh, Dhiren and Rob who have been un-waivering in their support. Many more thanks to all the friends at UMIST and all the people who have inhabited the rooms on Q-floor.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

Stochastic Modeling of Gene Regulatory Networks

by

Hitesh Mistry

Gene Regulatory Networks (GRNs) describe how chemical species within a cell interact with one another, thereby governing the rates at which key genes are expressed. This thesis is concerned with modeling a particular GRN, *Arabidopsis thaliana Circadian Clock*, by considering three different approaches; discrete stochastic, continuous stochastic and parameter variation. By considering these different methods we will see if the desired behavior required from our network is robust to biological noise.

Through employing stochastic approaches we found the GRN under question is robust to biological noise to a point; the results of our study led to a couple of interesting questions to people within the field. When the number of molecules involved in the reactions were reduced sufficiently the biological noise in the system destroyed the desired circadian rhythm. To the biologists we would ask how low are the molecule numbers involved in such reactions and to the modelers how appropriate is it to use Michaelis-Menten type kinetics for low molecule numbers.

Copyright Statement

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.

CHAPTER I

Introduction

1.1 Biological Background

One of the key aspects of developmental biology is to know which genes are activated at certain times and to what extent. By knowing this an understanding of how an organism functions can be achieved. Interactions between DNA, RNA, protein and other smaller molecules are responsible for the regulation of gene expression. To be able to develop a mathematical model [10] of such regulatory networks is indeed a very difficult challenge as the pathways can be very complex: a sound understanding of the underlying biology of the process is essential [1]. The main example under consideration in this thesis is the modeling of the circadian clock in *Ararbidopsis thaliana*. The model is a prime example of how experiments and mathematics can be used together to help understand biological processes. A group of mathematicians and biologists in Warwick University, Locke et al., have developed a system of first-order non-linear differential equations to model the *Arabidopsis Circadian Clock* [31, 32, 24], consisting of a set of genes that are responsible for maintaining rhythmic expressions of key genes, discussed further in Chapter 3.

Before we specialize to *Arabidopsis* a quick review of what Francis Crick has called the "Central Dogma" [9] is in order. The dogma holds that genetic information flows

"from DNA to RNA to protein": DNA is the store of hereditary information within a cell and the carrier of this information from one generation to the next. These DNA molecules are very long, unbranched polymers and encode the genetic information as linear sequence of *nucleotides*. The nucleotides consist of three elements (see, e.g. [5] for details): the sugar deoxyribose, a phosphate group and a nitrogenous base which is one of adenine(A), cytosine (C), guanine (G) or thymine (T). DNA is a three-dimensional structure, in which two DNA complementary polymer chains (the bases are associated in pairs, C with G and A with T) are held together by a hydrogen bonds. There is a sense of direction on DNA defined by the way the links in the polymer chain fit together. Genetic information is encoded by the sequence of the nucleotides along each chain. Think of each base—A, C, G or T—as a letter drawn from a four letter alphabet which will spell out biological messages. Every organism will then differ, to a certain degree if clones are ignored, because their respective DNA molecule will consist of different sequences of nucleotides and hence will spell out different messages.

Transcription of a DNA molecule is an intricate process, through which the cell prepares a kind of working copy of some of its genetic information, copying the base sequence of a segment of its DNA into an *RNA* molecule. *RNA polymerases* are the enzymes that are required to perform transcription. They are responsible for unraveling the DNA chains and copying the genetic information. For transcription to take place, several additional ingredients are required: these help the RNA polymerase to remain attached and to start the copying process. Firstly the RNA polymerase must realize where on the genome to start transcription and where to finish.

There are at least three different types of RNA polymerase: each transcribes different RNA genes. The one that is of main interest here is RNA polymerase II; it can

transcribe all the protein coding genes. On the DNA sequence there exist *promoter regions*, sites to which the polymerase will bind, and which occur just upstream of the start site of transcription. In these regions a number of *transcription factors* will attach to then help position the RNA polymerase correctly at the promoter. The RNA polymerase then starts the process of pulling the two DNA strands apart and so allows the start of the transcription process: Figure 1.1 describes the process of transcription in a simple manner.

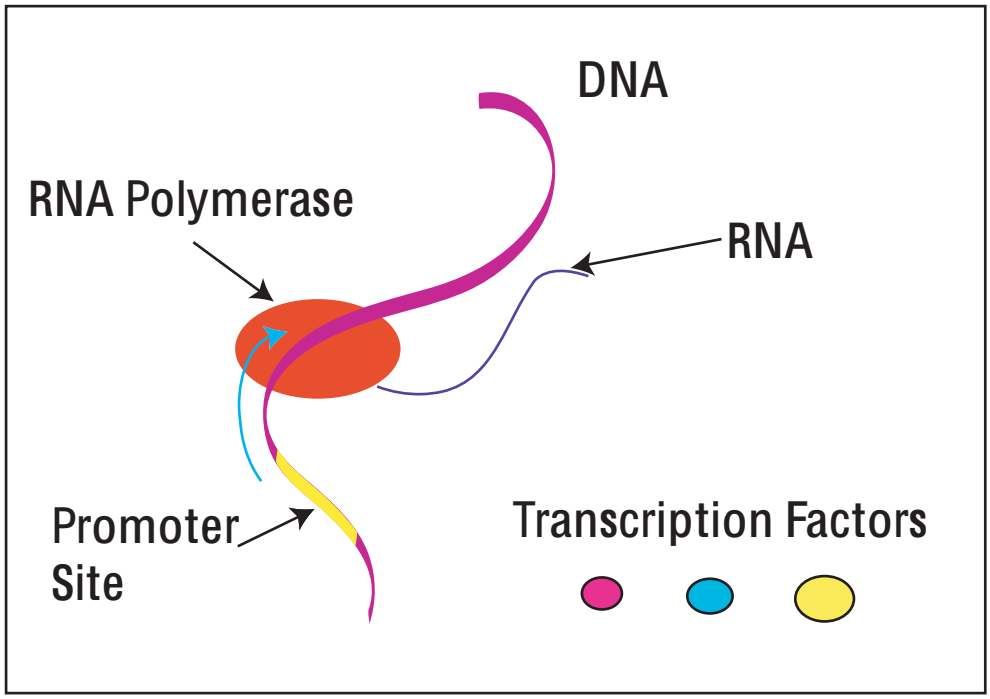

Figure 1.1: Transcription: Here we can see that once the transcription factors are bound to the promoter site the RNA Polymerase moves up the DNA strand copying the DNA to produce RNA.

This seems to suggest that DNA transcription does not progress in a straightforward manner. The process stops and starts depending on whether or not the transcription factors are in place. Additionally the presence of other small molecules colliding with the bound transcription factors can affect the rate of transcription. Finally, one should also consider that the numbers of molecules involved in tran-

scription are very small: a typical plant cell may contain very few RNA polymerases.

There are several types of RNA that are produced in the cell. The RNA molecules that are copied from the section of the DNA that correspond to the amino acid sequence of proteins are called *messenger RNA* (mRNA).

The sketch above is really only correct for the relatively simple molecular biology of *prokaryotic* cells—those, such as bacteria, that lack a nucleus. In *eukaryotes*—cells that have a nucleus—the transcription cycle is a lot more complicated. The RNA molecule whose production is described above is called the *primary transcript*, but this initial transcript undergoes several processes before it is known as mRNA. The primary transcript is located in the cell's nucleus while the ribosomes, on which protein synthesis takes place, are located in the *cytoplasm*. The eukaryotic cell is thus comprised of two main compartments. The nucleus found in the center where we usually find the chromosomes while the cytoplasm is a jelly-like substance which, though contained within the *plasma membrane* or outer wall of the cell, lies outside the nucleus. The primary transcript undergoes RNA *capping* and *polyadenylation*, two processes that increase the stability of the mRNA molecule and aid its transport from nucleus to cytoplasm.

RNA capping involves a guanine (G) nucleotide being attached to the front end of the transcript (known as the 5' end) while polyadenylation adds a special structure to the tail of the transcript also known as the 3' end. The 3' end is first trimmed by an enzyme and then a second enzyme adds on a repeated number of adenine (A) nucleotides and is sometimes referred to as the poly(A) tail. This RNA then decreases in size due to RNA splicing. The primary transcript contains a lot of non-coding regions, whose function is not fully understood. Theses non-coding sequences are known as *introns*, while the coding sequences are called *exons*. The end result

of splicing is a much shorter RNA molecule, which contains an uninterrupted coding sequence. After splicing a functional mRNA molecule is now able to leave the nucleus and be translated into protein.

After transport to the cytoplasm, the mRNA molecules are eventually degraded by the cell. The amount of protein produced by a single mRNA molecule depends, among other things, on its lifespan in the cytoplasm. The lifespan varies depending on what sort of cell it is produced in.

Once the mRNA molecule has left the nucleus and before it is degraded it is available for *translation*: conversion of the information from one chemical language to another. The translation is not a one-to-one correspondence between a nucleotide in RNA and an amino acid in protein. The reason for this is that there are 20 different amino acids and only four nucleotides. The nucleotide sequence is translated into protein–another polymer whose basic elements are amino acids—sequence by a set of rules, the genetic code.

The code reads out the nucleotides of an RNA molecule in groups of three so in total there are sixty-four possible combinations of three nucleotides but only twenty amino acids. Most of these nucleotide triplets are redundant in the sense that most amino acids are specified by more than one triplet. Each of these triplets is known as a *codon*, and each corresponds to an amino acid or to a special marker indicating the stop-point for translation.

Transfer RNAs (tRNAs) are an adaptor molecule that binds to both the codon and the corresponding amino acid. Accurate and rapid translation of mRNA into protein requires a molecular machine which can travel along the mRNA chain, holding the associated tRNA molecules in place and bonding together the associated amino acids to form a protein chain. This machine is a very large complex consisting of more

than fifty proteins, several RNA molecules (rRNAs) and is called the *ribosome*.

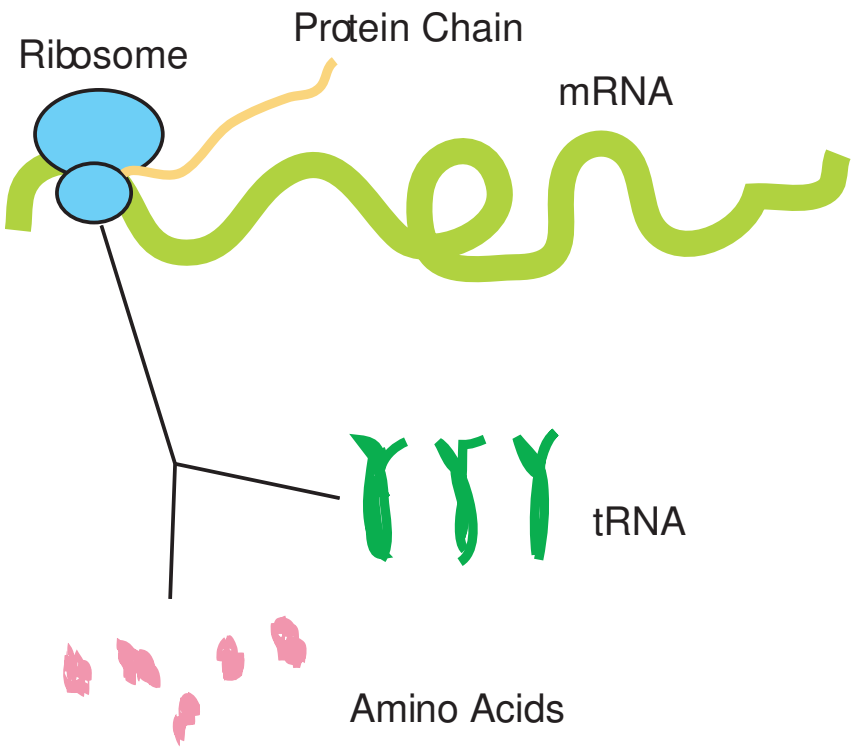

Figure 1.2: Translation

Once protein synthesis has started, a new amino acid is added to the chain in a cycle of reactions. Translation starts with a certain codon (AUG) and a special initiator tRNA. The translation process stops with the presence of one of several codons (UAA, UAG, or UGA), that signal to the ribosome to stop translation. The protein chain is then released into the cytoplasm. The RNA molecule is then released by the ribosome ready to start another round of protein synthesis. This suggests that translation, unlike transcription could be a linear process.

1.2 Gene Regulatory Networks

Having introduced key biological ideas in the previous section we may now introduce the notion of *gene regulatory networks*. A gene regulatory network describes how a collection of proteins and their corresponding mRNAs and DNA sequences

interact within a cell.

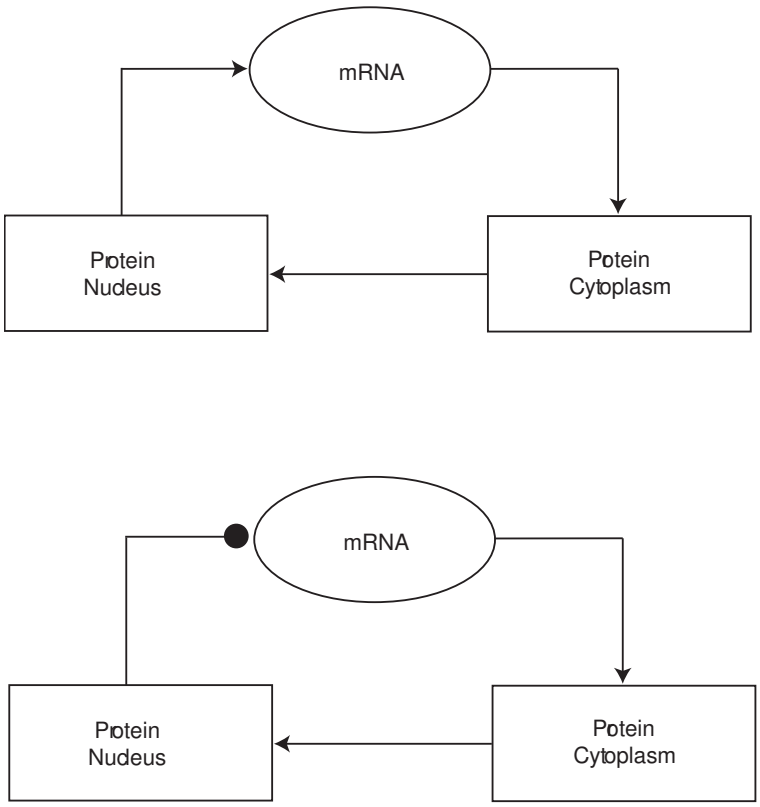

Figure 1.3: Examples of autoregulatory networks

Figure (1.3) illustrates two simple autoregulatory networks. Arrows correspond to an enhancing effect and the dots an inhibitory effect—this notation will be used throughout this thesis. The top network in the figure is known as a positive feedback loop while the one below as a negative feedback loop. A more complex example, the *segment polarity network*, which is active during the early development of the fruit fly *Drosophila melanogaster*, is illustrated in figure (1.4).

The network consists of a set of genes that are responsible for establishing certain repeated structures–precursors of the adult insect's segments—during early development. A group of mathematicians and biologists at Washington University, von Dassow et al., have developed a system of first-order non-linear differential equations to model the segment polarity network [45, 47, 46].

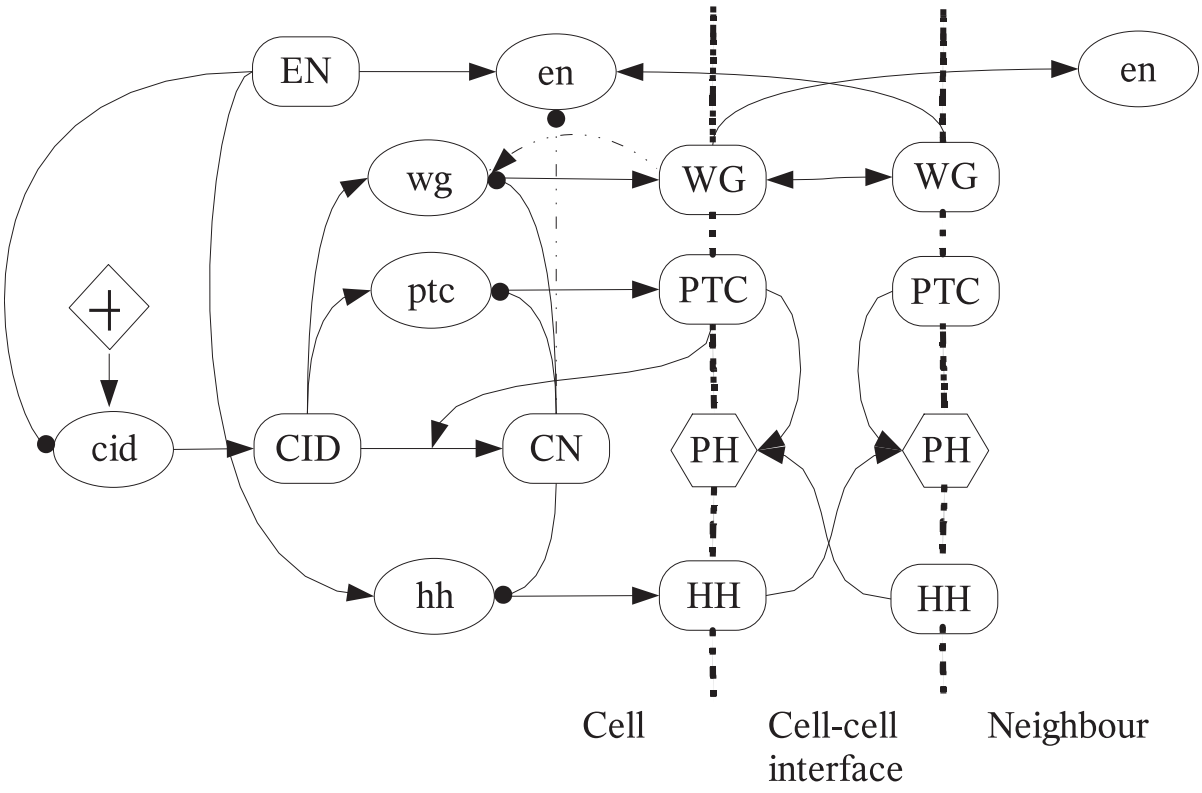

Figure 1.4: Segment Polarity Network

The network shows how proteins activate and inhibit the production of genes within adjacent cells. Simply put, proteins inhibit or activate the transcription process. The ellipses correspond to mRNAs and the rectangles the proteins. The model also includes some *trans-membrane* proteins: these are the proteins that pass through a cell's membrane and mediate communication with adjacent cells. Interactions between the following five genes have been considered in their model: EN, engrailed; WG, wingless; HH, hedgehog; PTC, patched; CID, cubitus interruptus; CN, repressor fragments of cubitus interruptus; PH, patched-hedgehog complex. The dashed lines correspond to interactions conjectured after initial numerical experiments and which have subsequently been verified experimentally. This figure represents the interactions von Dassow et al. thought would replicate the assembly patterns of the segment polarity genes.

Networks, like the segment polarity network, can be modeled with the use of ODEs. The segment polarity network consists of a system of thirteen non-linear ODEs, requiring 48 free parameters, including half-lives of the molecules, binding rates and cooperativity coefficients. Experimental values of these parameters are unknown and plausible ranges for these parameters—established by very crude arguments about the duration of the developmental period and energetic constraints on maximum rates of protein production—-span several orders of magnitude. A *solution* in the sense of von Dassow et al. is a set of parameters gives rise to behavior that is qualitatively consistent with the behavior of real cells in an embryo: their main finding is that the regulatory network is highly robust in the sense that almost every remotely plausible parameter value appears as part of some solution. For more detail with regards to the biology and modeling behind this network we refer the reader to their papers [45, 47, 46].

Modeling gene regulation with the use of differential equations assumes that the concentrations of molecules vary deterministically and continuously. This does not seem to be the case, as the number of molecules is small and may vary from cell to cell. A growing body of literature, discussed in the rest of this chapter, examines the role of stochasticity in gene networks. Much of this work draws a distinction between *extrinsic noise* and *intrinsic noise*. The former includes cell-to-cell variations in concentrations of such cellular components as metabolites, polymerases and ribosomes. By contrast, intrinsic noise refers to fluctuations in the timing of cellular events within a single cell, including variations in the timing of he initiations of transcription.

1.3 Stochasticity in Gene Expression

Considering stochasticity in gene networks is currently a very hot topic in mathematical biology. Stochastic models have been considered since biological processes such as transcription usually involve small numbers of molecules. A key question is to ask how low can we take molecule numbers without affecting the desired qualitative behavior of the models.

The review by Kaern et al. [27] on *Stochasticity in Gene Expression* provides useful insight into how stochasticity affects gene expression from a theoretical point of view. Gene expression needs to be a robust process as it is invariably under a lot of stress from constant environmental changes. Genetically identical cells do vary substantially in terms of molecular content even when under the influence of similar environmental conditions. Kaern et al. measure gene expression noise by considering the standard deviation divided by the mean, of protein concentrations, and noise *strength* as the variance divided by the mean. Describing the noise strength

in this way however does seem rather crude.

Kaern et al. look at stochastic models and describe circumstances under which a stochastic model converges to a deterministic one. The two points in question are system size and reaction speeds. By system size we mean the numbers per molecular species in a fixed volume.

We will start with the *finite number effect* which examines system size. Consider an eukaryotic cell in which a chemical species X is said to be in equilibrium in the sense that the concentration of species is the same in the nucleus and in the cytoplasm. But the nucleus is a much smaller entity than the cytoplasm. So if a molecule left the nucleus and entered the cytoplasm, it would cause a bigger percentage change in nuclear concentration of X than the cytoplasmic concentration. This change is what is known as the *finite number effect*. For example, if 10 molecules are present in the nucleus and 100 in the cytoplasm, the relocation of 1 molecule from the nucleus to the cytoplasm has a 10 percent change in nucleus concentration but only 1 percent change in the cytoplasmic concentration. In general, let N denote the average molecular abundance in the cytoplasm and nucleus respectively, $\eta = \sigma/N$ be the coefficient of variation, where σ is the standard deviation of the number of protein molecules. A decrease in abundance results in a $1/\sqrt{N}$ scaling of the noise.

In a hypothetical experiment we would like to decrease system size in such a way that it does not affect the steady state protein number; so we can see the affect a change in system size has on the protein distribution. This can be done by decreasing the transcription rate but increasing the translation rate in an appropriate manner. We find a decrease in system size leads to an increase in noise which gives a broader protein distribution. This proportional change is known as *translational bursting*. Which implies we have an increase in heterogeneity i.e. cell-to-cell variation

of protein production in genetically identical cells.

Very few of the reactant parameters, concentrations and rates have been measured. Worse still, many are not yet experimentally accessible and those quantities that can be measured, for example, metabolic fluxes, don't give any real insight into underlying chemical rates; since fluxes depend on both concentrations and rates. That is, the experimental data provide constraints of the parameters, but do not determine them. As we will see below, rate variation, especially in those affecting the binding of regulatory elements, have important consequences for the strength of stochasticity in gene expression.

The review also looks at a second point, fast reaction speeds. Fast reaction kinetics for reversible reactions leads to a deterministic approach. Slow reaction kinetics mean we stay in a state for a longer amount of time. For example if a promoter is in an active state, leading to the transcription of mRNA, then more mRNA is produced quickly. This is known as a *transcription burst*. If the promoter remains active long enough, protein production will follow the state of the promoter. Which leads to a bimodal protein distribution, where protein is produced at either a very high or a very low rate.

Kaern et al. find that translational bursting is valid when transitions between promoter states are quick. The biochemical processes regulating transcription initiation are much faster than protein synthesis and degradation. They find this is the main cause of gene expression noise in prokaryotic cells. Eukaryotic cells have slow reaction kinetics in the nucleus, because all the DNA is packed tightly due to the presence of nucleosomes. *Nucleosomes* package DNA into chromosomes inside the cell's nucleus and control gene expression.

A recent review by Paulsson [34] discusses the similarities between the vast amount

of stochastic models over the last 30 years for gene expression. A set of inspirational papers were first produced in the early 1970s by David Rigney [39, 38, 37] and Otto Berg [33] describing mRNA and protein fluctuations in cells. One of their main findings was the discovery of translational bursting which we described earlier. Their analytical stochastic models showed that if mRNAs are either translated or degraded with constant probabilities leads to a widening of the protein distribution.

Over twenty years later Kepler and Elston [28] developed a model which was in the same vein as Rigney and Bergs. Their model assumed that mRNAs are degraded quickly and they discovered that genes which switched on and off slowly produced large protein fluctuations, widening of the distribution of protein concentrations. Thattai and van Oudenaarden [43] extended the model further by examining a negative feedback loop and noise propagation. They found that for an autoregulatory protein, negative feedback effectively led to a reduction in system noise. Stochastic models developed in recent times have been simulated numerically with the aid of the Gillespie Algorithm [14, 15, 16, 17], which is discussed in Chapter 2. The conclusions drawn from the numerical simulations agree well with the earlier analytical studies.

1.4 Experimental Studies of Stochasticity

Elowitz [12] examined strains of *E. coli* to try and discriminate between intrinsic and extrinsic noise. Their results show that both types of noise contribute to cell-to-cell variation in gene expression. They were able to look at protein production with varying amounts of noise by constructing strains of *E. coli* where the reporter genes, yellow fluorescent protein (YFP) and cyan fluorescent protein (CFP), were controlled by identical promoters and regulatory sequences.

These modified strains of *E. coli* produced only 3 to 6 percent as much protein as

the wild type, because they had much lower transcription rates. Elowitz constructed plots of normalized CFP against normalized YFP, for both the wild type and the other strains (see figure(1.5)). Here variation along the main diagonal corresponds to extrinsic noise and variation perpendicular to the main diagonal corresponds to intrinsic noise.

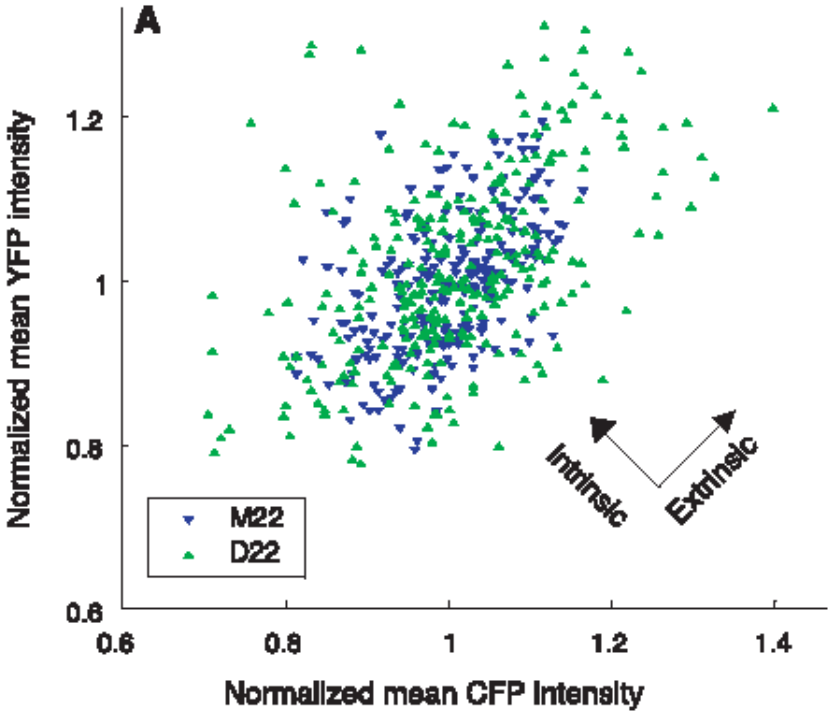

Figure 1.5: Plot of fluorescence in two strains: (M22) quiet and (D22) noisy. Every point represents mean flourescence intensities from one cell.

The two strains under consideration in Figure (1.5) are M22, the least noisy strain, and D22, the noisy wild type. Each point in the figure represents mean fluorescent intensities from one cell, over some small time interval. Note that the wild type had a much lower transcription rate than M22.

They found both intrinsic and extrinsic noise increased by a factor of 5 in the wild type. Figure (1.6) are plots of noise versus rate of transcription for strain M22 in (B) and D22 in (C). The x-axis represents fluorescence levels, where the rightmost

point was used to normalize all fluorescence intensities. η_{tot}, η_{int} and η_{ext} represent, respectively, total, intrinsic and extrinsic noise. Intrinsic noise increases for both strains as transcription rate decreases, as expected. This agrees with the analytical models proposed by Swain et al. [42] and Paulsson et al. [35], which predict an increase in intrinsic noise as the transcription rate decreases.

Notice in Figure (panel C 1.6) that extrinsic noise behaves in a very different way to intrinsic noise. η_{int} decreases monotonically however η_{ext} exhibits a maximum value for some intermediate rate of transcription in the wild-type. These results suggest that cell-to-cell variability is not totally due to intrinsic noise. The take home point is that both extrinsic and intrinsic noise together give rise to variation.

Colman-Lerner et al. [7], working with yeast cells, did a similar series of rigorous experiments that suggest cell-to-cell variation comes mainly form the amount of translational machinery available in the cytoplasm. They found that random fluctuations during transcription and translation accounted for only a small amount of cell-to-cell variation.

They looked at the response of *Saccharomyces cerevisiae*, (Brewer's yeast) cell to an α factor, a pheromone that triggers a signalling cascade whose ultimate outcome is a decision whether to switch from normal, vegetative growth to the initiation of mating events. They divide the signal transduction system into two subsystems. The first, the *pathway subsystem*, shown in a blue box in figure (1.7), includes all the events leading up-to, but not including, the initiation of transcription. The second compartment, the *expression subsystem*, shown in a red box in figure (1.7), includes the initiation of transcription and protein translation.

They conducted two experiments, whose results can be seen in Figure (1.8). The first was designed to measure gene expression noise. They produced cell lines in which

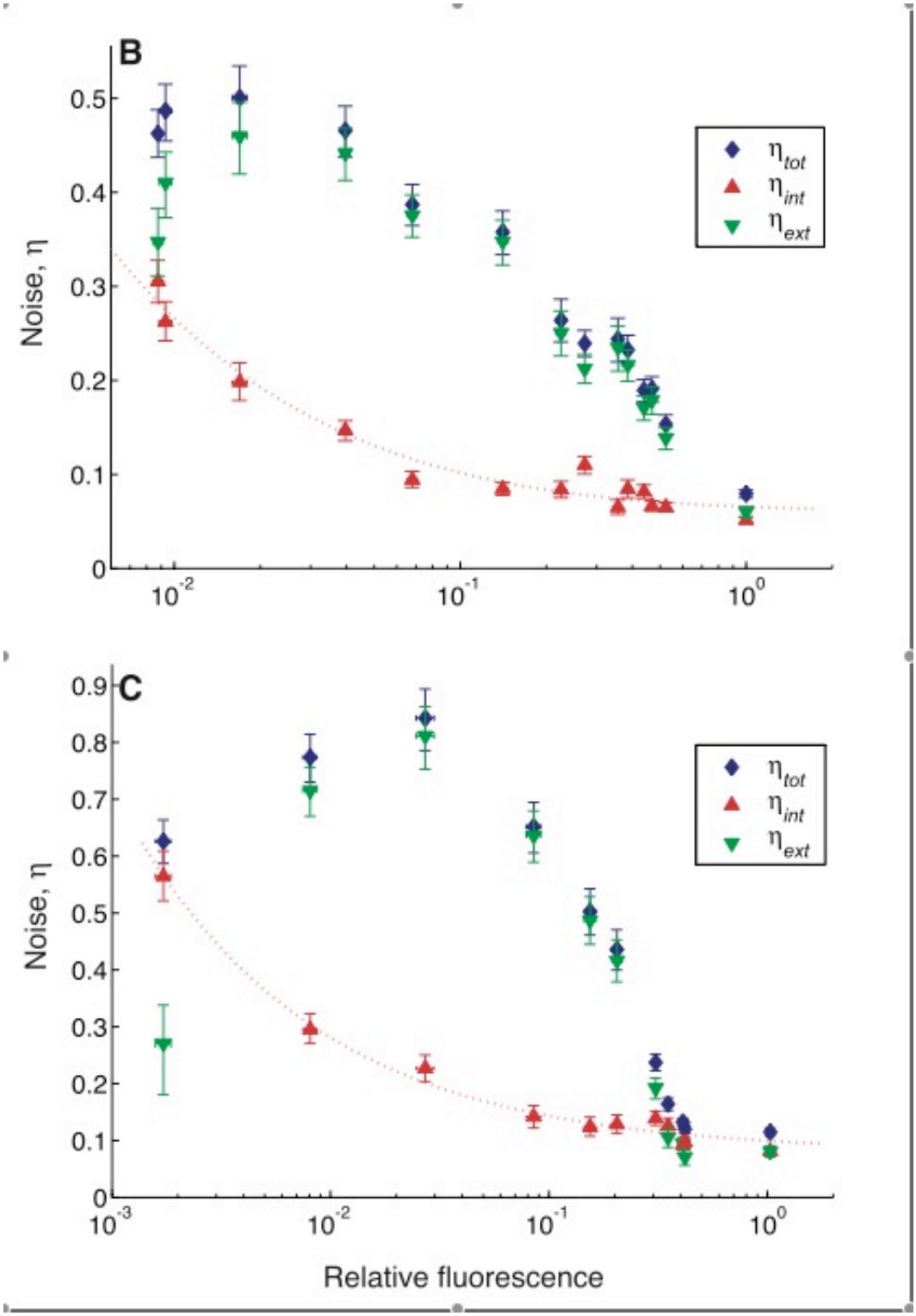

Figure 1.6: Panel B: Noise versus transcription rate for M22. Panel C: Noise versus transcription rate for D22 (see text for in-depth discussion).

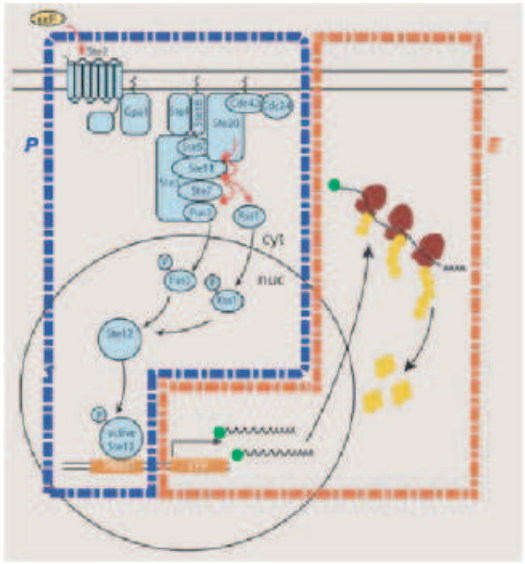

Figure 1.7: The mating pheromone response system.

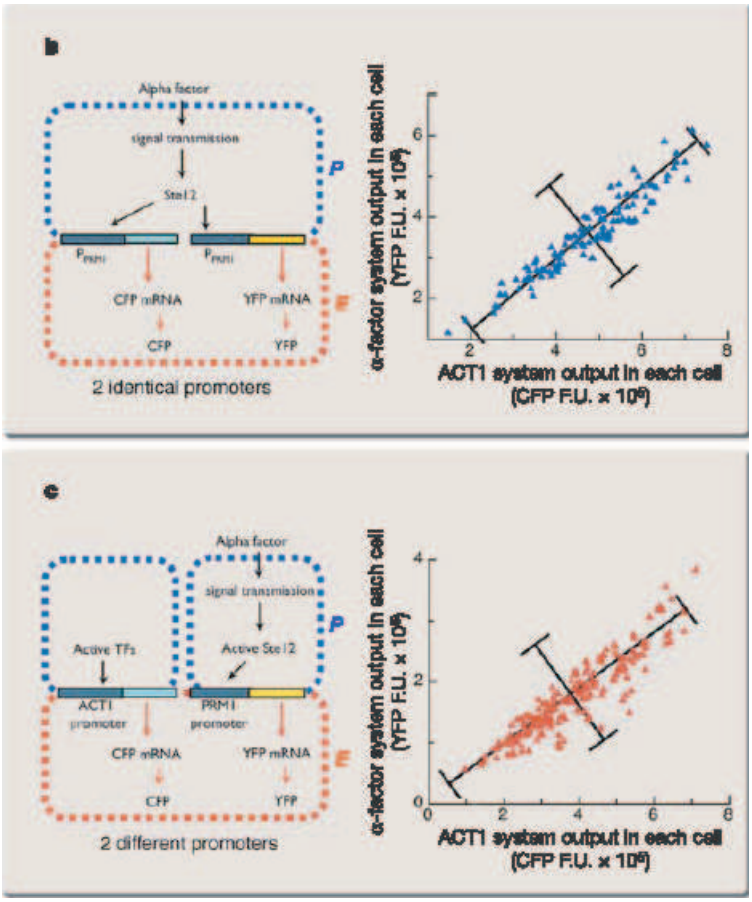

Figure 1.8: Experimental results. Note that the x-axis in panel b should be labeled α-factor system output and not ACT1 system output.

fluorescent reporter genes yellow fluorescent protein (YFP) and cyan fluorescent protein (CFP) were regulated by the same α-responsive promoter. The righthand plot in panel (1.8b) has a point for each of a large number of cells. The x-coordinate is the intensity of CFP fluorescence while the y-coordinate is YFP intensity. If intrinsic noise were very strong, the two fluorescent signals would be uncorrelated, even within a single cell, and so one would expect a circular cloud of points. Instead, Colman-Lerner see strongly correlated variation, so that, for example, cells that express CFP strongly also express YFP strongly.

In the second experiment—designed to explore the extent of cell-to-cell variation in the α-transduction machinery—the fluorescent reporters are under the control of different promoters, one (controlling YFP expression) that is α-responsive and one (governing CFP expression) that is not. The results, displayed in the right part of panel (1.8c) are similar to those of the first experiment: expression levels within a single cell are strongly correlated. The results from both experiments implied that the variations in the capacity of cellular "machinery", which includes, for example, the RNA polymerases used during transcription as well as the ribosomal complexes needed for translation, is the main source of cell-to-cell-variation.

The foregoing observations demonstrate that there is considerable stochasticity in gene expression and prompt the questions; how do cells cope with fluctuating signals at macroscopic level and what effect does this noise have on regulatory networks? Blake et al. [4] examine noise in eukaryotic gene expression, also in *Saccharomyces cerevisiae*. They developed a gene cascade with two regulatory steps to look in detail at the way in which transcription noise propagates through a simple network.

For the experiment they looked at the expression of a gene, *GAL1*, whose role is to adapt the cell's metabolism, allowing the yeast to use the sugar galactose when its

preferred carbon source, glucose, is unavailable. They modified the gene's regulatory region so that, in addition to *GAL1*'s natural regulation by galactose, the gene can also be reversibly repressed by a reagent not normally present in cells. As in the experiments described above, they used a fluorescent reporter (here, *yEGFP*—yeast Enhanced Green Fluorescent Protein). They also constructed a stochastic model, which took into account key transcriptional processes such as fluctuations in mRNA production due to premature polymerase detachment and re-initiation of translation and which also encompassed slow transitions between promotor states: such slow activator/repressor kinetics led, as expected, to noisy, bursting responses and prolonged bistable expression states.

1.5 Summary

In this chapter we have introduced the biological concepts needed to understand gene regulatory networks. A very complex example of such a network was discussed in section (1.2). But the segment polarity network is just the tip of the iceberg in terms of complexity, gene networks can involve hundreds of genes with thousands of unknown parameters.

We have seen, through Colman-Lerner's experiments, that extrinsic noise is a key player with regards to providing a reason for cell-to-cell variation for genetically identical cells. Elowitz suggested that both intrinsic and extrinsic noise lead to variation, suggesting there exists a correlation between the two. Blake shows that stochasticity arising from transcription is a large contributor to cell-to-cell variation among genetically identical cells. Furthermore they explored the propagation of noise through a gene network with the aid of elaborate experiments. Finding that an increase in noise in transcribing a regulatory protein led to an increase in cell-to-cell

variability.

The reviews by Kaern and Paulsson describe the current trend to model transcription events using discrete stochastic models. These models are being considered due to the low numbers of molecules involved within interactions in the networks. We have seen that the experimental findings agree with the analytical models as seen in Blake. We may conclude that when modeling biological networks biological noise due to low copy numbers needs to be considered.

CHAPTER II

Mathematical Background

In this chapter we will introduce the mathematical concepts behind modelling gene networks deterministically [8, 41] and stochastically [44, 13, 48]. By considering a simple example, an enzyme reaction , we will introduce Michaelis-Menten kinetics. We will then introduce an idealized example of transcription and translation which we will model stochastically in three ways testing for robustness. Note that concentrations of species will be denoted by, [], and numbers of a chemical species, N_k where k is the species in question. All the numerical simulations throughout the thesis were done using C++ [36, 29, 26] and the graphical representation of the data with MATLAB [25].

2.1 Enzyme Kinetics

Enzymes are biological catalysts, proteins, that help speed up chemical reactions. Reactions still occur without the presence of an enzyme but at a rather slower rate known as the basal rate. The *activation energy*, the energy required to move from one state to the next, is higher when no enzyme is present and lower when the appropriate enzyme is present. Enzyme-catalyzed Reactions can become over a million times faster! Enzymes are key building blocks of life. What exactly happens when an enzyme and substrate come together? We can imagine an enzyme to be a ball with

a piece missing on the surface. The right substrate will be able to fit in this hole rather like a puzzle. Once bound we have an enzyme-substrate complex. From the complex to the release of the enzyme and its product is known as the catalytic step.

The standard mathematical framework for the study of enzyme kinetics is so-called Michaelis-Menten kinetics which we will illustrate with an example. The following exposition owes a great deal to John Gillespe [18].

The enzyme Lactose dehydrogenase,(LDH) is a key enzyme involved in the *Cori-cycle*, it is responsible for recycling lactic acid produced by skeletal muscle. Lactic acid, also known as lactate, is transported through the blood from the muscle into the liver where it is converted into an acid called *pyruvate*, this reaction is catalyzed by LDH. Adaptations of LDH to different temperatures has been studied a great deal over the years in a number of different species ranging from Antarctic fish to desert lizards. One of the most interesting parameters to come out of the study was the *Michaelis* constant, K_m. To see what this parameter is and why it was so interesting let us consider the following reaction, known as the *Michaelis-Menten* model for enzyme catalysis.

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_{cat}} E + P \tag{2.1}$$

Where $k_i's$ are rate constants. The binding and dissociation of the enzyme, E, to substrate, S, is treated as a reversible reaction and is considered to be very fast. The second step is a simple first order reaction describing the catalytic step and release of the product, P.

Having written down our first chemical reaction, we now implement the Law of Mass Action, which states that the rate of a chemical reaction is the product of

the concentrations of the chemical species involved, to write down the following rate equations.

$$\frac{d[ES]}{dt} = [S][E]k_1 - k_{-1}[ES] - k_{cat}[ES] \tag{2.2}$$

$$\frac{d[P]}{dt} = k_{cat}[ES] \tag{2.3}$$

We will express enzyme and substrate concentrations in terms of their initial concentrations $[E_0]$ and $[S_0]$, by using the following stoichiometric relationships $[E_0] = [E] + [ES]$ and $[S_0] = [S] + [ES]$, such that we can eliminate $[S]$ and $[E]$ from (2.2). We will now calculate the steady state, $d[ES]/dt = 0$, value for $[ES]$.

$$\begin{aligned} \frac{d[ES]}{dt} &= 0 \\ \Rightarrow [ES] &= \frac{k_1[E][S]}{k_{-1} + k_{cat}} \end{aligned} \tag{2.4}$$

Define,

$$K_m = (k_{-1} + k_{cat})/k_1 \tag{2.5}$$

then,

$$[ES] = \frac{[E][S]}{K_m} \tag{2.6}$$

We know that the rate of production of $[P]$ is, $d[P]/dt = v = k_{cat}[ES]$. Using stoichiometric relationships and plugging in the steady state value for $[ES]$ we obtain the Michaelis-Menten Equation (2.7).

$$v = \frac{V[S]}{K_m + [S]} \tag{2.7}$$

$$V = k_{cat}[E_0] \tag{2.8}$$

Where K_m is the Michaelis constant, mentioned earlier, is a measure of the strength of the binding between the substrate and the enzyme. A low value for K_m corresponds to strong binding and a high value for weak binding. V is known as the *maximum rate*, which occurs when the concentration of the substrate is high enough to saturate the enzyme. Figure (2.1) describes how the velocity of the reaction varies with changing substrate concentrations. Notice that when the substrate concentration is equal to K_m, velocity of the reaction is equal to half the maximum rate. If the substrate concentration was much lower than K_m we can see that the reaction behaves like a first order reaction.

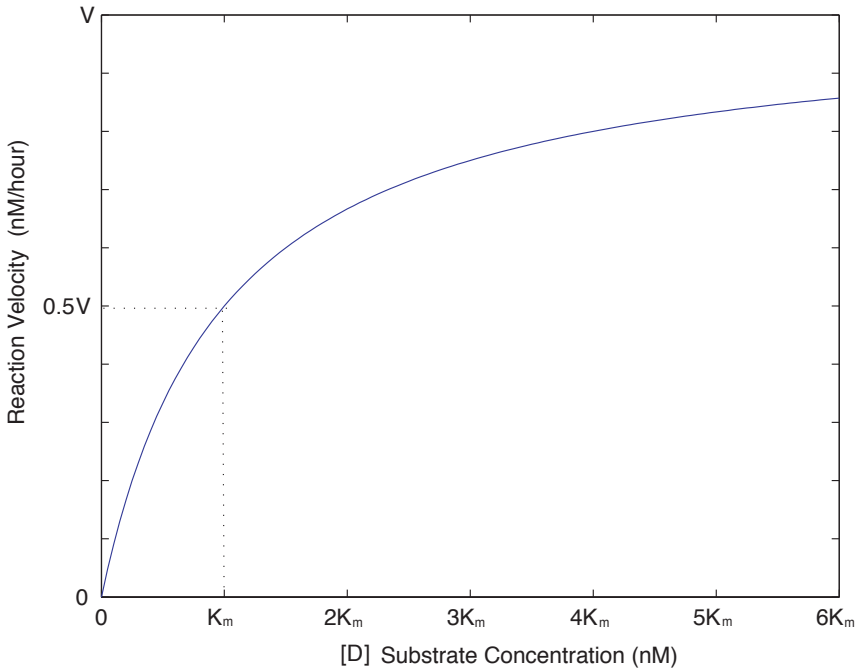

Figure 2.1: Example of Michaelis Menten Reaction

A generalization of equation (2.7) is,

$$v = \frac{V[S]^n}{K_m^n + [S]^n}. \tag{2.9}$$

Where n is known as the *Hill coefficient* such that,

- $n > 1$: Positive cooperation, once one substrate molecule binds to an enzyme its affinity with other substrate molecules increases.
- $n < 1$: Negative cooperation, once one substrate molecule is bound its affinity decreases.
- $n = 1$: No cooperation, the affinity doesn't matter on whether any substrate molecules are bound.

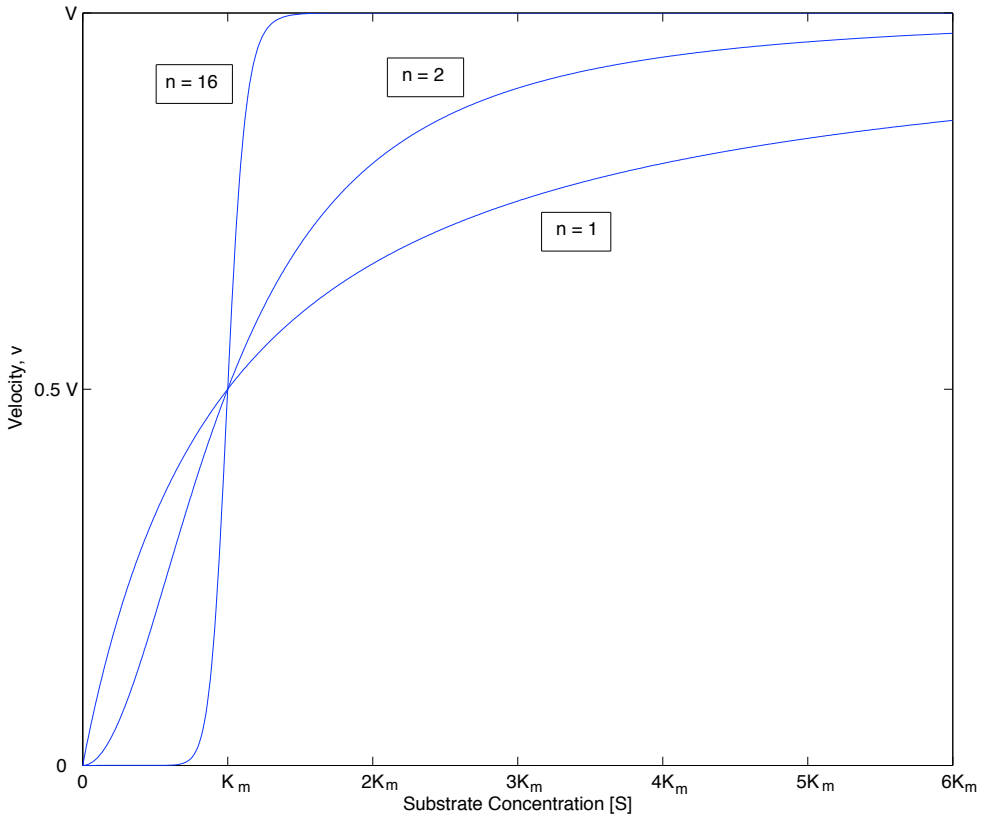

Figure 2.2: Example of Michaelis Menten Reaction with enzyme cooperativity

Figure(2.2) shows what effect Hill coefficient can have on an enzyme reaction. So as we increase n a more switch like behavior like a heaviside function occurs. From both figures(2.1,2.2) we can see how Michaelis Menten kinetics are useful in describing changes of states in enzyme kinetics. This ends our brief introduction to

Michaelis-Menten and deterministic enzyme kinetics.

2.2 Stochastic Processes

In this section we will give a brief review of stochastic processes, starting with definitions of a *Wiener Process* and *Itô Process*, leading onto the concept of Stochastic Differential Equations. Finally at the end of the section we introduce the *Euler-Maruyama* method used to solve SDEs numerically.

2.2.1 Wiener Processes and Itô process

Definition II.1. A standard *Wiener Process*, a stochastic process, on $[0, T]$ is a random variable, $W(t)$ which depends continuously on $t \in [0, T]$ such that:

1. W(0) = 0
2. For any $0 \leq t_1 < t_2 \leq T$, the increment $W(t_2) - W(t_1) \sim N(0, t_2 - t_1)$
3. All increments on non-overlapping time intervals are independent. $W(t_2) - W(t_1)$ and $W(t_4) - W(t_3)$ are independent for all $0 \leq t_1 < t_2 < t_3 < t_4 \leq T$.

From parts (2) and (1) of definition (2.2.1), choosing $t_2 = t$ and $t_1 = 0$, we see that $W(t) \sim N(0, t)$. A variable z follows a Wiener process if it satisfies the properties of definition (2.2.1). From part (2) of definition (2.2.1) a change Δz, during a small time period Δt, follows a normal distribution with mean zero and variance Δt. The third property implies that z follows a *Markov process*. A process is a Markov process if it has the *Markov property*, the probability distribution of future states only depends on the current state and not on past states.

Define a *generalized Wiener process* for a variable X by,

$$dX = a\,dt + b\,dz \tag{2.10}$$

where a and b are constants and dz follows a Wiener process. The term adt is the mean behavior and bdz is the variability to the path followed by X. So bdz is seen as adding noise to the deterministic equation, $dX/dt = a$. The amount of noise being b times a Wiener process. For a small time interval Δt, a change ΔX follows a normal distribution with mean $a\Delta t$ and variance $b^2\Delta t$.

Note that equation (2.10) is not a differential equation as such, it simply represents a trajectory subjected to random fluctuations. To solve such an equation we would transform it into a P.D.E. such as the Fokker-Planck Equation, whose solution would describe the temporal evolution of the probability distribution.

An *Itô process* is a generalized Wiener process where the parameters a and b are functions of the underlying variable X and time t. The mean and variance are $a(t, X)$ and $b(t, X)^2$ respectively.

2.2.2 Stochastic Differential Equations

For the deterministic case we have,

$$\frac{dX}{dt} = f(t, X(t)) \tag{2.11}$$

$$I.C. X(0) = X_0 \tag{2.12}$$

$$\Leftrightarrow X(t) = X_0 + \int_0^t f(s, X(s))ds, \tag{2.13}$$

where I.C. represents the initial condition, similarly for the stochastic case,

$$X(t) = X_0 + \int_0^t f(s, X(s))ds + \int_0^t g(s, X(s))dW(s), \tag{2.14}$$

where f, g are scalar functions and X_0 (I.C.) can be interpreted as an SDE.

$$dX = f(t, X)dt + g(t, X)dW \tag{2.15}$$

$$I.C. X(0) = X_0 \tag{2.16}$$

This is an Itô SDE, if the integral is interpreted as an Itô integral. Note that dW has mean zero and variance dt. From definition (2.2.1) we found that $\Delta W_k \sim \sqrt{\Delta t}N(0,1)$ which we can re-write as,

$$dW = \phi\sqrt{dt} \tag{2.17}$$

where ϕ, $-\infty < \phi < +\infty$, is a random variable drawn from the normal distribution, whose p.d.f. is,

$$\frac{\exp(-\frac{1}{2}\phi^2)}{\sqrt{2\pi}}. \tag{2.18}$$

Also we find that the expected value of ϕ is zero and has variance equal to one. In a small time interval $(t, t+\Delta t)$ (2.15) becomes,

$$\Delta X = f(t,X)\Delta t + g(t,X)\phi\sqrt{\Delta t}. \tag{2.19}$$

This equation involves taking small approximations and assumes that the drift and variance are still $f(t,X)$ and $g^2(t,X)$ respectively in the small interval $(t, t+\Delta t)$.

2.2.3 Numerical Simulation of S.D.E.s

We will now look at a numerical method [29, 36] for solving the following Itô S.D.E.

$$dX = f(X,t)dt + g(X,t)dW(t) \quad t \in [0,T] \tag{2.20}$$

For computer simulations, a discrete time model for $W(t)$ is obtained as follows:

Define grid: $t_k = k\Delta t$, $k = 0, 1, ..., N$

Denote $W_k = W(t_k)$, then from part (1) of definition $\Rightarrow W_0 = 0$.

Parts (2) and (3) $\Rightarrow W_{k+1} = W_k + \Delta W_k$, $k = 0, 1, 2, ..., N-1$, where the ΔW_k

are independent R.V.'s such that $\Delta W_k \sim N(0, \Delta t)$, or $\Delta W_k \sim \sqrt{\Delta t} N(0,1)$ (this is useful since it can be calculated by programs).

Definition II.2. Setup the following grid for the *Euler-Maruyama Method*:

$t_k = k\Delta t$, $k = 0, 1, ..., N$ with $N\Delta t = T$.

$X_k \equiv \tilde{X}(t_k)$: numerical solution at t_k.

$X(t_k)$: exact solution at t_k.

$$X_{k+1} - X_k = f(X_k, t_k)\Delta t + g(X_k, t_k)\Delta W_k \tag{2.21}$$

with $\Delta W_k \equiv W_{k+1} - W_k = z\sqrt{\Delta t}$, where $z \sim N(0,1)$.

The Euler-Maruyama method is said to have strong order of convergence of a $1/2$,

$$E(|X_N - X_{t_N}|) = O(\Delta t^{1/2}). \tag{2.22}$$

Thus strong order of convergence is based on the mean of the error at a fixed time.

The problems discussed throughout are stiff systems so a predictor-corrector method is needed. We implement the explicit Euler Method first, then implement the Euler-Maruyama method:

$$\tilde{X}_{k+1} - X_k = f(X_k, t_k)\Delta t \tag{2.23}$$

$$X_{k+1} - \tilde{X}_{k+1} = f(\tilde{X}_{k+1}, t_k)\Delta t + g(X_k, t_k)\Delta W_k. \tag{2.24}$$

Note that Mersenne Twister was used to generate random numbers for all numerical algorithms considered.

2.3 Stochastic Modelling of Enzyme Kinetics

When we are dealing with small copy numbers of molecules, should we still be dealing with concentrations or should we be dealing with actual numbers of a chemical species? In this section we will discuss two stochastic approaches to enzyme kinetics which are designed to deal with small and larger numbers of molecules respectively. We will also introduce a model to discuss the effects white noise perturbations has on parameter values, namely the transcription and translation rates.

2.3.1 Chemical Master Equation

We are going to talk about simulations that account for every particle in some fixed volume, which is assumed small enough that it is well-mixed in the sense that all the particles encounter each other frequently. If there are n chemical species, denoted by S_i for $i = 1....n$, in play we will specify the state vector, X_i, of the system with whole numbers. Define $P(X,t)$, be the probability that there will be X_i of chemical species S_i in a fixed volume at time t. Let us now consider a very small time interval $(t, t+dt)$ where at most one reaction can occur. If there are m reactions, $a_1(X), ..., a_m(X)$, we will define the *stoichiometric vectors* $v_1, ..., v_m$ to be the result of reactions on a state vector X_i. Thus we have a possibility of $m+1$ distinct states we could be in at time t that takes us to state X at $(t+dt)$.

$$P(X,t+dt) = P(\text{no states change over dt})P(X,t) + \sum_{j=1}^{m} P(X-v_j,t)P(\text{state change to X over dt}) \tag{2.25}$$

From Gardiner[13] this gives us the following Chemical Master Equation.

$$\frac{\partial P(X,t)}{\partial t} = \sum_{j=1}^{m} a_j(X - v_j, t) - a_j(X)P(X,t) \tag{2.26}$$

We will now implement the CME for the types of reactions considered earlier in the chapter and how we can solve such equations analytically.

Uni-molecular Reaction

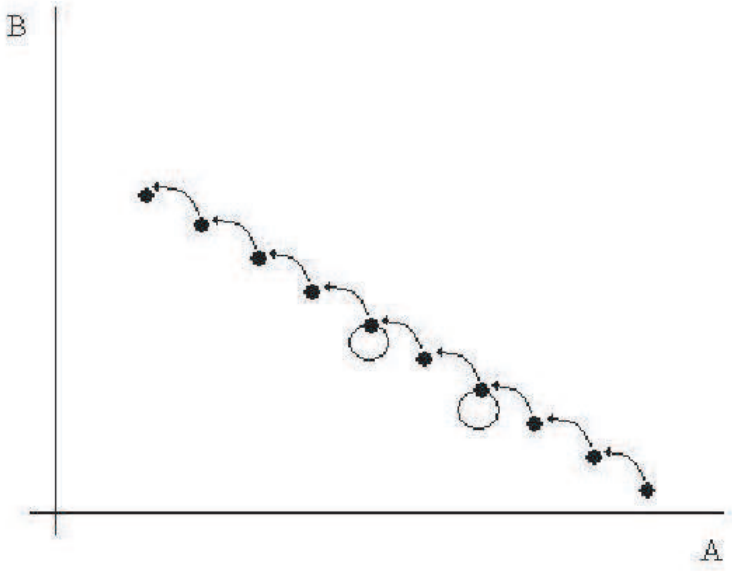

Figure 2.3: Random path of a unimolecular reaction

$$A \xrightarrow{k} B \tag{2.27}$$

Take a simple uni-molecular reaction (2.27). Eventually every molecule will end up at state B, where B is a large fixed number. Figure (2.3) resembles one particular path this chemical reaction can take. The circles represent no change in state. We can derive the following CME for this type of reaction,

$$\frac{\partial P(A,t)}{\partial t} = k(A+1)P(A+1,t) - kAP(A,t) \tag{2.28}$$

We have stated B is a fixed concentration, so that the total number of molecules at any time in the system is $K = A+B$, $B = K-A$. We now calculate the derivative of the first moment.

$$\frac{\partial \langle B \rangle}{\partial t} = \sum_A B \frac{\partial P}{\partial t} \tag{2.29}$$

$$\begin{aligned} \frac{\partial \langle B \rangle}{\partial t} &= \sum_A Bk(A+1)P(A+1,t) - kBAP(A,t) \\ &= \sum_A (K-A)k(A+1)P(A+1,t) - k(K-A)AP(A,t) \\ &= \sum_A kAP(A,t) \\ &= k\langle A \rangle \end{aligned} \tag{2.30}$$

Where $\langle\rangle$ correspond to taking the expectation. Notice how the equation is similar to the corresponding rate equation,

$$\frac{d[B]}{dt} = k[A] \tag{2.31}$$

where $[A]$ and $[B]$ are concentrations of their respective species. If we were to look at the steady state, $\frac{\partial \langle B \rangle}{\partial t} = 0$, we find the mean to be zero which is what you would expect because the reaction (2.27) is unidirectional and so, eventually all of species A would have been transformed to B. Next we consider the temporal evolution of the second moment.

$$\frac{\partial\langle B^2\rangle}{\partial t} = \sum_A B^2 \frac{\partial P}{\partial t} \tag{2.32}$$

$$\begin{aligned}\frac{\partial\langle B^2\rangle}{\partial t} &= \sum_A B^2 k(A+1)P(A+1,t) - kB^2AP(A,t) \\ &= \sum_A (K-(A-1))^2 kAP(A,t) - k(K-A)^2AP(A,t) \\ &= \sum_A (2K-2A+1)kAP(A,t) \\ &= (2K+1)k\langle A\rangle - 2k\langle A^2\rangle\end{aligned} \tag{2.33}$$

Now we have the rate of change of the second moment and are in a position to look at the time-varying variance.

$$\frac{\partial\langle\sigma_B^2\rangle}{\partial t} = \frac{\partial(\langle B^2\rangle - \langle B\rangle^2)}{\partial t} \tag{2.34}$$

$$= (2K+1)k\langle A\rangle - 2k(\langle A^2\rangle + \langle A\rangle^2) \tag{2.35}$$

As we approach the steady state, $\frac{\partial\langle\sigma_B^2\rangle}{\partial t} \to 0$ the second moment $\langle A^2\rangle \to 0$.

Simple bi-molecular reaction

We will now look at a bimolecular reaction,

$$P_1 + P_2 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} Z. \tag{2.36}$$

Figure (2.4) shows a random path a bimolecular reaction can take, again circles as before represent no change in state. We now derive the following Chemical Master Equation.

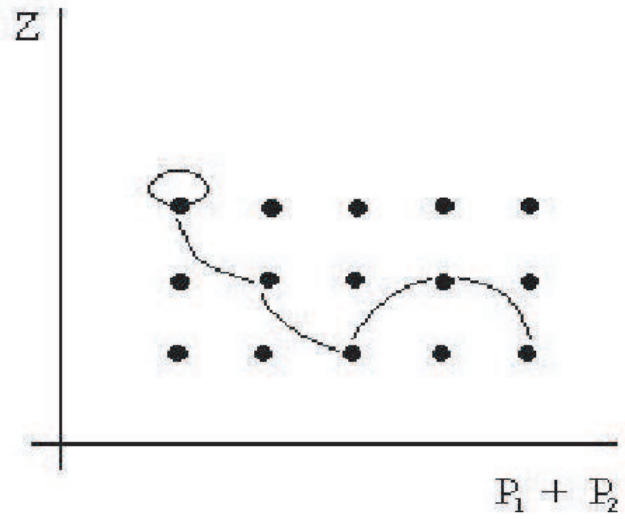

Figure 2.4: Random path of a bimolecular reaction

$$\frac{\partial P(P_1,P_2,Z,t)}{\partial t} = k_1(P_1+1)(P_2+1)(Z-1)P(P_1+1,P_2+1,Z-1,t) \tag{2.37}$$

$$-k_1P_1P_2P(P_1,P_2,Z,t)$$

$$+k_2(Z+1)P(P_1-1,P_2-1,Z+1,t) \tag{2.38}$$

$$- \quad k_2ZP(P_1,P_2,Z,t)$$

Next we calculate the derivative of the first moment,

$$\frac{\partial\langle Z\rangle}{\partial t} = k_1\langle P_1P_2\rangle - k_2\langle Z\rangle. \tag{2.39}$$

Which is again similar to its corresponding rate equation. At steady state the mean is found to be,

$$\langle Z\rangle = \frac{k_1}{k_2}\langle P_1P_2\rangle. \tag{2.40}$$

Thus at the steady state we have reached a limiting value. Let us now consider the derivative of the second moment to find an expression for the variance at steady state.

$$\begin{aligned} \frac{\partial\langle Z^2\rangle}{\partial t} &= 2k_1\langle P_1P_2Z\rangle + k_1\langle P_1P_2\rangle \\ &\quad -2k_2\langle Z^2\rangle + k_2\langle Z\rangle \end{aligned} \tag{2.41}$$

$$\begin{aligned} \frac{\partial\sigma_Z^2}{\partial t} &= 2k_1\langle P_1P_2Z\rangle + k_1\langle P_1P_2\rangle - 2k_2\langle Z^2\rangle \\ &\quad +k_2\langle Z\rangle - 2\langle Z\rangle k_1\langle P_1P_2\rangle + 2k_2\langle Z\rangle^2 \end{aligned} \tag{2.42}$$

$$\sigma_Z^2 = \frac{k_1}{k_2}\langle P_1P_2Z\rangle + \langle Z\rangle - \langle Z\rangle^2. \tag{2.43}$$

We now have an analytical expression for the variance, which we know is non-zero. Again we can use moment generating functions to solve the CME.

As the reactions become more complex the CME becomes very difficult to solve analytically since the P.D.E. which arises from using moment generating functions gets much harder to solve. Also imagine a network involving many reactions, it would be near impossible to look at the system analytically. But we can examine the system numerically.

2.3.2 Gillespie Algorithm: Discrete Markov Process

We can solve the Chemical Master Equations (2.28,2.37) numerically using a Stochastic Simulation Algorithm(SSA). One of the original algorithms developed was that of Daniel Gillespie [14, 15] back in the 1970's. He showed that it was possible to simulate the kind of chemical reactions discussed here using an effective and efficient stochastic algorithm. His algorithm has been used frequently to describe

biochemical network systems such as the ones discussed in Chapter 1 and Chapter 3. We will give a brief description and application here of the Gillespie algorithm, for a more detailed version of the method we refer you to his papers [14, 15, 16, 17].

Gillespies Algorithm is a very straightforward idea. The algorithm works with whole number of particles rather than concentrations. Already we can imagine this concept to be ideal for modelling reactions with small copy numbers such as transcription. The algorithm involves choosing two random numbers at each time step. The first predicts when the next reaction will occur and the second decides which reaction will occur.

If their exists are number of reactions, $\mu = 1, 2, 3, ...$ We suppose that for some very small time dt that only two species may interact at the same spot and time. The chance of three species interacting is very unlikely. Also we assume this time interval $(t, t+dt)$ is so small that only one reaction can take place. Let $a_\mu(t)dt$ be the reaction probability that at a time t a reaction μ will occur in the next time interval $(t, t + dt)$ in a volume V. Reaction probability, $a_\mu(t)$, is a product of the reaction rate c_μ and the number of possible reactions μ in volume V.

To calculate c_μ we need to transform the reaction rates, k_i, since we are now dealing with molecules and not concentrations, which are assumed to be in moles per litre. To do this we need to know the volume V and *Avogadro's number*, L, the number of molecules in a mole of substance. Introduce the parameter ω to be,

$$\omega = LV \tag{2.44}$$

as an external parameter to the system such that we can control the size of the system.

At time t we now know the state of the system, provided we know the reaction rates c_μ i.e. the number of molecules of each species is known. Therefore we know $a_\mu(t)$ for each reaction μ and thus $a_0(t)$, sum of all $a_\mu(t)$. We can now implement the algorithm.

- Firstly we need to find time the time τ after t when the next reaction will occur. This is done by drawing a random number from an exponential probability density function of rate a_0.

$$p(\tau) = a_0 exp(-a_0\tau) \tag{2.45}$$

- We now choose a random number from the uniform distribution between 0 and 1, to decide which reaction at occurs at time τ. Such that if the random number lies in the interval; $(0, a_1/a_0$ reaction 1 will occur, $(a_1/a_0, (a_1 + a_2)/a_0)$ reaction 2 occurs and so on.
- Finally we update the system after implementing the chosen reaction, which will alter the number of species in the system. We then go back to step 1 and continue the process for as long as we require.

We will now describe an idealized example of transcription and translation. Let $[D]$, $[X]$ represent the concentrations of the DNA promoter site and a transcription factor respectively. The following reversible reaction then describes the binding and dissociation of the transcription factor to the DNA promoter site.

$$D + X \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} DX \tag{2.46}$$

$[DX]$ represents the concentration of the transcription factor DNA complex. The next two reactions represent transcription of mRNA, with concentration $[M]$, and translation of protein, with concentration $[P]$.

$$DX + E \xrightarrow{k_2} DX + E + M \tag{2.47}$$

$$M \xrightarrow{k_3} M + P \tag{2.48}$$

Where $[E]$ represents the concentration of RNA polymerase. These reactions along with the next two, describing degradation of protein and mRNA, are considered irreversible.

$$M \xrightarrow{d_m} \emptyset \tag{2.49}$$

$$P \xrightarrow{d_p} \emptyset \tag{2.50}$$

Note we will let $\emptyset$ denote the degradation of a species here and throughout the rest of this thesis. The reactions just described can be expressed in terms of the following rate equations.

$$\frac{d[DX]}{dt} = k_1[D][X] - k_{-1}[DX] \tag{2.51}$$

$$\frac{d[M]}{dt} = k_2[DX][E] - d_m[M] \tag{2.52}$$

$$\frac{d[P]}{dt} = k_3[M] - d_p[P] \tag{2.53}$$

Let us assume that the copy numbers of the chemical species in the reactions described are very low and vary in a stochastic manner. In which case we can now implement the Gillespie Algorithm. We initially break down the original model into the following 6 reactions.

$$D + X \longrightarrow DX \tag{2.54}$$

$$DX \longrightarrow D + X \tag{2.55}$$

$$DX + E \longrightarrow DX + E + M \tag{2.56}$$

$$M \longrightarrow M + P \tag{2.57}$$

$$M \longrightarrow \emptyset \tag{2.58}$$

$$P \longrightarrow \emptyset \tag{2.59}$$

We now need to transform the rate constants. Let us consider the forward reaction (2.54), which is a second order reaction and by neglecting all the other reactions we obtain the following rate equation.

$$\frac{d[DX]}{dt} = k_1[D][X] \tag{2.60}$$

To convert into molecules we multiply through by ω, (2.3.2).

$$\frac{d(\omega[DX])}{dt} = k_1[D][X]\omega \tag{2.61}$$

$$\frac{dN_{DX}}{dt} = k_1[D]N_X \tag{2.62}$$

$$\frac{dN_{DX}}{dt} = \frac{k_1}{\omega}N_D N_X \tag{2.63}$$

Our first reaction rate c_1 is simply k_1/ω. For a first order reaction it is clear the reaction rate would simply be the rate constant k_i and for zeroth order reactions $k_i\omega$. We will now solve the deterministic idealized model giving a solution with which we can compare stochastic simulations with. By choosing arbitrary parameter values, k_i's, degradation rates, d_m and d_p, a suitable initial condition we can see the temporal

evolution of the number of protein molecules which eventually reaches a steady state in figure (2.5).

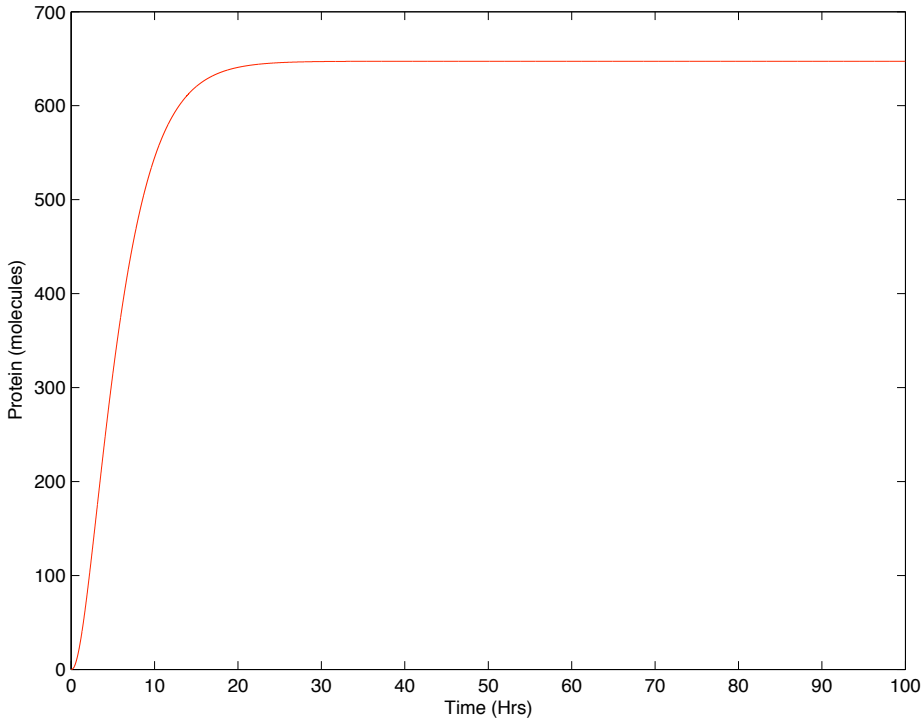

Figure 2.5: Deterministic Solution: $k_1 = 1.3294, k_{-1} = 0.8085, k_2 = 0.1445, k_3 = 0.2089, d_m = 0.3187, d_p = 0.3505$

Let us continue now with developing the discrete stochastic model. After transforming all the rate constants we can then write down the following set of reaction probabilities.

$$a_1(t) = \frac{k_1}{\omega} N_D N_X \tag{2.64}$$

$$a_2(t) = k_{-1} N_{DX} \tag{2.65}$$

$$a_3(t) = \frac{k_2}{\omega} N_{DX} N_E \tag{2.66}$$

$$a_4(t) = k_3 N_M \tag{2.67}$$

$$a_5(t) = d_m N_M \tag{2.68}$$

$$a_6(t) = d_p N_P \tag{2.69}$$

To decide when a reaction occurs we evaluate the following expression. Where T corresponds to the time of the next reaction, u_1 is a uniformly distributed number and a_0 the sum of all the reaction probabilities.

$$T = (1/a_0)\log(1/u_1) \tag{2.70}$$

We then pick a second uniformly distributed random number, u_2, which tells us which reaction takes place. We can now show the effect translational bursting, as described in Chapter 1, has on the protein distribution in figures(2.6,2.7). The red line corresponds to an average over a 100 realizations, blue line corresponds to a stochastic simulation with a fixed volume, V. The histograms show how translational bursting can effect the protein distribution. Translational bursting gives us a broader protein distribution which implies the biological noise in the system has increased but the average protein number remains unchanged. So our simple model is robust with respect to translational bursting.

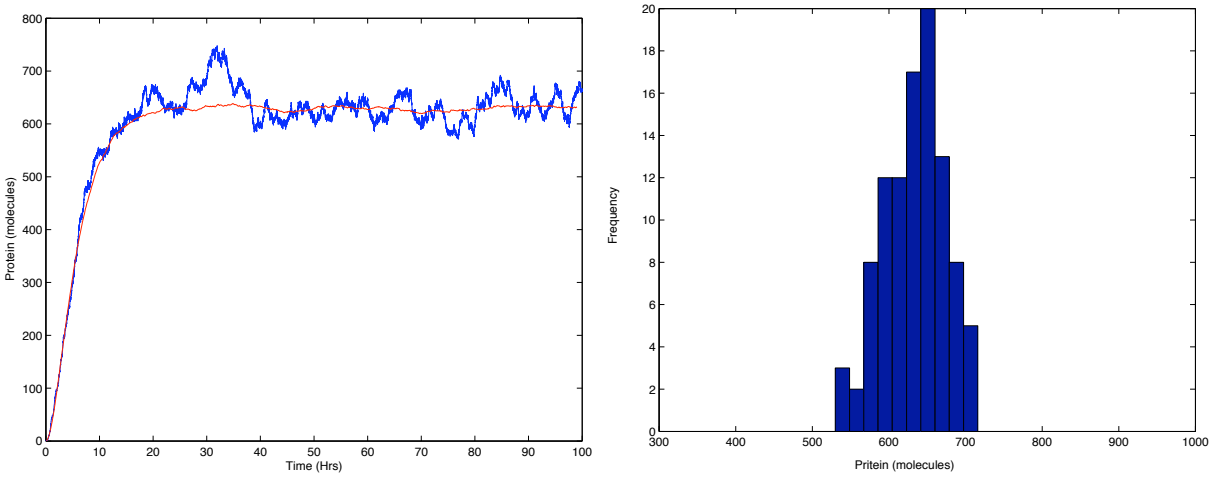

Figure 2.6: Gillespie Simulations: $k_1 = 1.3294, k_{-1} = 0.8085, k_2 = 0.1445, k_3 = 0.2089, d_m = 0.3187, d_p = 0.3505$

The Gillespie algorithm has one major advantage over conventional deterministic

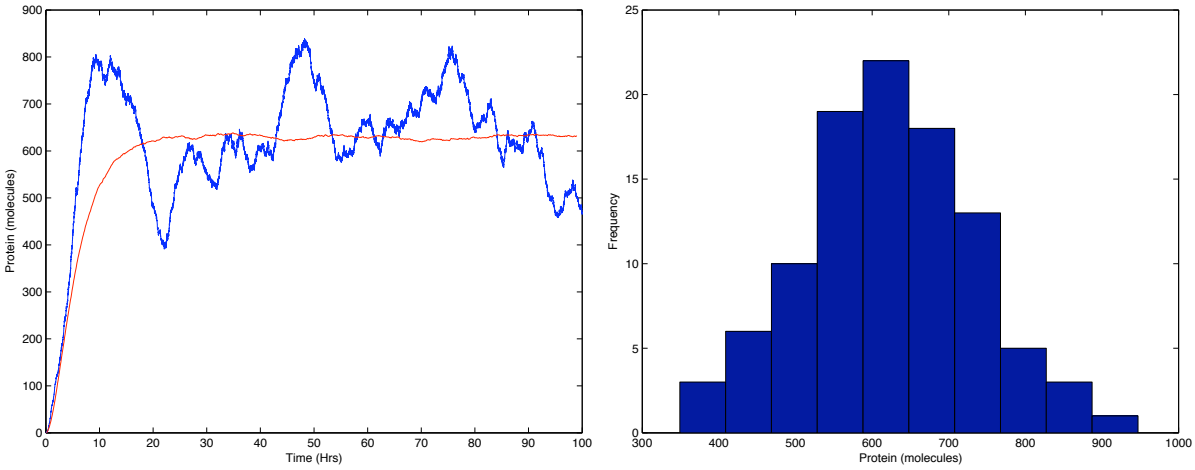

Figure 2.7: Gillespie Simulations: $k_1 = 1.3294, k_{-1} = 0.8085, k_2 = 0.01445, k_3 = 2.089, d_m = 0.3187, d_p = 0.3505$

modeling. We have seen that it is capable of modeling random reaction events at random times. The decisions to decide when and which reaction takes place are determined by rate constants and population size of each chemical species. By avoiding discretising time the Gillespie algorithm does not waste time doing simulations when no reaction occurs, it is able to treat time continuously.

The precision the algorithm can achieve can only work in chemical systems with certain properties. For example it is unable to represent each molecule in the system separately. So in a signalling pathway, like the ones we will be discussing, it cannot trace the path of the molecule so its fate is unknown. Also it cannot associate physical quantities to each molecule such as position and velocity, therefore it cannot simulate diffusion.

From a biological point of view. Protein molecules change frequently in a cell, they undergo phosphorylation to change states in signalling pathways. A protein complex may also contain lots of sites which could be modified independently. Which is then able to influence how the complex will participate in a reaction. So if a complex had 10 sites, 2^{10} unique states! Each of which react in a different way.

2.3.3 Gillespie to S.D.E.s: A continuous Markov Process

We have seen that the algorithm developed by Gillespie is a good candidate for modeling transcription and translation, when we have low copy numbers of molecules involved in these events. However what happens if the copy numbers of molecules are no longer small such that the waiting time to the next reaction becomes extremely small. Computationally using the Gillespie algorithm becomes very expensive and an alternative approach is required.

In the discrete model of chemical reactions, described in the previous section, the number of times that a reaction j occurs in a small time interval $(t, t+\tau)$ can be approximated by a Poisson random variable, $P_j(a_j(x)\tau)$. Where, $X(t) = x$, the number of molecules in state X at time t. If $a_j(x)\tau \gg 0$ then we can approximate the Poisson random variable with a Normal random variable,

$$P_j(a_j(x)\tau) \approx N(a_j(x)\tau, a_j(x)\tau) \tag{2.71}$$

$$= a_j(x)\tau + \sqrt{a_j(x)}\sqrt{\tau}N(0,1) \tag{2.72}$$

$$= a_j(x)\tau + \sqrt{a_j(x)}\Delta W_\tau \tag{2.73}$$

Thus the discrete Markov process is approximated by a continuous Markov process, described by the following SDE of Itô form.

$$dX = \sum_{j=1}^{M} v_j a_j(X)dt + \sum_{j=1}^{M} v_j \sqrt{a_j(X)}dW_j(t) \tag{2.74}$$

Where v_j is the stoichiometric vector and $W_j(t)$ are independent Wiener processes. Implementing this idea transforms the rate equations discussed in the previous section into the following system of SDEs.

$$dN_{DX} = \left(\frac{k_1 N_D N_X}{\omega} - k_{-1} N_{DX}\right) dt \tag{2.75}$$

$$+\sqrt{\left(\frac{k_1 N_D N_X}{\omega}\right)} dW_1 + \sqrt{(k_{-1} N_{DX})} dW_2 \tag{2.76}$$

$$dN_M = \left(\frac{k_2 N_{DX} N_E}{\omega} - d_m N_M\right) dt \tag{2.77}$$

$$+\sqrt{\left(\frac{k_2 N_D N_X}{\omega}\right)} dW_3 + \sqrt{(d_m N_m)} dW_4 \tag{2.78}$$

$$dN_P = (k_3 N_M - d_p N_P) dt \tag{2.79}$$

$$+\sqrt{(k_3 N_M)} dW_5 + \sqrt{(d_p N_p)} dW_6 \tag{2.80}$$

We will simulate the SDEs using the Euler-Maruyama method which will again show the effect of translational bursting in figures (2.8 , 2.9), using the same parameter values as in the Gillespie model described previously. In this case we notice no real difference between either the Gillespie approach or the SDE approach. However the SDE approach is computationally far more efficient when dealing with large number of reactions and so proves to be a good alternative to the Gillespie approach.

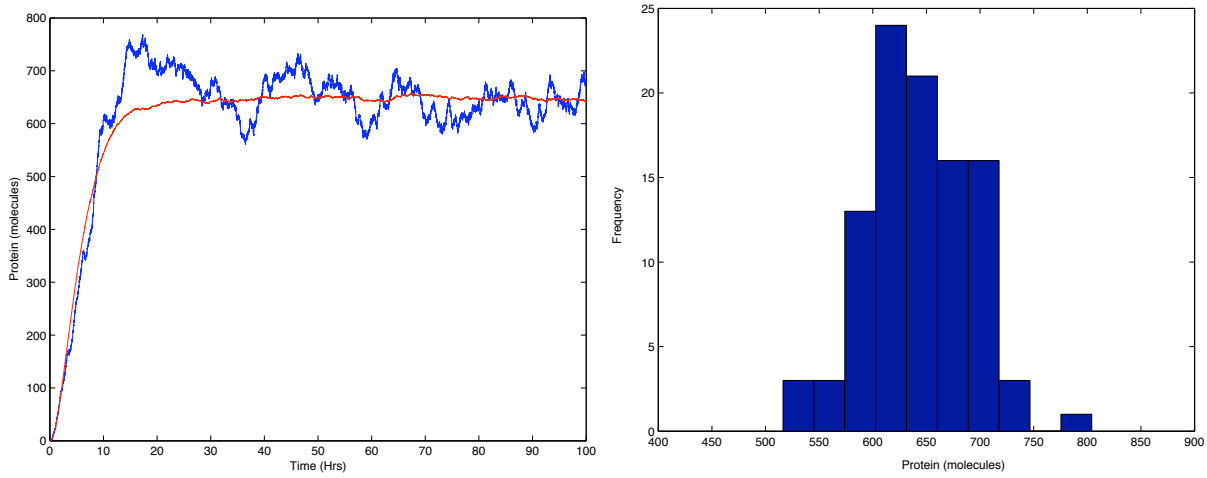

Figure 2.8: SDE Simulations: $k_1 = 1.3294, k_{-1} = 0.8085, k_2 = 0.1445, k_3 = 0.2089, d_m = 0.3187, d_p = 0.3505$

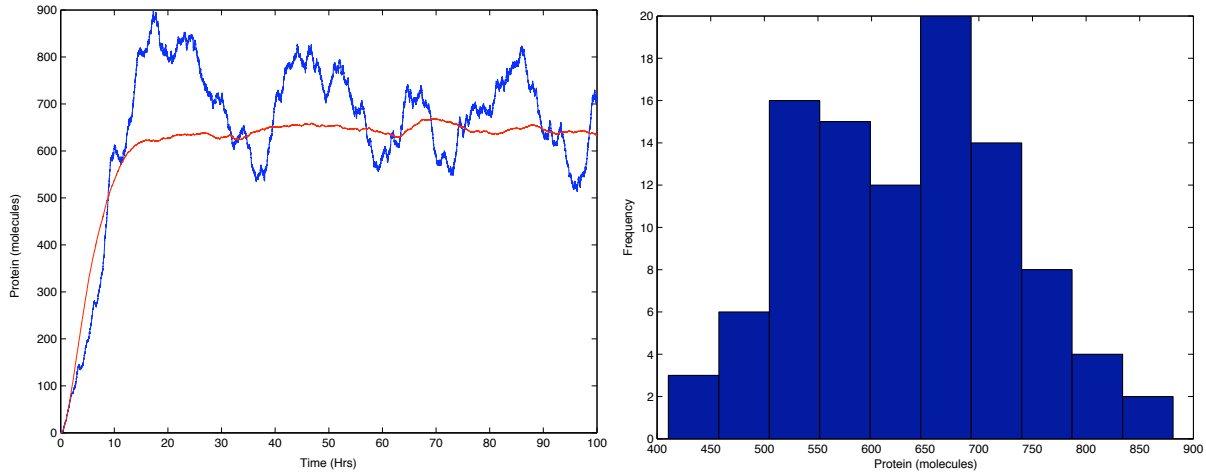

Figure 2.9: SDE Simulations: $k_1 = 1.3294, k_{-1} = 0.8085, k_2 = 0.01445, k_3 = 2.089, d_m = 0.3187, d_p = 0.3505$

2.3.4 Sensitivity to Parameter Variation

In this section we will consider white noise stochastic perturbations on the transcription and translation rate constants, k_3 and k_4 respectively. The question we will be looking to answer is whether the dynamical behavior of such a simple model, reactions (2.46,2.47,2.48,2.49,2.50), is robust to parameter variations one might see in such biological environments.

Let us consider stochastic perturbations of the two parameters, k_3 and k_4,

$$k_3 \longrightarrow k_3(1 + \alpha W_1(t)) \tag{2.81}$$

$$k_4 \longrightarrow k_4(1 + \alpha W_2(t)) \tag{2.82}$$

where $\alpha \in \mathbb{R}$ scales the stochastic fluctuations and $W(t)$ is a standard Wiener process. These transformations give rise to the following system of ODE and SDEs,

$$\frac{dN_{DX}}{dt} = \frac{k_1 N_D N_X}{\omega} - k_{-1} N_{DX} \tag{2.83}$$

$$dN_M = \left(\frac{k_2 N_{DX} N_E}{\omega} - d_m N_M\right) dt + \left(\frac{\alpha k_2 N_D N_X}{\omega}\right) dW_1 \tag{2.84}$$

$$dN_P = (k_3 N_M - d_p N_P)dt + (\alpha k_3 N_M) dW_2 \tag{2.85}$$

This system was solved using the Euler-Maruyama method using the same parameter values that were used in the previous models. The results can be seen in figures (2.10 , 2.11 , 2.12) where the blue line corresponds to a single realization and the red line is the average taken over 100 realizations. The results suggest that the averaged dynamical behavior of the model we have developed in this section is robust to stochastic white noise fluctuations. To be able to really decide if this idealized model of transcription and translation is truly robust to parameter variation we would need experimental data to provide us with a more realistic value for α.

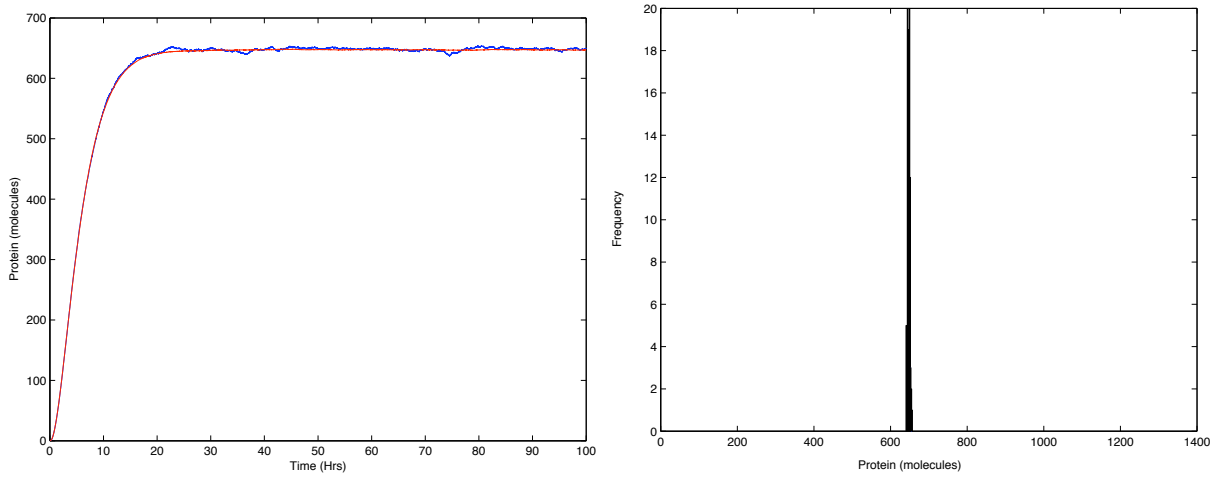

Figure 2.10: Solution of system with $\alpha = 0.01$: $k_1 = 1.3294, k_{-1} = 0.8085, k_2 = 0.1445, k_3 = 0.2089, d_m = 0.3187, d_p = 0.3505$

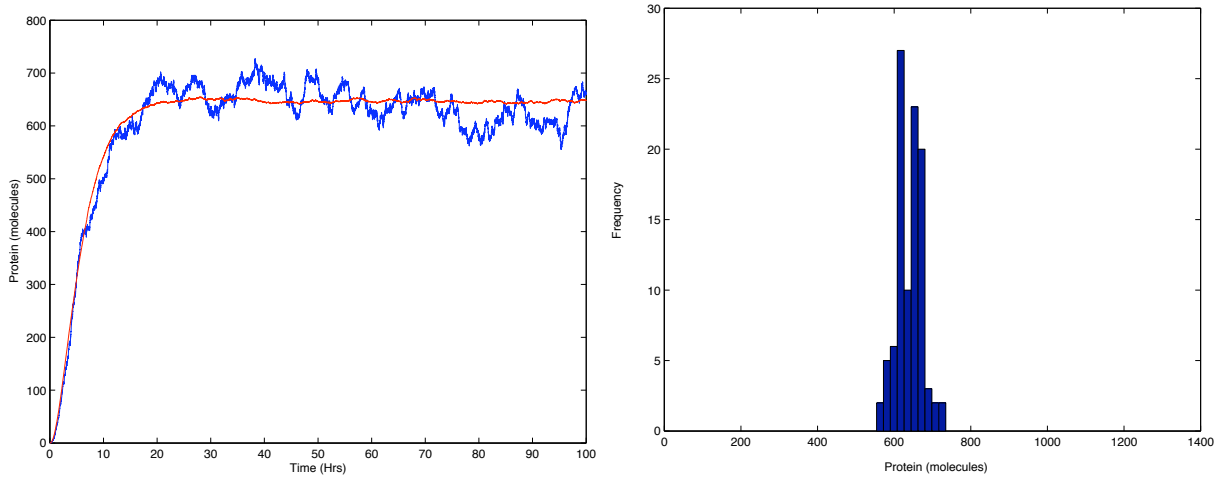

Figure 2.11: Solution of system with $\alpha = 0.1$: $k_1 = 1.3294, k_{-1} = 0.8085, k_2 = 0.1445, k_3 = 0.2089, d_m = 0.3187, d_p = 0.3505$

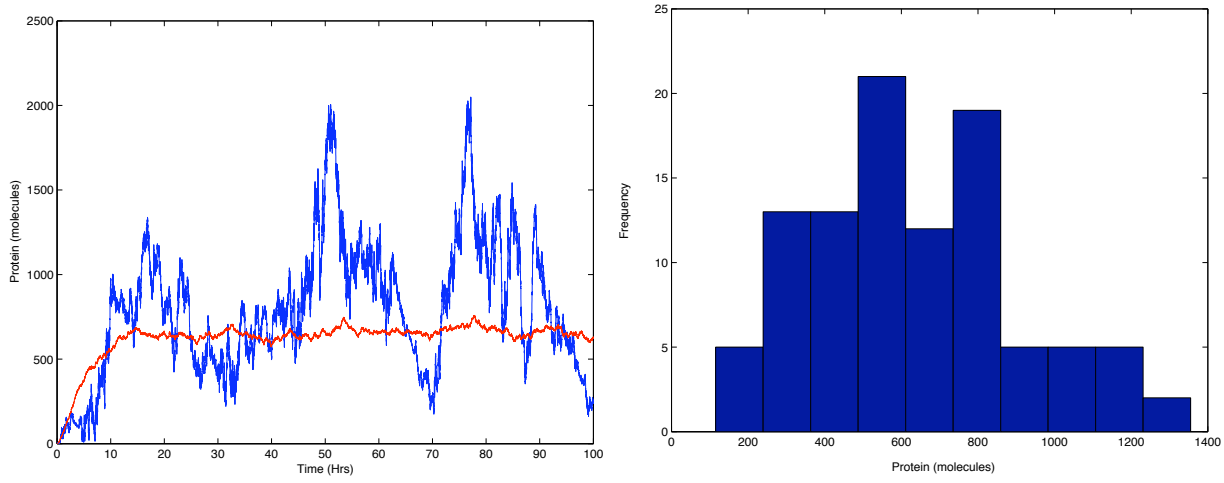

Figure 2.12: Solution of system with $\alpha = 1$:$k_1 = 1.3294, k_{-1} = 0.8085, k_2 = 0.1445, k_3 = 0.2089, d_m = 0.3187, d_p = 0.3505$

2.4 Summary

In this chapter we have introduced three stochastic methods that could be used to test the robustness of gene regulatory networks. We have the Gillespie algorithm which can be used to model situations where the copy numbers of species is very low and so provide us with an option to modelling intrinsic fluctuations. As the reactions become more frequent in a small time interval we are able to switch to an SDE approach which does not loose the effect of intrinsic fluctuations but is computationally a lot more efficient. Also the SDE approach discretises time which will be useful later on when we begin to analyze our results more rigourously. Finally we introduced a model which could be used to examine the effects white noise perturbations can have on key parameters. In the following Chapters we will use these mathematical tools to examine the robustness of the circadian clock in *Arabidopsis thaliana*.

CHAPTER III

Arabidopsis Circadian Clock

Circadian clocks are seen throughout nature and are key to life on earth as we see it today. Dunlap [11] discusses the history of circadian clocks and links between clocks in different species. He proposes that circadian oscillators are all composed of both inhibitive and promotive reactions configured into some sort of network. So far networks have been presumed to be based on some sort of transcriptional regulation. One could say that there are many periodic mechanisms in organisms that display 24hr periods under the influence of external stimulus; circadian clocks have additional constraints:

- The period must remain relatively constant over a wide range of temperatures [40].
- The clock must be able to function in the presence of intrinsic noise.
- The circadian rhythm must be able to withstand global changes in transcription and translation rates caused by, for example, variations in nutrition.

These considerations suggest that models of circadian clocks need to be robust. The second point could be tested by developing a discrete stochastic model and examining its behavior. The final point can be studied by developing a parameter variation model, similar to one discussed in the previous Chapter.

So what role does mathematics play in this large and unpredictable biological world? One answer maybe found in the work of Locke et al. [31, 32, 24]. They produced a number of bio-mathematical papers establishing links between experiments and computer simulations of the circadian clock in *Arabadopsis thaliana*, a much-studied organism regarded as a model for the flowering plants.

In their first paper [31] Locke et al. produced a deterministic model that used only 3 genes governed by Michaelis-Menten Kinetics and sought to reproduce certain key qualitative features of the experimental observations. This original model did not match the *in vivo* experiments very well, so they returned to the theoretical model and proposed changes, mainly the inclusion of a second loop [32]. The mammalian and Drosophila circadian clocks are composed of two interlocking feedback loops for transcription and translation. One may therefore draw the conclusion that the second loop in the Arabidopsis clock is not much of a surprise. However they also found that the putative extra gene, initially called Y, which was involved in the second loop must have two peaks of expression in each day: a burst at dawn and then a much broader burst again at dusk. By then going back to the experiments, they were able to identify a possible candidate for this mysterious gene, *GIGANTEA*. Their third paper [24] explores the effect temperature has on all these key genes. They begin to examine how circadian clocks are able to maintain their 24 hour rhythm across a wide range of temperatures despite the strong influence of temperature on chemical reactions. These papers by Locke et al. show that the combination of raw experimental data from biologists and the more theoretical mathematical approach yields a deeper understanding of the circadian clock in the *Arabidopsis* plant.

As none of the parameters in the mathematical model are known, the mathematical problem under examination here is a kind of inverse problem. The aim is to

choose network architectures and parameters to match the experimental findings. In Locke et al.'s first model [31], a set of parameters could not be found to match their experimental findings, leading to their exploration of the possibility of there being more genes in play. The moral here is that the underlying mathematics improved experimental design, which in turn helped give a greater understanding of the networks involved.

Currently understood circadian oscillators mainly rely on a negative feedback loop [11]. Deterministic models for circadian rhythms [31, 32, 30] develop oscillations of the limit cycle type, which are entrained by light/dark cycles. The papers produced by Gonze et al. [19, 22, 20, 21, 23] consider the effects of molecular fluctuations arising due to small number of molecules involved in transcriptional regulation. Gonze et al. developed both deterministic and stochastic models for the *Drosophila* and *Neurospora* circadian clocks and found that circadian rhythms persisted even with very low copy numbers. However, they found that if they considered sufficiently low copy numbers the circadian oscillation was overpowered by the noise so that the periodic behaviour was destroyed.

In this chapter we will look at the original *Arabidopsis* clock model and examine the effect of introducing biological noise in the same vein as Gonze et al. The fluctuations described in Chapter 1 can have detrimental effects on cell signalling and may corrupt circadian clocks, so models describing such biological systems need to be robust [3, 6]. Modeling the network using Gillespie type kinetics [2], it will be possible to see the effect molecular noise on circadian rhythm. It will be seen how the circadian oscillations behave when the maximum numbers of mRNA and protein molecules are of the order of only a few tens or hundreds. We finally discuss the effects of white noise perturbations to the transcription and translation parameters.

In perturbing these parameters, the robustness of the original deterministic model may be ascertained.

3.1 Deterministic Model

Here we review briefly the feedback models of Locke et.al. [31, 32].

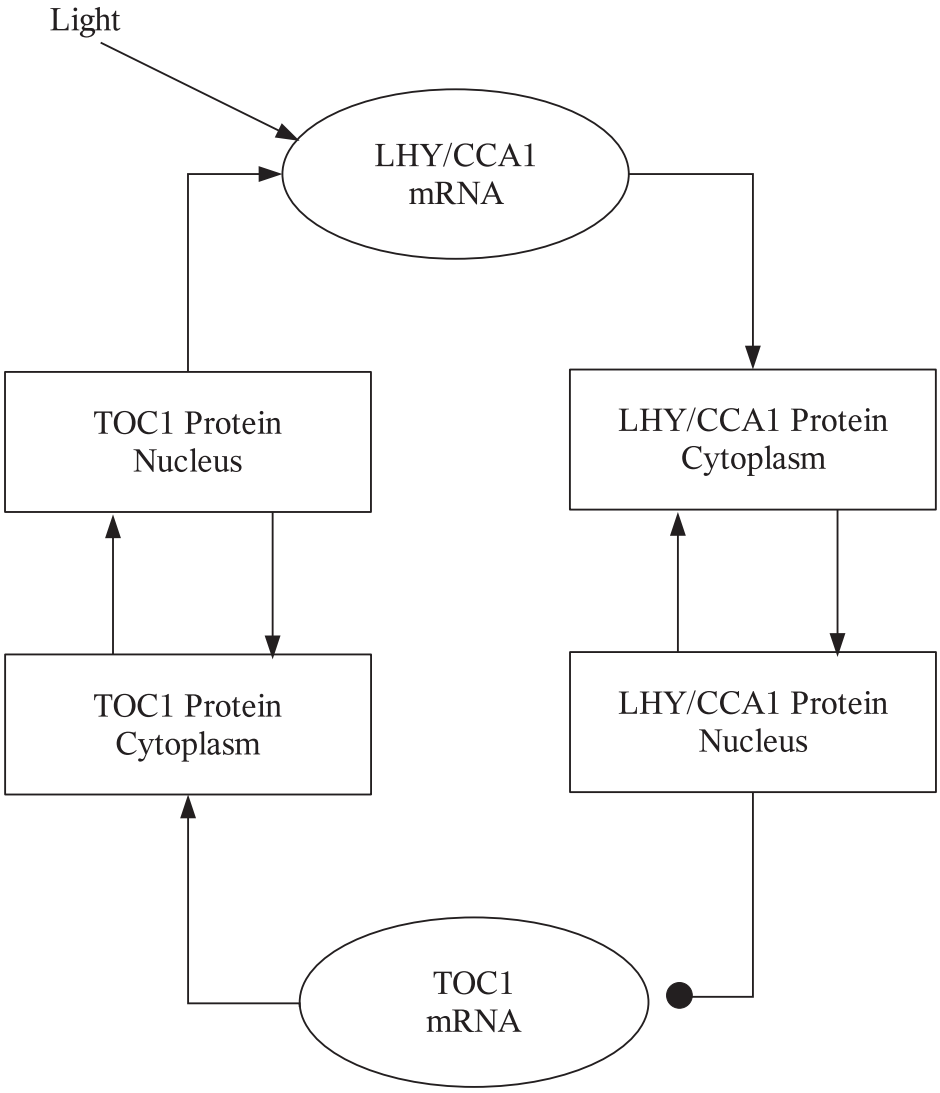

Figure 3.1: Arabidopsis Clock Network

Figure 3.1 shows the network used to model the *Arabidopsis* clock. mRNA's and proteins are represented by ovals and boxes respectively. Arrows represent a positive effect, the dot an inhibitory effect.

Light is known to activate the expression of LHY/CCA1 genes, so either light or TOC1 protein are required to activate the transcription of LHY/CCA1 mRNA which then travels into the cytoplasm, where it is translated into protein. LHY/CCA1 protein then travels into the nucleus and inhibits the production of TOC1 mRNA: establishing a negative feedback loop. If we were to take a day which consisted of 12 hours light and 12 hours of darkness we would expect periodic expressions of both genes. A light/dark pattern is considered to try and replicate a plant in its natural environment. LHY/CCA1 production would peak as light became available, inhibiting and thereby reducing the levels of TOC1. As TOC1 levels decrease, levels of LHY/CCA1 would decrease. It maybe inferred then that LHY/CCA1 concentrations should peak at dawn and then die away as the day wore on. Consequently we would expect TOC1 levels to have a much broader peak.

A large amount of experimental data is available, providing us with approximate values for the phase and period of the mRNA oscillations. Plants were subjected to constant light then constant darkness (12 hours of each). Concentrations of the protein are obtained by attaching a Luciferase reporter gene to the genes in question. Even given the abundance of experimental data, the ubiquity of noise makes parameter fitting a significant problem. This is a common problem when modeling many biological systems. In response, Locke et. al. developed *score functions*, an approach which considers qualitative agreements between the model and the experimental data, see Chapter 4.

Locke et. al. developed a model based on a system of O.D.E.'s of the form,

$$\frac{d[x]}{dt} = synthesis \pm transport - decay. \tag{3.1}$$

Synthesis, represents transcription and translation, the former modeled using Hill

functions and the latter just a simple linear kinetics. *Transport* describes the transport of proteins and mRNA between, nucleus and cytoplasm. *Decay*, is simply the decay of proteins and mRNA modeled using Michealis-Menten kinetics.

The following equations describe the clock model in figure 3.1 developed by Locke et. al [31].

$$
\begin{aligned}
f_1(T_n, L_m) &= \frac{dL_m}{dt} = Light(t) + \frac{n_1 T_n^a}{g_1^a + T_n^a} - \frac{m_1 L_m}{k_1 + L_m} \\
f_2(L_c, L_m, L_n) &= \frac{dL_c}{dt} = p_1 L_m - r_1 L_c + r_2 L_n - \frac{m_2 L_c}{k_2 + L_c} \\
f_3(L_n, L_c) &= \frac{dL_n}{dt} = r_1 L_c - r_2 L_n - \frac{m_3 L_n}{k_3 + L_n} \\
f_4(T_m, L_n) &= \frac{dT_m}{dt} = \frac{n_2 g_2^b}{g_2^b + L_n^b} - \frac{m_4 T_m}{k_4 + T_m} \\
f_5(T_m, T_c, T_n) &= \frac{dT_c}{dt} = p_2 T_m - r_3 T_c + r_4 T_n - \frac{m_5 T_c}{k_5 + T_c} \\
f_6(T_n, T_c) &= \frac{dT_n}{dt} = r_3 T_c - r_4 T_n - \frac{m_6 T_n}{k_6 + T_n}
\end{aligned} \tag{3.2}
$$

The labels L and T correspond to LHY/CCA1 and TOC respectively and the subscripts m, n and c correspond to mRNA, nuclear protein and cytoplasmic protein concentrations respectively. Here the rate constants n_k, g_k parameterize transcription; m_k and k_k, degradation; p_k translation and r_k transport of mRNA's and proteins. The Hill coefficient b was set to 2 following biological evidence suggesting that LHY and CCA1 proteins bind as a dimer to TOC1s promoter. No experimental evidence exists as yet for the Hill coefficient a so it was set to 1. $Light(t)$ represents the input of light into the system. A light sensitive protein, P_n, is known to interact with the LHY gene promoter such that,

$$
Light(t) = q_1 P_n \Theta_{light}, \tag{3.3}
$$

where $\Theta_{light} = 1$ when light is present and 0 when its not. P_n satisfies the following equation,

$$f_7(P_n, \Theta_{light}) = \frac{dP_n}{dt} = (1 - \Theta_{light})p_3 - \frac{m_7 P_n}{k_7 + P_n} - q_2 \Theta_{light} P_n, \tag{3.4}$$

Here the four parameter values are chosen to express light protein in a similar way to the experiments, to give an acute, transient activation response for the expression of LHY/CCA1. Overall there are 29 free parameters, for which only 6 have been accounted for by the light transduction system. The remaining 23 were found after an extensive parameter search, details of which are found in [31].

Results from Locke et al.s first model encapsulated certain aspects of the experimental data quite well, in particular periodicity and profile of the LHY peaks seemed correct. However the model seemed to anticipate dawn, more so than the experiment. Also the TOC1 peak in the model does not agree with the experiment, arriving around two hours later.

Figures (3.2) and (3.3) are phase plots showing that the oscillations arising from the model are of a periodic nature and correspond to limit-cycle behavior in the light/dark case and a spiral fixed point in dark/dark case (black lines). In these plots the fixed time step used in the integration was 0.005 hours and the phase space was divided up into 100 equally sized compartments. The plots show how long a trajectory spends in that compartment of phase-space. Note that the two figures plotted here will be used as a comparison for similar plots obtained from the stochastic simulations.

The data for figure (3.2) was taken from simulations in which the day was divided into alternating periods of light and darkness, each 12 hours long. The plots involve

data from 100 hours of simulated time, collected after integration through an initial 200 hours (to allow transients associated with entrainment to the diurnal cycle to relax).

After 300 hours of simulated time, the pattern of illumination changed to continuous darkness for a further 300 hours of simulated time: the data displayed in figure (3.3) comes from this "dark/dark" phase of the simulation. We will be comparing these phase space figures to subsequent ones obtained from the stochastic models. Note that the parameter values used throughout this chapter were the same set that gave Locke et al. the best qualitative fit to experiments.

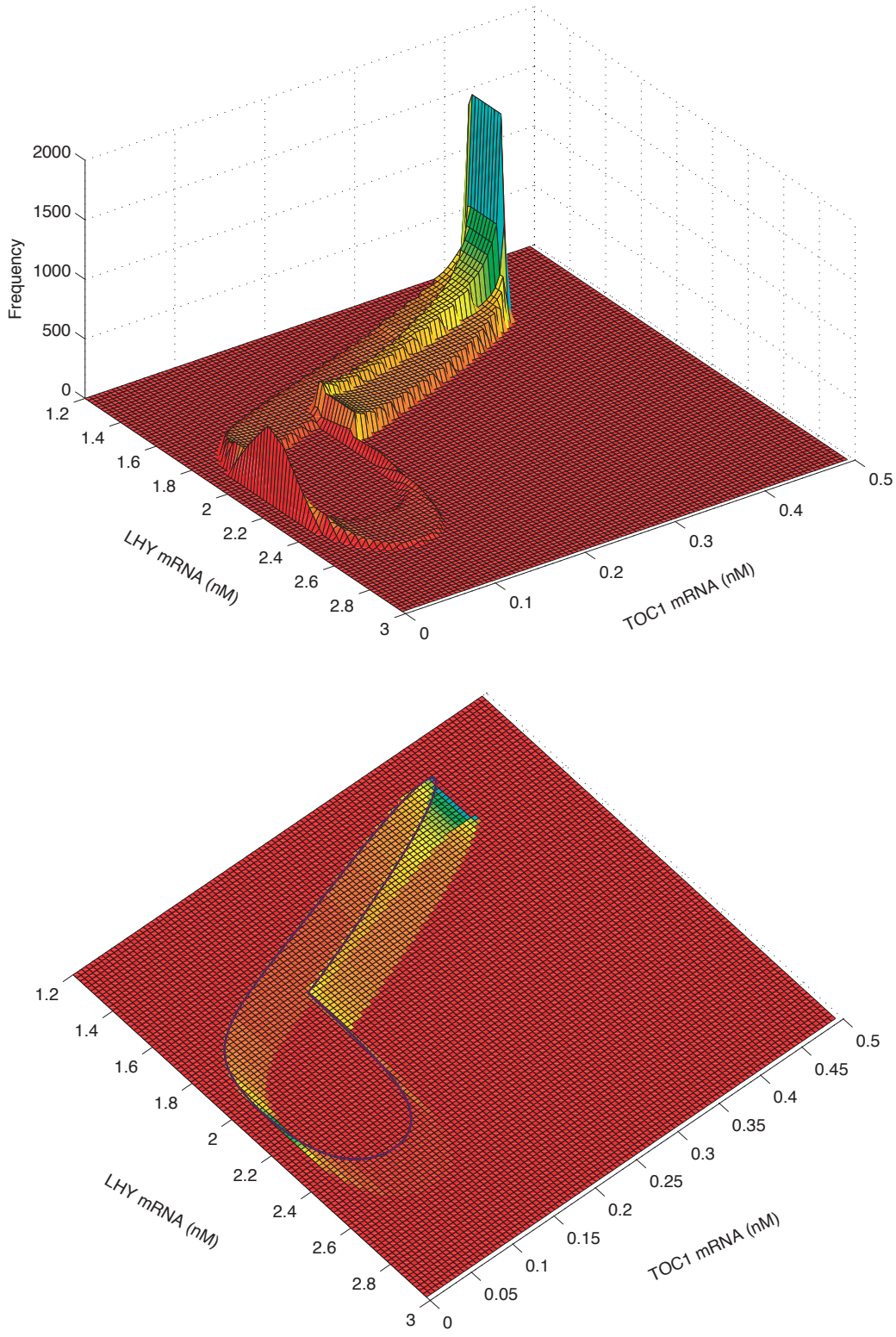

Figure 3.2: ODE simulation 3.2: 3-D and 2-D phase portraits for the 12hr light/dark cycle over 100 hour period.

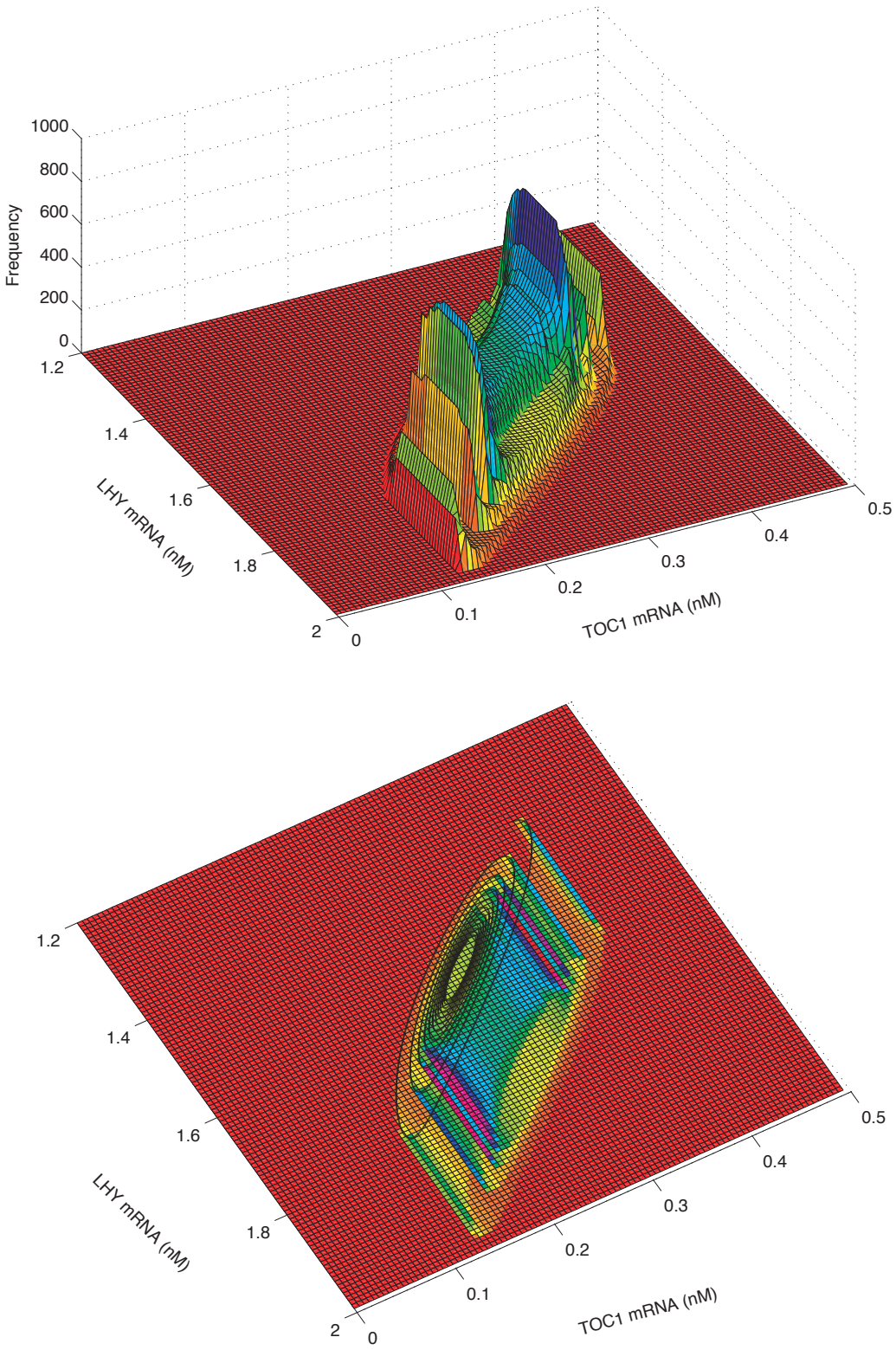

Figure 3.3: ODE simulation 3.2: 3-D and 2-D phase portraits for the continuous darkness cycle over a 300 hour period.

3.2 Discrete Stochastic Model of the Arabidopsis Clock

To see the effect of molecular noise on the *Arabidopsis* clock, we constructed a discrete stochastic model and used the Gillespie algorithm to simulate the model numerically. We transformed the deterministic model into Gillespie type steps which include non-linear kinetic functions in their probabilities. We broke down the original model into the following elementary reactions:

$$
\begin{aligned}
Dark &\longrightarrow P_n \\
P_n &\longrightarrow L_m + P_n \\
T_n &\longrightarrow L_m + T_n \\
L_m &\longrightarrow L_c + L_m \\
L_c &\longrightarrow L_n \\
L_n &\longrightarrow L_c \\
* &\longrightarrow T_m \\
T_m &\longrightarrow T_c + T_m \\
T_c &\longrightarrow T_n \\
T_n &\longrightarrow T_c \\
P_n &\longrightarrow \emptyset \\
P_n &\longrightarrow \emptyset \\
L_m &\longrightarrow \emptyset \\
L_n &\longrightarrow \emptyset \\
L_c &\longrightarrow \emptyset \\
T_m &\longrightarrow \emptyset \\
T_n &\longrightarrow \emptyset \\
T_c &\longrightarrow \emptyset
\end{aligned}
$$

Note that light sensitive protein (P_n) is only produced in darkness and that there are two degradation terms, representing a general steady-state degradation as well as a stronger, light-mediated degradation process. Also note that P_i, T_i and L_i are now numbers of molecules rather than concentrations.

We introduce the parameter ω to transform the concentrations apperaring in 3.2 into whole numbers of molecules. Having thus controlled system size, the following reaction probabilities corresponding to the aforementioned reactions are obtained:

$$
\begin{aligned}
a_1(\Theta_{light}) &= p_3(1-\Theta_{light})\omega \\
a_2(P_n,\Theta_{light}) &= q_1 P_n \Theta_{light} \\
a_3(T_n) &= \frac{n_1\omega T_n}{g_1\omega + T_n} \\
a_4(L_m) &= p_1 L_m \\
a_5(L_c) &= r_1 L_c \\
a_6(L_n) &= r_2 L_n \\
a_7(L_n) &= \frac{n_2\omega^3 g_2^2}{g_2^2\omega^2 + L_n^2} \\
a_8(T_m) &= p_2 T_m \\
a_9(T_c) &= r_3 T_c \\
a_{10}(T_n) &= r_4 T_n \\
a_{11}(P_n) &= \frac{m_7\omega P_n}{k_7\omega + P_n} \\
a_{12}(\Theta_{light},P_n) &= q_2 \Theta_{light} P_n \\
a_{13}(L_m) &= \frac{m_1\omega L_m}{k_1\omega + L_m} \\
a_{14}(L_c) &= \frac{m_2\omega L_c}{k_2\omega + L_c} \\
a_{15}(L_n) &= \frac{m_3\omega L_n}{k_3\omega + L_n} \\
a_{16}(T_m) &= \frac{m_4\omega T_m}{k_4\omega + T_m} \\
a_{17}(T_c) &= \frac{m_5\omega T_c}{k_5\omega + T_c} \\
a_{18}(T_n) &= \frac{m_6\omega T_n}{k_6\omega + T_n}
\end{aligned}
\tag{3.5}
$$

We chose not to decompose the Hill functions into more elementary steps due to the findings of Gonze et al. [20], who have shown that similar results can be obtained without breaking down the reactions further.

The following equation is the Chemical Master Equation for the *Arabidopsis* clock, where $P \equiv P(Light_{np}, LHY_m, LHY_{cp}, LHY_{np}, TOC1_m, TOC1_{cp}, TOC1_{np})$.

$$
\begin{aligned}
\frac{\partial P}{\partial t} &= a_1(\Theta_{light})P(...,LHY_{np}-1,...) - a_1\Theta_{light}P(...,LHY_{np},...) \\
&+ a_2(Light_{np},\Theta_{light})P(...,LHY_m-1,...) - a_2(Light_{np},\Theta_{light})P(...,LHY_m,...) \\
&+ a_3(TOC1_{np})P(...,LHY_m-1,...) - a_3(TOC1_{np})P(...,LHY_m,...) \\
&+ a_4(LHY_m)P(...,LHY_{cp},...) - a_4(LHY_m)P(...,LHY_{cp},...) \\
&+ a_5(LHY_{cp}+1)P(...,LHY_{cp}-1,LHY_{np}+1,...) - a_5(LHY_{cp})P(...,LHY_{cp},LHY_{np},...) \\
&+ a_6(LHY_{np}+1)P(...,LHY_{cp}+1,LHY_{np}-1,...) - a_6(LHY_{np})P(...,LHY_{cp},LHY_{np},...) \\
&+ a_7(LHY_{np})P(...,TOC1_m-1,...) - a_7(LHY_{np})P(...,TOC1_m,...) \\
&+ a_8(TOC1_m)P(...,TOC1_{cp}-1,...) - a_8(TOC1_m)P(...,TOC1_{cp},...) \\
&+ a_9(TOC1_{cp}+1)P(...,TOC1_{cp}-1,TOC1_{np}+1,...) \\
&- a_9(TOC1_{cp})P(...,TOC1_{cp},TOC1_{np},...) \\
&+ a_{10}(TOC1_{np}+1)P(...,TOC1_{cp}+1,TOC1_{np}-1,...) \\
&- a_{10}(TOC1_{np})P(...,TOC1_{cp},TOC1_{np},...) \\
&+ a_{11}(Light_{np}+1)P(...,Light_{np}+1,...) - a_{11}(Light_{np})P(...,Light_{np},...) \\
&+ a_{12}(\Theta_{light},Light_{np}+1)P(...,Light_{np}+1,...) - a_{12}(\Theta_{light},Light_{np})P(...,Light_{np},...) \\
&+ a_{13}(LHY_m+1)P(...,LHY_m+1,...) - a_{13}(LHY_m)P(...,LHY_m,...) \\
&+ a_{14}(LHY_{cp}+1)P(...,LHY_{cp}+1,...) - a_{14}(LHY_{cp})P(...,LHY_{cp},...) \\
&+ a_{15}(LHY_{np}+1)P(...,LHY_{np}+1,...) - a_{15}(LHY_{np})P(...,LHY_{np},...) \\
&+ a_{16}(TOC1_m+1)P(...,TOC1_m+1,...) - a_{16}(TOC1_m)P(...,TOC1_m,...) \\
&+ a_{17}(TOC1_{cp}+1)P(...,TOC1_{cp}+1,...) - a_{17}(TOC1_{cp})P(...,TOC1_{cp},...) \\
&+ a_{18}(TOC1_{np}+1)P(...,TOC1_{np}+1,...) - a_{18}(TOC1_{np})P(...,TOC1_{np},...)
\end{aligned}
$$

To solve such a system analytically would indeed be a very difficult task. However

one can examine the rate of change of the first and second moments as discussed in Chapter 2. This calculation was performed and did not provide us with any further insight into the model. We will now continue to look at the model from a numerical point of view.

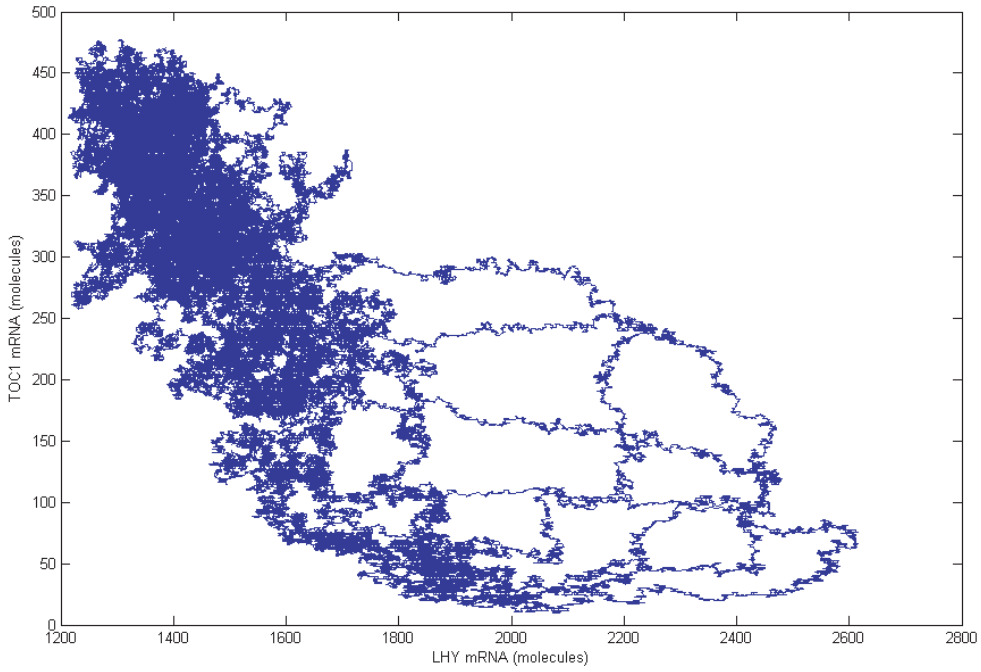

Figure 3.4: Gillespie simulation: Phase portrait, $\omega = 1000$, 12 hours light/dark

Recall that in the deterministic model, circadian oscillations evolved towards a limit cycle shown in figures (3.2 and 3.3). Let us now consider the dynamic behavior of the discrete stochastic model in figure (3.5), where $\omega = 1000$. This value of ω is a realistic biological value obtained for an average plant cell. For this value of ω the numbers of molecules for LHY mRNA, LHY cytoplasmic protein, TOC1 mRNA and TOC1 cytoplasmic protein range from 100 to 2800. Figure (3.5) shows that the circadian rhythm persists, as one would expect from the deterministic model. For this value of ω, phase trajectories from realizations of the discrete process give rise to

a density concentrated near the deterministic limit cycle: see figure 3.4. This further indicates the robustness of the circadian oscillations in the face of molecular noise when the maximum number of molecules is no more than a few thousand.

The histogram of periods taken over 100 consecutive cycles in figure (3.6) is very broad, but has a maximum close to 24 hours for both LHY mRNA and TOC1 mRNA. This is in good agreement with the deterministic case, implying the period is robust to molecular noise.

Let us examine the frequency of passage through different areas of the phase space, (see figure (3.7)). If we compare this with the corresponding deterministic version (figure (3.2)) we can see that the area through which the trajectories pass is increased as is the frequency of passing through certain areas. The most frequently visited regions seem to be those where LHY mRNA levels are low and TOC1 mRNA is at its peak. The corresponding time of day is within a few hours on either side of dusk, which agrees well with the deterministic case.

After the circadian rhythm is entrained to a 12 hour light/dark cycle we subjected the model to constant darkness for 300 hours. Figure (3.8) shows the temporal evolution of LHY mRNA, LHY cytoplasmic protein, TOC1 mRNA and TOC1 cytoplasmic protein numbers. The score functions used to fix the parameters for the deterministic model (see Chapter 4 for details) require that the oscillations decay under conditions of constant darkness, but that they remain approximately periodic with a period of around 25 hours. This does not seem to be the case with the stochastic model. We notice the amplitude of the oscillations does not dampen over time. In fact it behaves in a rather random manner. This is rather worrying from a modelling point of view. If we follow a state on the trajectory of a particular stochastic spiral fixed point, the noise in the system seems to be large enough to perturb the state onto an outlying

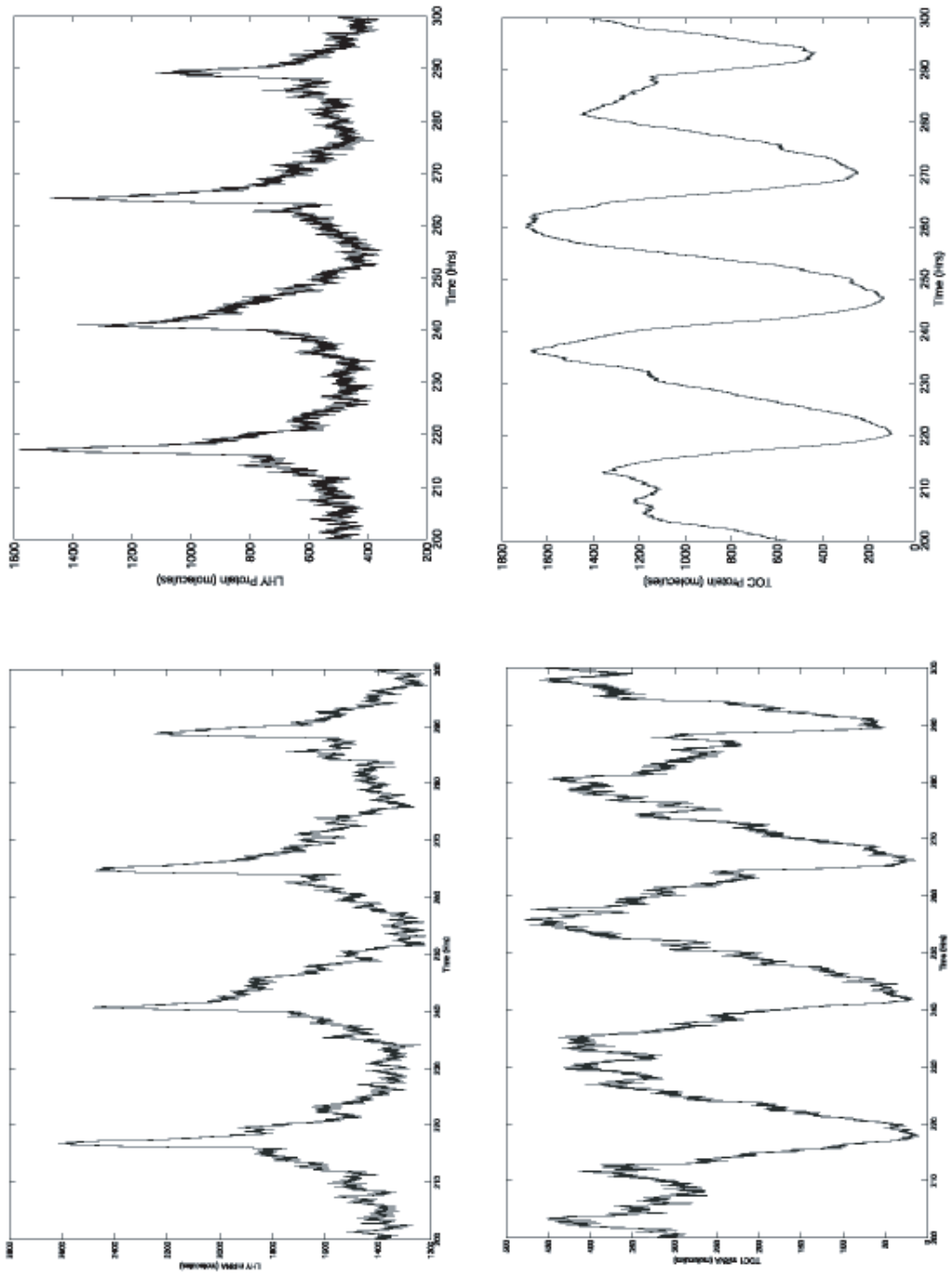

Figure 3.5: Gillespie simulation: $\omega = 1000$, 12 hours light/dark

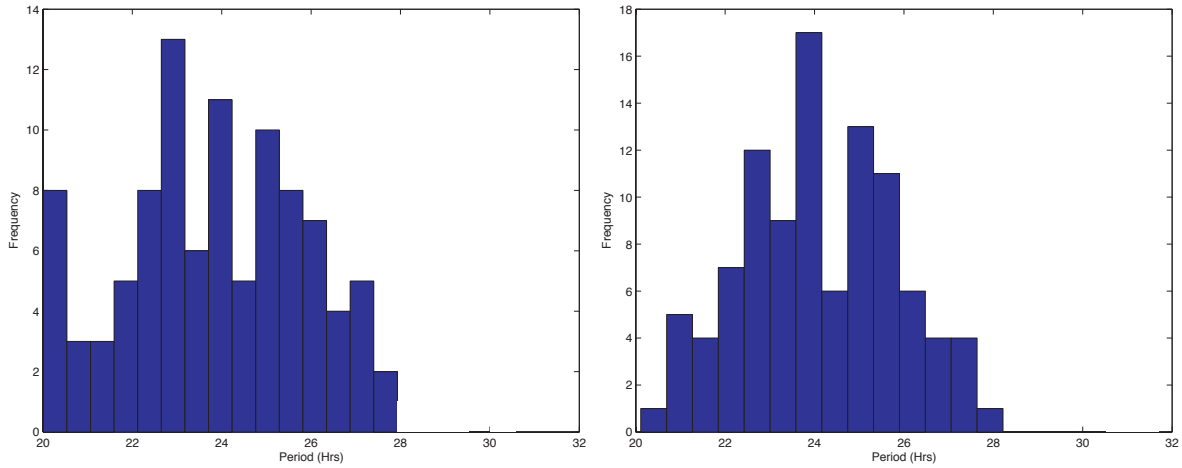

Figure 3.6: Gillespie simulation: $\omega = 1000$, 12 hours light/dark

part of the spiral. This would explain why the amplitude of the oscillations does not dampen and the change in phase. This suggests that the noise in the system is too large.

Comparison of the stochastic phase space plot, figure (3.9), with the deterministic one (figure (3.3)) tells a rather peculiar story. In the stochastic case, no clear areas of phase space stand out as being as frequently visited as in the deterministic case. This further corroborates our previous statement that the noise seems to be too large. We expect by increasing ω, we will be able to better mimic the deterministic dynamics. Increasing ω further in the discrete model becomes computationally very expensive. This will be further investigated when we consider the continuous stochastic model in the next section.

Instead, let us consider the limit of very strong stochasticity and decrease the system size further, thus decreasing the number of molecules involved in the reactions. In figure (3.10) $\omega = 100$ and we are still able to make out a circadian rhythm for numbers of molecules in the few hundreds and lower. Decrease ω further to 10 however, and the entrainment has been destroyed by molecular noise.

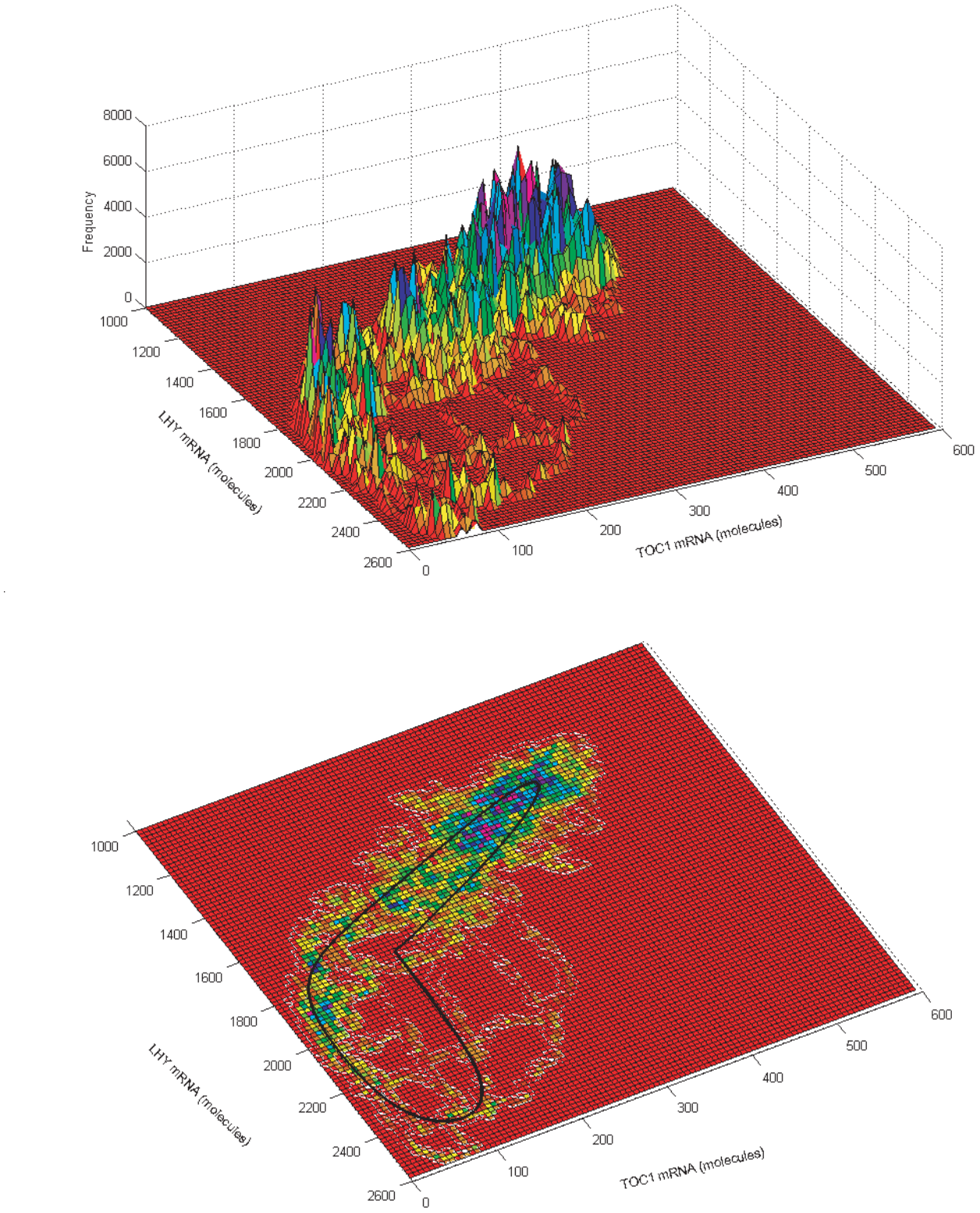

Figure 3.7: Gillespie simulation :$\omega = 1000$, 3-D and 2-D phase space for 12 hours light/dark over a 100 hour period.

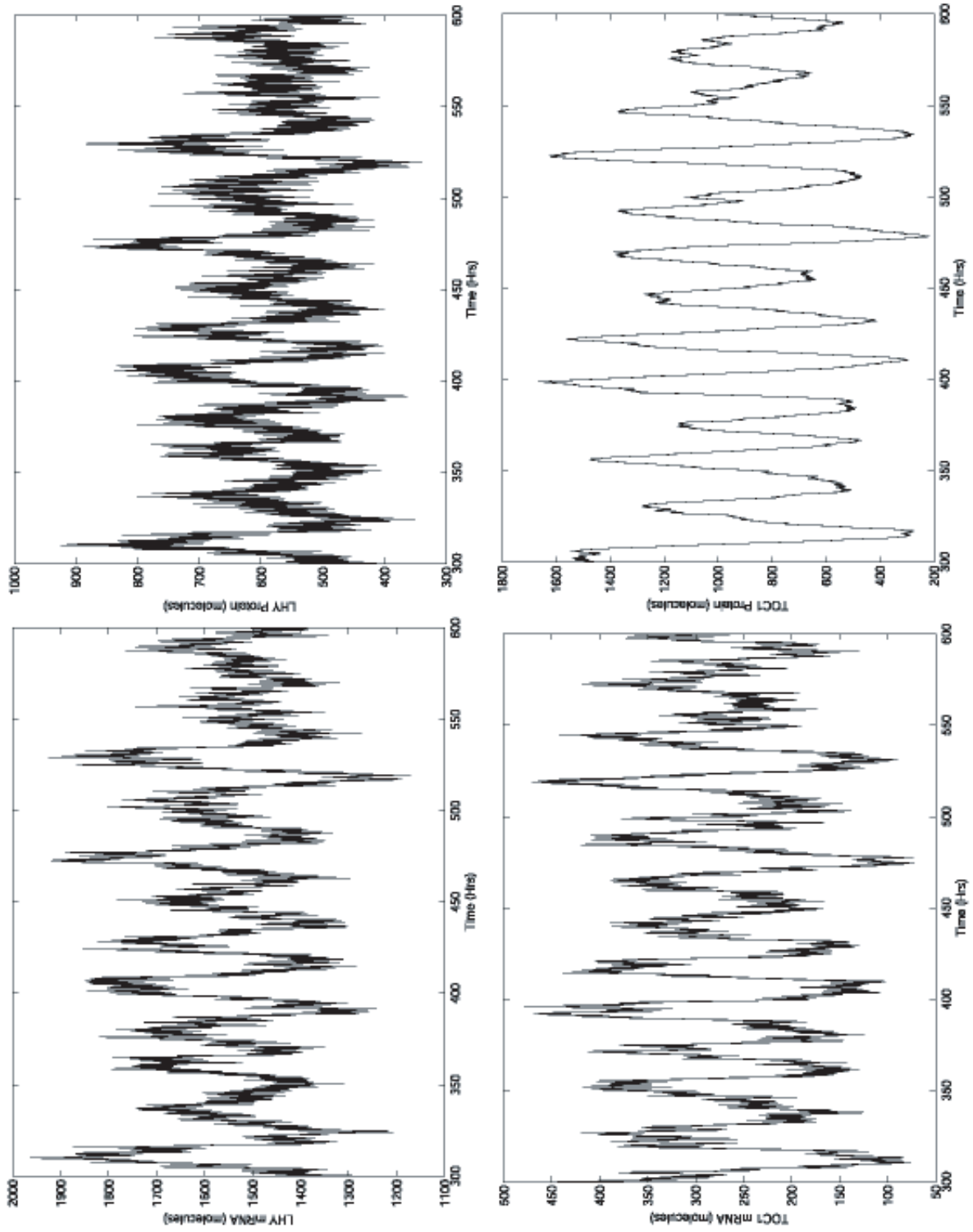

Figure 3.8: Gillespie simulation :$\omega = 1000$, constant darkness

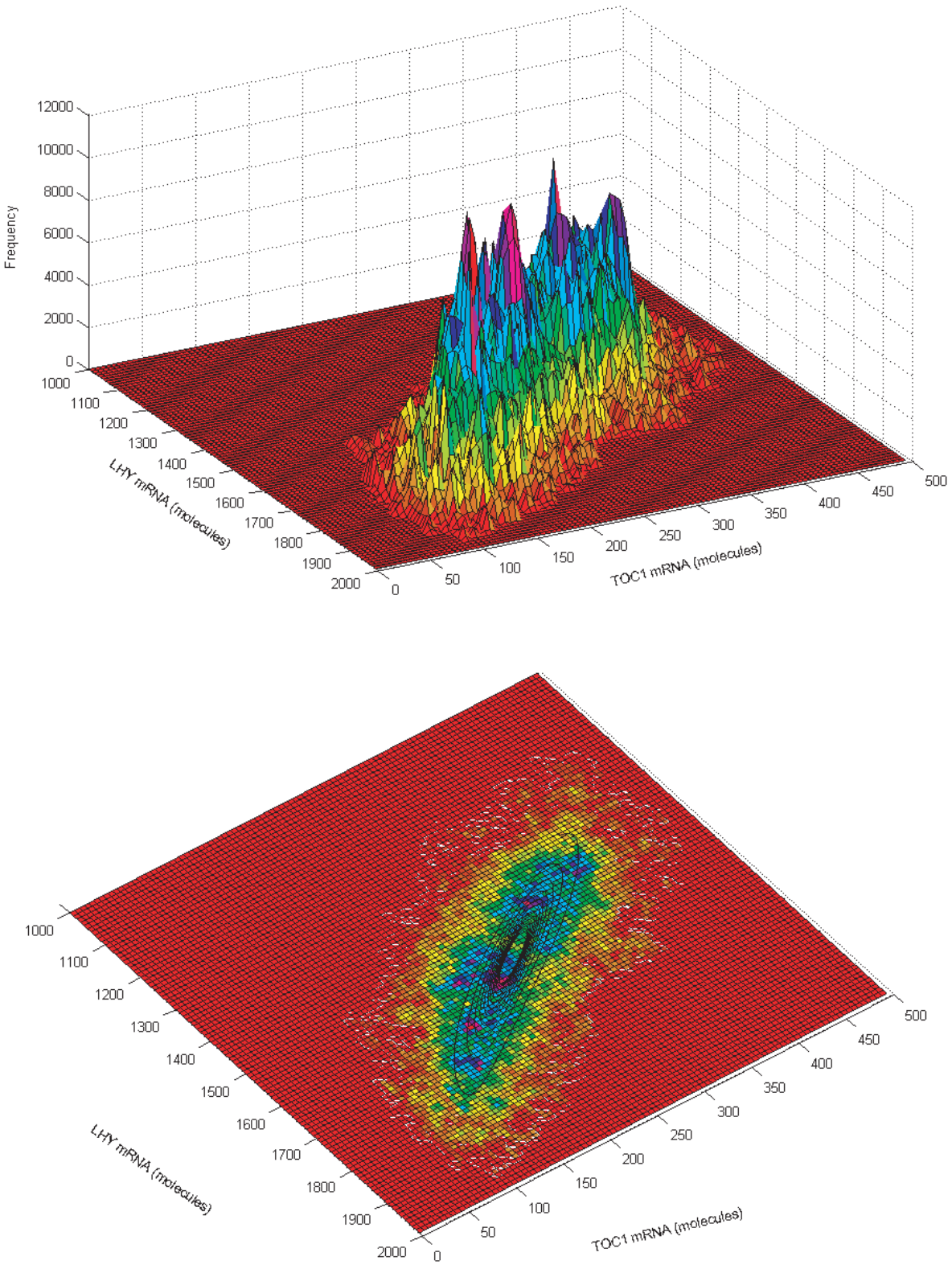

Figure 3.9: Gillespie simulation:$\omega = 1000$, 3-D and 2-D phase space for continuous darkness over a 300 hour period.

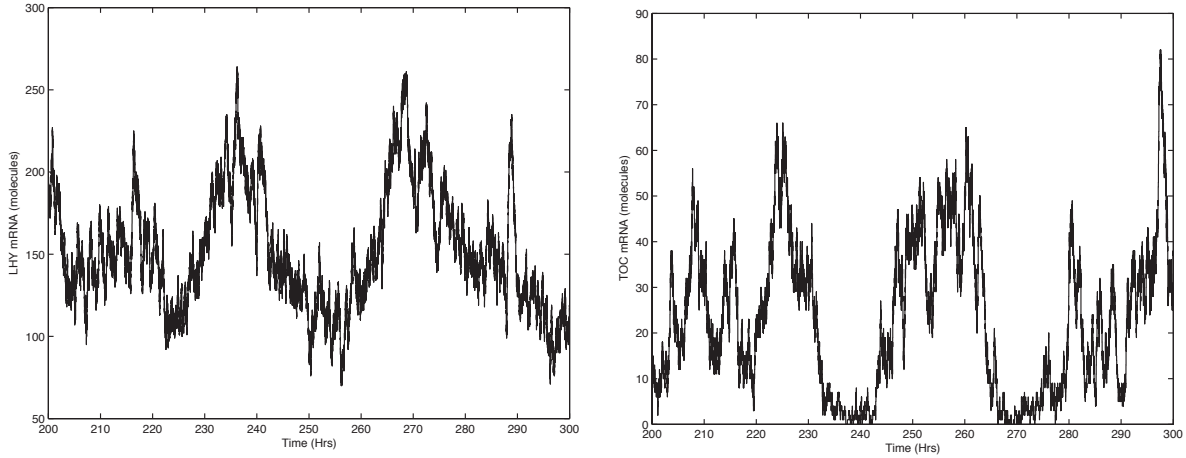

Figure 3.10: Gillespie simulation:$\omega = 100$, 12 hours light/dark.

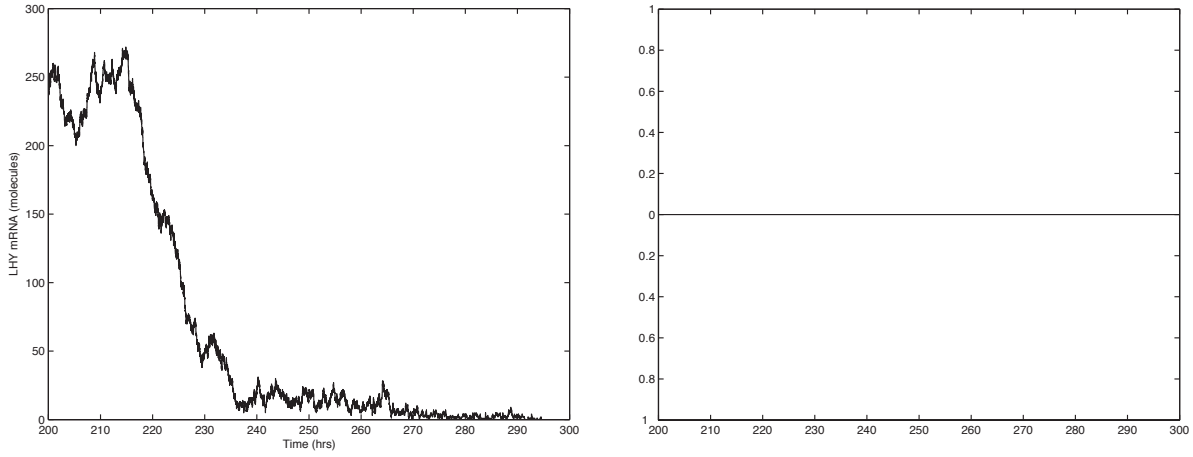

Figure 3.11: Gillespie simulation:$\omega = 10$, 12 hours light/dark.

In this section we have applied a well known numerical algorithm to solve a discrete stochastic model of the *Arabidopsis* circadian clock. Our results show that during the 12 hour light/dark cycle, the stochastic model retains the circadian rhythm for sufficiently large values of the system size ω. However we were unable to observe any sort of periodic behavior for $\omega = 10$, i.e. for a small system. This was not the case for Gonze et al. whose stochastic model for the circadian clock in *Drosophila* still displayed circadian rhythms similar to the deterministic model for numbers of molecules in their tens. This poses a question about whether the network has been modeled correctly or indeed if the network itself is actually correct. Ultimately such a question can only be settled by biologists quantifying the numbers of molecules present. Biologically what are the lowest possible numbers of molecules we may consider before the circadian rhythm is no longer visible?

3.3 SDE Model Arabidopsis Clock

In the discrete stochastic model in the previous section we observed that the circadian rhythm only appeared for moderate to large values of the system size ω. In this section we will continue to increase ω until we begin to observe circadian rhythms similar to the deterministic case for both light/dark and constant darkness cases. Increasing ω is equivalent to increasing the number of molecules, which implies that reactions are likely to occur more frequently. In Chapter 2 we discussed that if a reaction was to occur more frequently in a small time interval then a switch to SDEs seems appropriate. Thus we can transform our discrete model into the following system of SDEs.

$$d\mathbf{X} = \sum_{j=1}^{M} v_j a_j(\mathbf{X})dt + \sum_{j=1}^{M} v_j \sqrt{a_j(\mathbf{X})} d\mathbf{W_j} \tag{3.6}$$

Here $\mathbf{X} \equiv X_i$, $i = 1, ..., 7$ is a vector of the chemical species. We will illustrate the notation by means of an example. X_3 corresponds to L_n which satisfies the following SDE.

(3.7)

$$dL_n = (a_5(L_c) + a_6(L_n) + a_{15}(L_n))dt + \sqrt{a_5(L_c)}dW_5 + \sqrt{a_6(L_n)}dW_6 + \sqrt{a_{15}L_n}dW_{15}.$$

We end up with a system of 7 SDEs which may be solved numerically. Note that we converted the solutions back into concentrations and will be dealing with concentrations for the rest of this Chapter.

Figures (3.12, 3.13, 3.14, 3.15) describe the temporal evolution of the concentrations of LHY mRNA, LHY cytoplasmic protein, TOC1 mRNA and TOC1 cytoplasmic protein respectively. In each of the figures, red denotes single stochastic

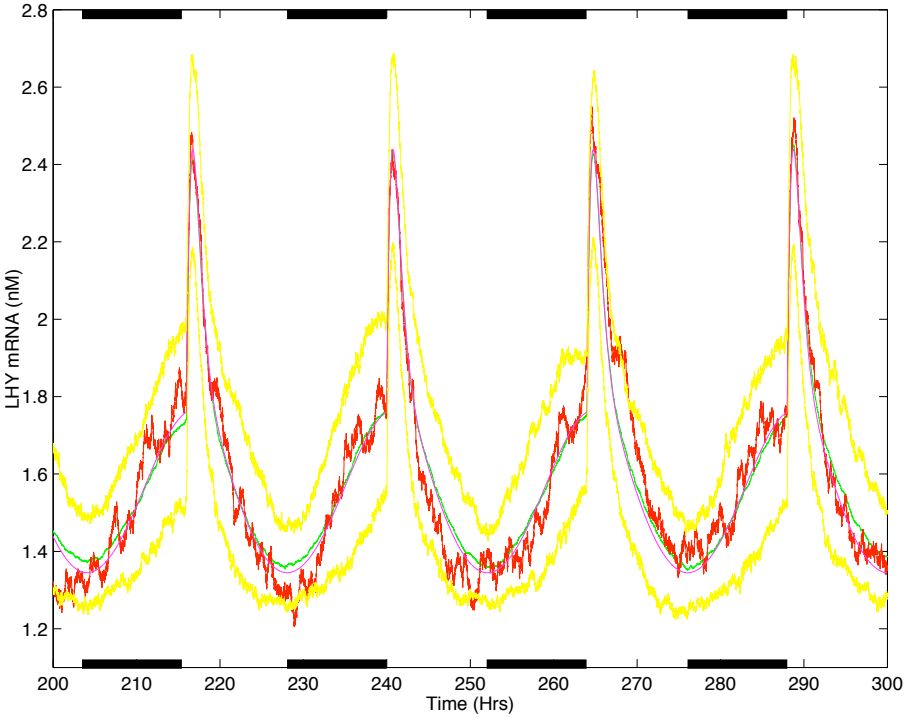

Figure 3.12: SDE simulation:$\omega = 1000$, LHY mRNA, 12 hours light/dark.

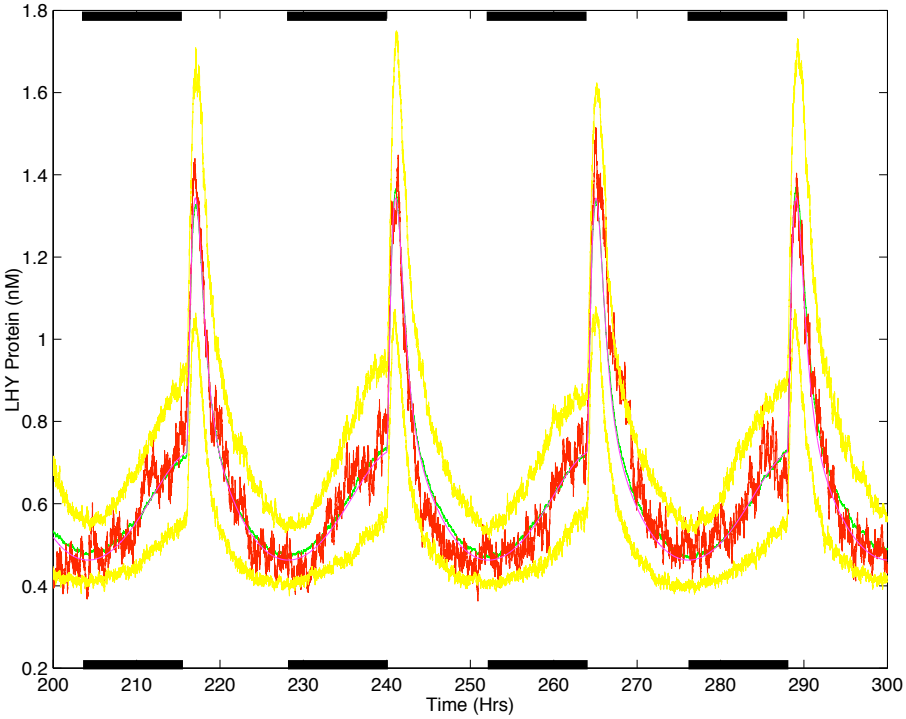

Figure 3.13: SDE simulation:$\omega = 1000$, LHY protein, 12 hours light/dark.

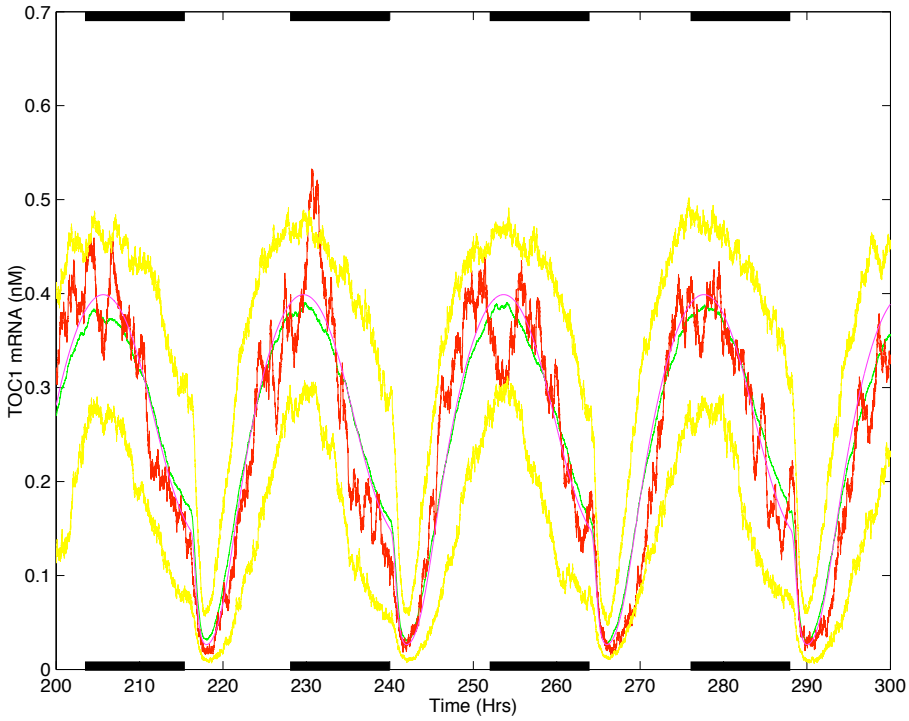

Figure 3.14: SDE simulation:$\omega = 1000$, TOC1 mRNA, 12 hours light/dark.

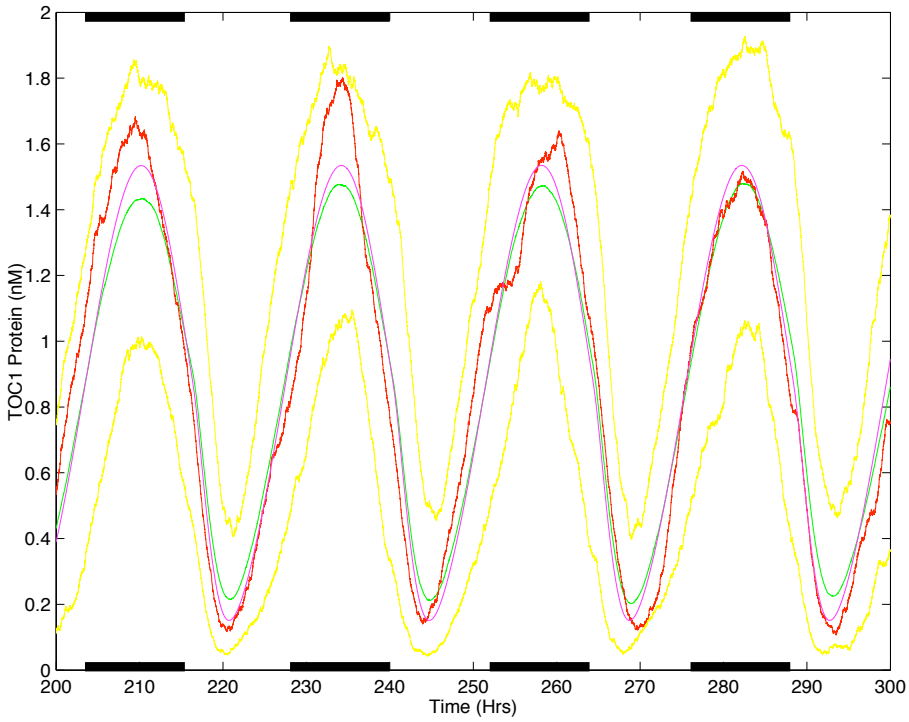

Figure 3.15: SDE simulation:$\omega = 1000$, TOC1 protein, 12 hours light/dark.

realization, magenta the deterministic solution and green the average of 100 stochastic realizations. The two yellow curves represent the 90 percent confidence intervals: that is, they show the distribution the single realizations take over the 100 simulations, ignoring the 5 lowest and 5 highest values at each time step. Notice that qualitatively, the figures (3.12, 3.13, 3.14, 3.15) agree well with the discrete simulations in the previous section. The average behavior for the 12 hour light/dark cycle provides us with further evidence of the stochastic model having established a circadian rhythm similar to the deterministic model in the presence of molecular noise. We obtain similar results qualitatively to the discrete stochastic model for constant darkness which can be seen in figures(3.16, 3.17, 3.18, 3.19). Qualitatively similar behavior is observed in the discrete and continuous stochastic models.

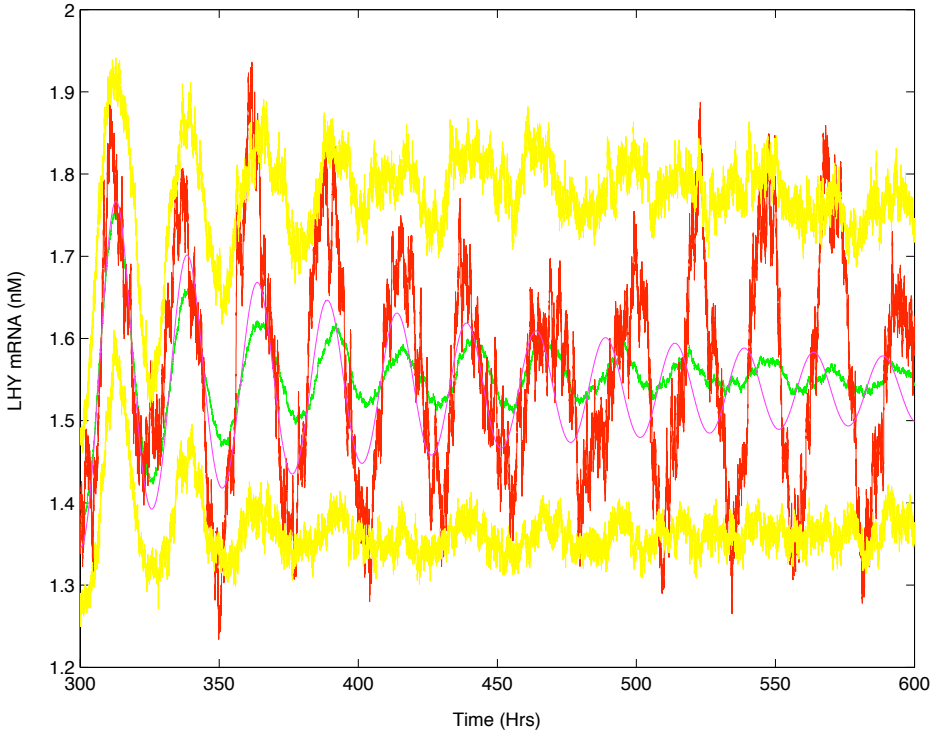

Figure 3.16: SDE simulation:$\omega = 1000$, LHY mRNA, continuous darkness.

However, if we examine the phase space plots seen in figures (3.21, 3.20) we notice that these do not agree well with the ones constructed for the discrete model in figures

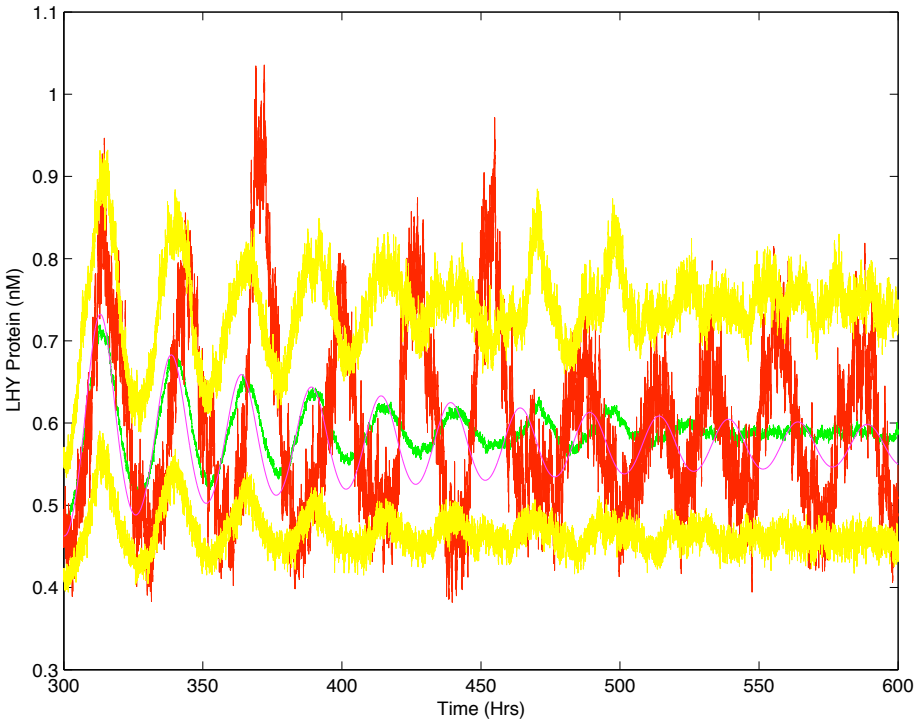

Figure 3.17: SDE simulation:$\omega = 1000$, LHY protein, continuous darkness.

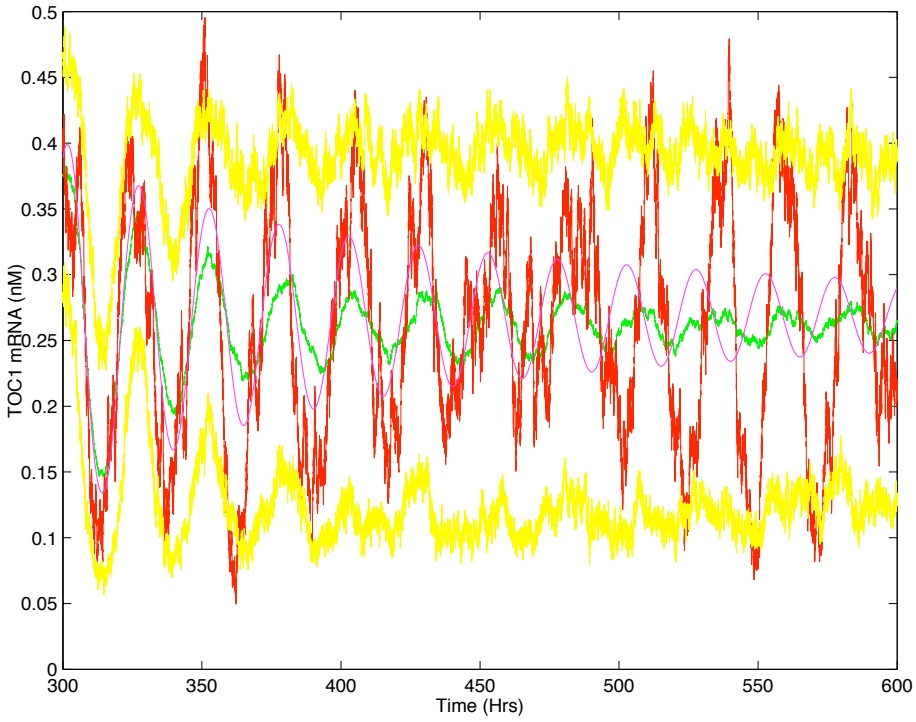

Figure 3.18: SDE simulation:$\omega = 1000$, TOC1 mRNA, continuous darkness.

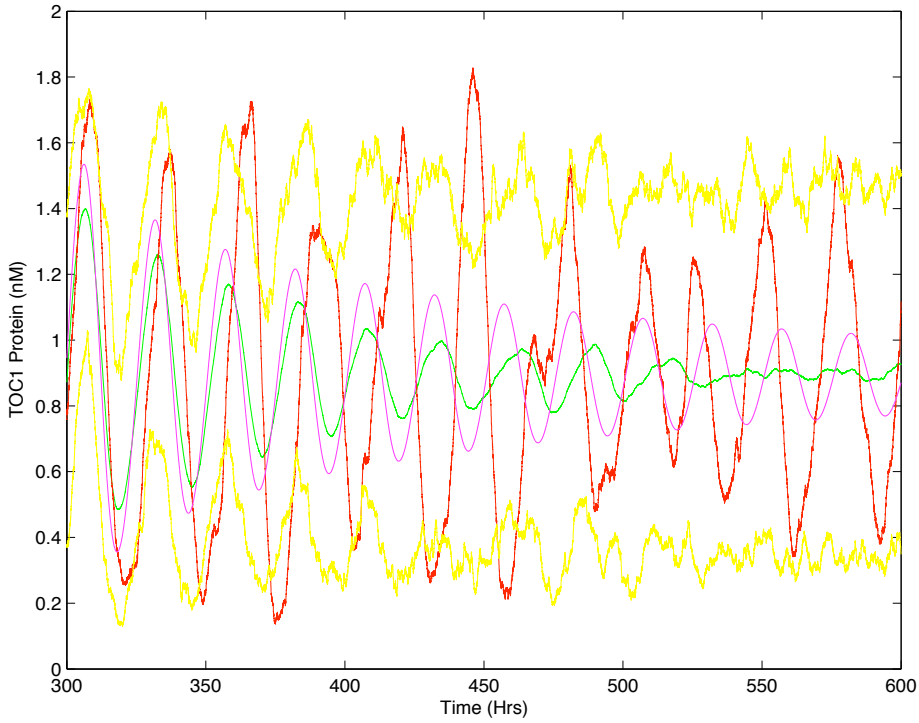

Figure 3.19: SDE simulation:$\omega = 1000$, TOC1 protein, continuous darkness.

(3.7, 3.9). This suggests that we were ill-advised to switch to an SDE approach for this level of system size.

We also performed Kolmogorov-Smirnov tests [29] to look for differences in the distribution of LHY mRNA and TOC1 mRNA. That is, we fixed certain times (shown in table 3.1) in the course of our standard light/dark simulation (alternating 12 hours periods of light and darkness; data collected after an initial 200 hours of simulated time) and, for each, collected values of LHY mRNA and TOC1 mRNA from each of 1000 independent realizations of both the Gillespie and SDE simulations. We then applied the Kolmogorov-Smirnov test to determine whether the two sorts of simulation produced different conditional (conditioned on time) distributions.

The p-values from the Kolmogorov-Smirnov test (seen in table 3.1) show that the two distributions obtained from the Gillespie Algorithm and from the SDE approach have approximately a 65 percent chance of being from the same distribution for $\omega =$

Table 3.1: P-values obtained from performing the Kolmogorov-Smirnov test

Time point (hrs)	$\omega = 1000$ LHY mRNA	$\omega = 1000$ TOC1 mRNA	$\omega = 100$ LHY mRNA	$\omega = 100$ TOC1 mRNA
204	0.6012	0.6451	0.0096	0.0017
216	0.6138	0.7318	0.0046	0.0090
228	0.6068	0.6660	0.0095	0.0079
240	0.5962	0.7186	0.0081	0.0084
252	0.6154	0.6462	0.0028	0.0086
264	0.6038	0.6252	0.0097	0.0037
276	0.6964	0.6046	0.0034	0.0036
288	0.6153	0.6731	0.0022	0.0066
300	0.6468	0.6381	0.0028	0.0079

1000. We also computed Kolmogorov-Smirnov tests for a smaller system size, $\omega = 100$. The p-values obtained for this system size demonstrate the two distributions are very much different, we expect that as we increase ω to 10000 the p-values to be much closer to one. Note that it is compuationally very expensive to increase ω in the discrete model, so we will not be able to perform any more Kolmogorov-Smirnov tests to see if the p-values converge to 1.

If we increase ω, we eventually see the dynamics in the constant darkness phase mimicing the deterministic model in figures(3.22,3.23). Essentially we decrease the amount perturbation of a particle following a stochastic limit cycle.

We have seen in this section that the system size needs to be quite large for the stochastic model to give similar dynamical behavior in the constant darkness phase. The continuous stochastic model also seems to give qualitatively similar behavior to the discrete model for large values of ω as expected.

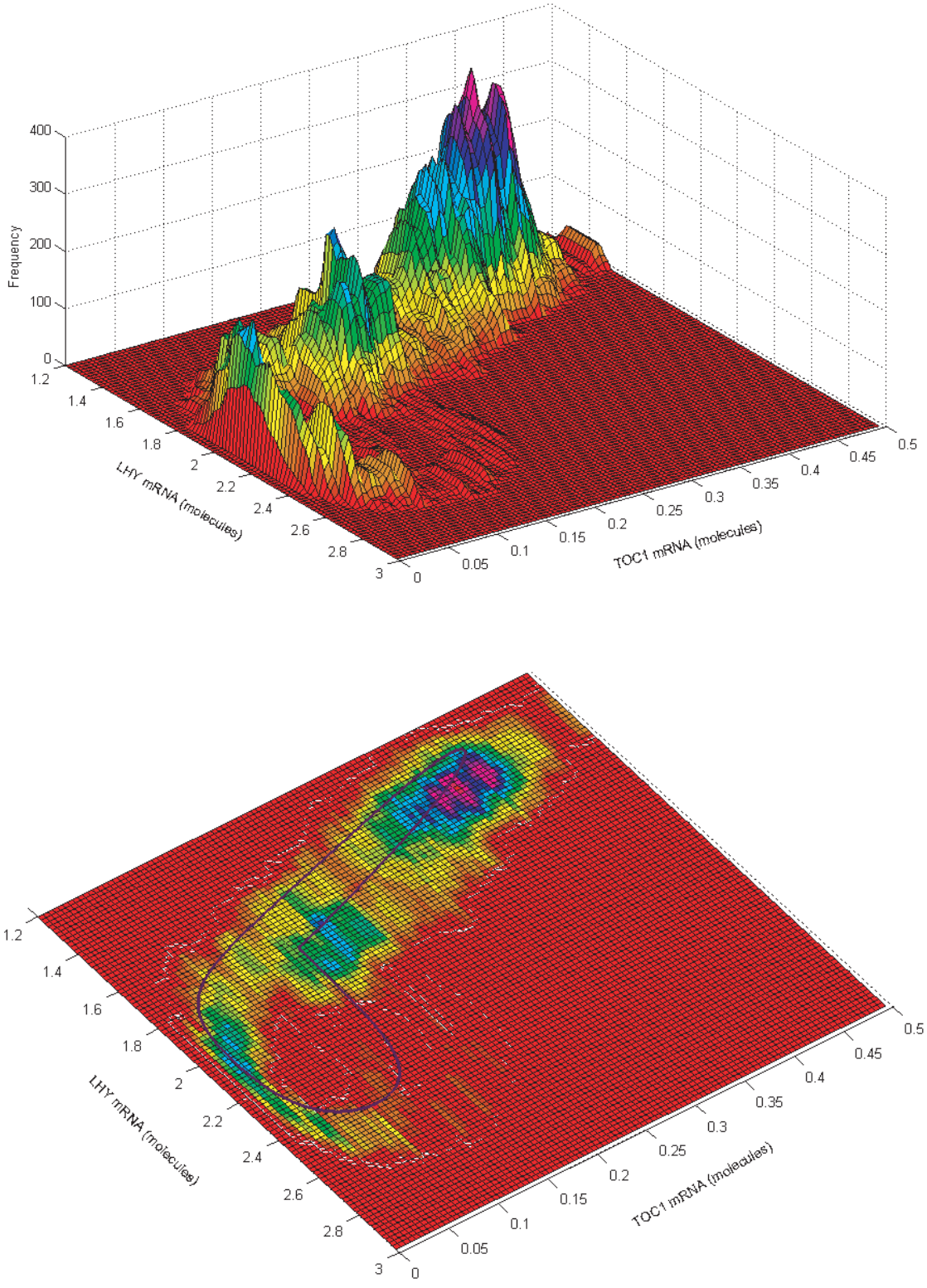

Figure 3.20: SDE simulation:$\omega = 1000$ 12 hours light/dark phase space.

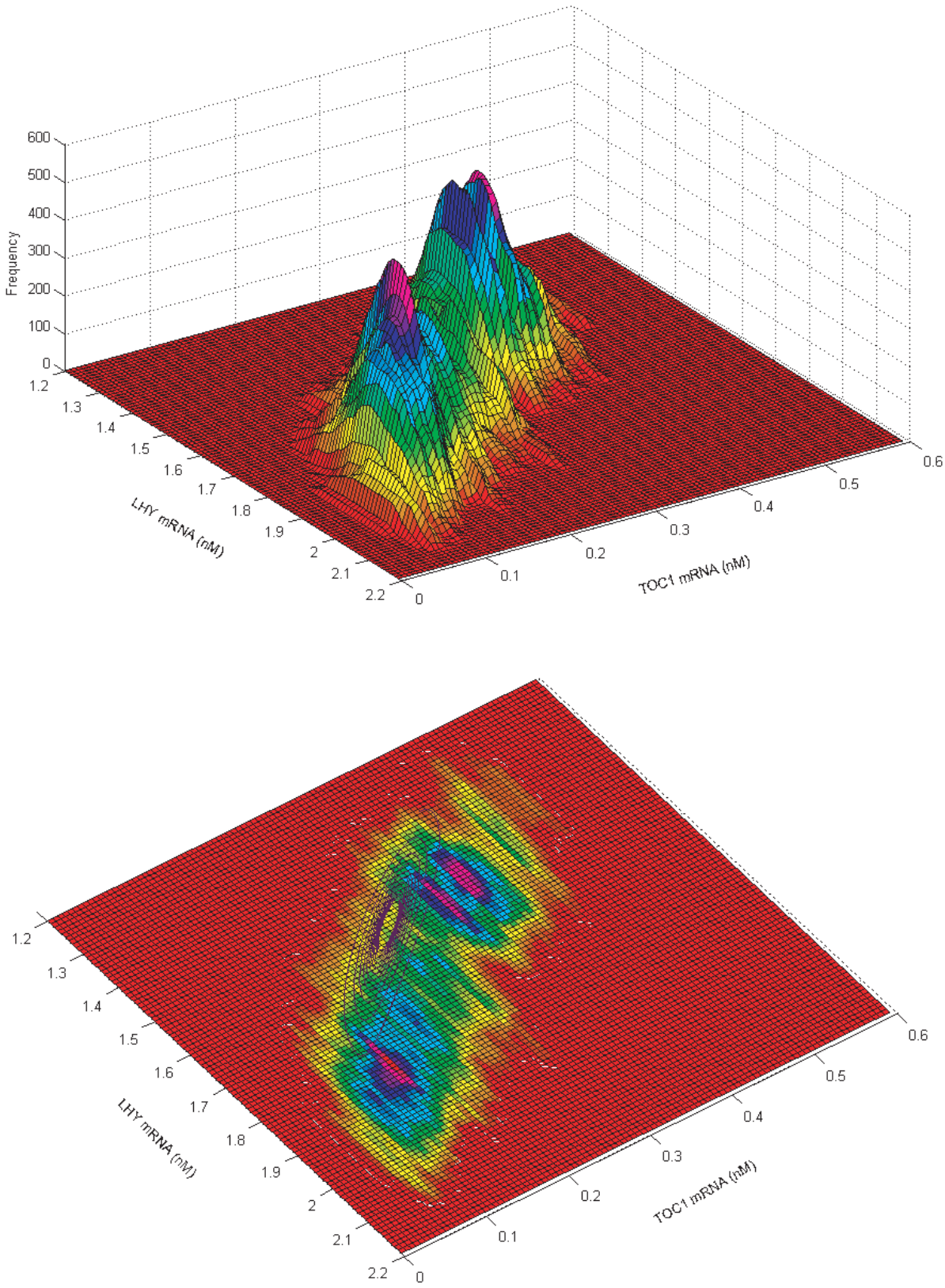

Figure 3.21: SDE simulation:$\omega = 1000$ continuous darkness phase space.

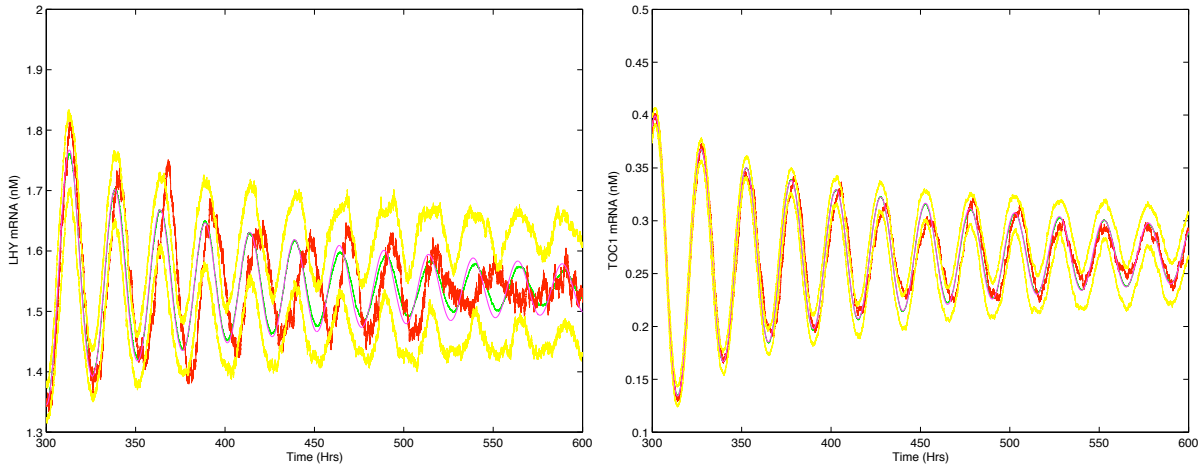

Figure 3.22: SDE simulation:$\omega = 10000$, LHY and TOC1 mRNA, constant darkness.

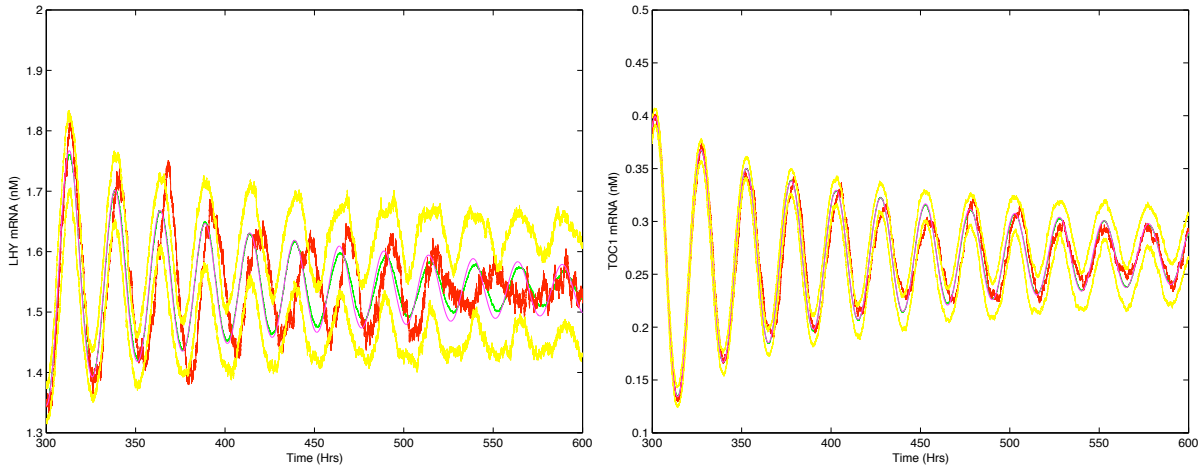

Figure 3.23: SDE simulation:$\omega = 100000$, LHY and TOC1 mRNA, continuous darkness.

3.4 Perturbation of Parameters

In this section, we are interested in seeing if the circadian rhythm produced by the deterministic model persists when we perturb transcription and translation rates in a stochastic way. Biologically, this is worth investigating as transcription and translation rates within a single cell are not constant and probably vary over time, as discussed in the introduction of this chapter. Note from the offset that we are now dealing with concentrations not whole numbers. We will take the transcription and translation rates and perturb them in the following way.

$$n_k \mapsto n_k(1 + \alpha W_{tc_i}(t)) \tag{3.8}$$

$$p_k \mapsto p_k(1 + \beta W_{tl_j}(t)) \tag{3.9}$$

This transforms certain ODEs in the original model, into the following SDE's,

$$\begin{aligned} dL_m &= (f_1)dT + \alpha\left(\frac{n_1 T_n^a}{g_1^a + T_n^a}\right) dW_{tc_1} \\ dL_c &= (f_2)\, dT + \beta\,(p_1 L_m)\, dW_{tl_1} \\ dT_m &= (f_4)\, dT + \alpha\left(\frac{n_2 g_2^b}{g_2^b + L_n^b}\right) dW_{tc_2} \\ dT_c &= (f_5)\, dT + \beta\,(p_2 T_m)\, dW_{tl_2} \\ dP_n &= (f_7)dT + \beta((1 - \Theta_{light})p_3)dW_{tl_3}, \end{aligned} \tag{3.10}$$

thereby giving us a mixed system of non-linear SDE's and ODE's. Here $\alpha, \beta \in \mathbb{R}$. α is chosen to represent intrinsic noise: fluctuations that are inherent to the system and β is chosen to represent extrinsic noise, both of which lead to cell-to-cell variability.

Choosing suitable levels for α and β is the next step. The experimental work of Elowitz and of Colman-Lerner discussed in Chapter 1 suggests that $\alpha \ll \beta$. In

an ideal scenario, we would have experimental observations to help decide the level of noise to be introduced. Unfortunately as is often the case when modeling biochemical systems, data is hard to come by. Without any experimental guidance we control the noise by hand. The following figures show that very small fluctuations in transcription and translation rates preserve the circadian rhythm seen in the deterministic model, figures(3.24, 3.25, 3.26, 3.27). As we increase the strength of the perturbations we find that the dynamics for constant darkness no longer produce the correct oscillatory behavior, figures (3.28, 3.29, 3.30, 3.31). Finally we can increase the perturbations to a point where the noise has destroyed the circadian rhythm completely for both the light/dark and constant dark cycles.

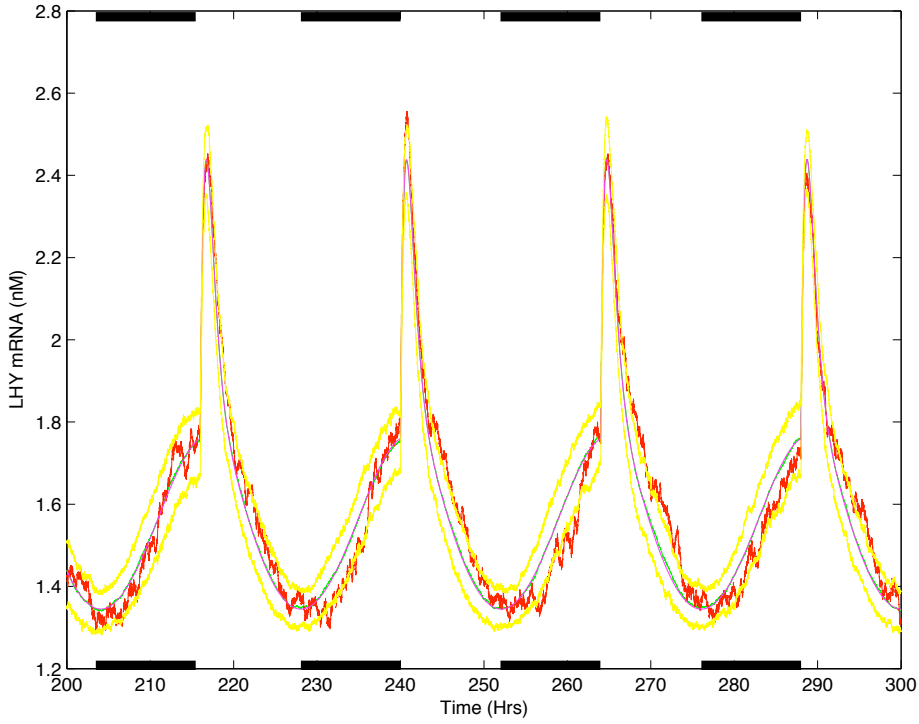

Figure 3.24: $\alpha = 0.001$, $\beta = 0.01$ LHY mRNA, 12 hours light/dark.

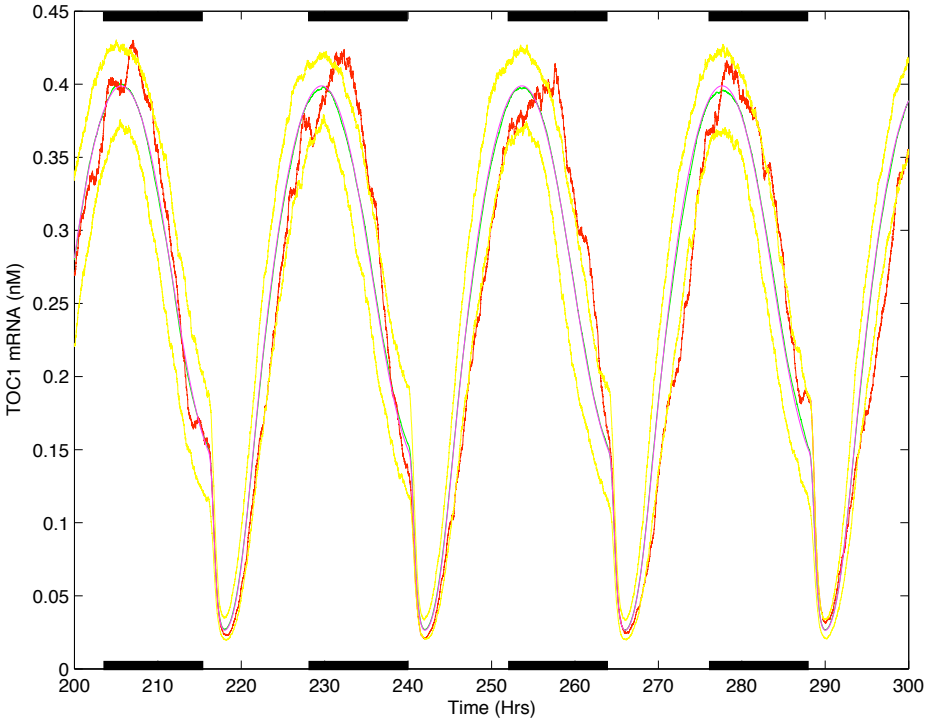

Figure 3.25: $\alpha = 0.001$, $\beta = 0.01$ TOC1 mRNA, 12 hours light/dark.

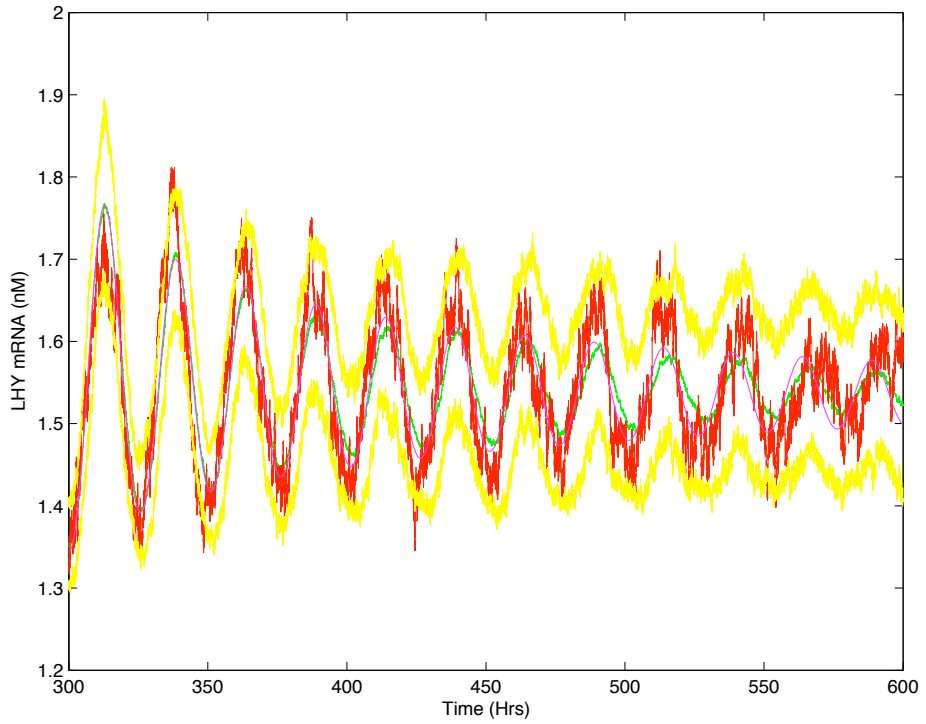

Figure 3.26: $\alpha = 0.001$, $\beta = 0.01$ LHY mRNA, constant darkness.

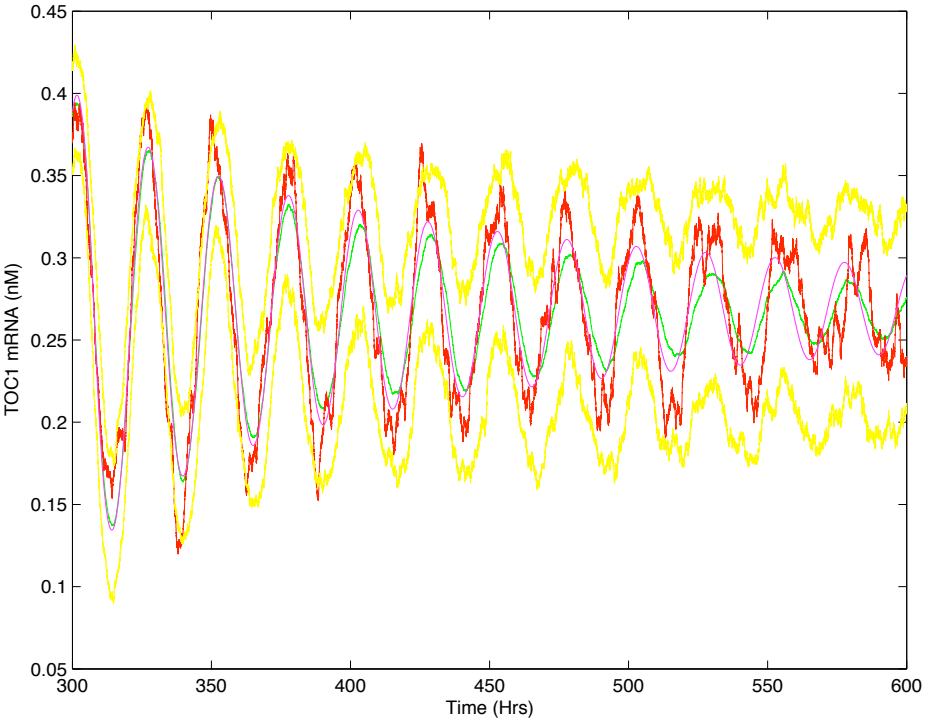

Figure 3.27: $\alpha = 0.001$, $\beta = 0.01$ TOC1 mRNA, constant darkness.

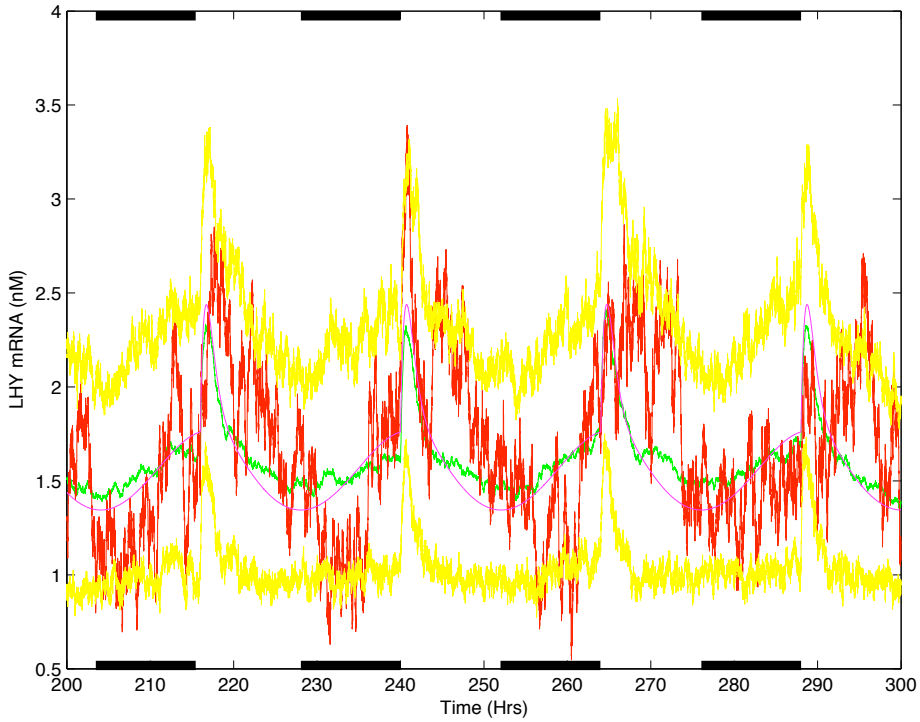

Figure 3.28: $\alpha = 0.01$, $\beta = 0.1$ LHY mRNA, 12 hours light/dark.

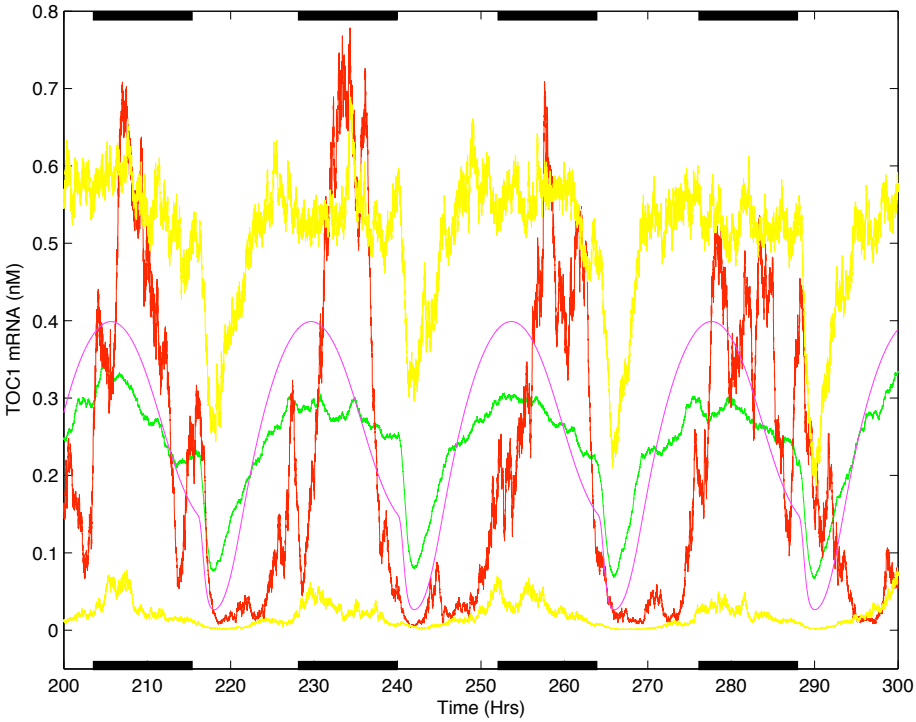

Figure 3.29: $\alpha = 0.01$, $\beta = 0.1$ TOC1 mRNA, 12 hours light/dark.

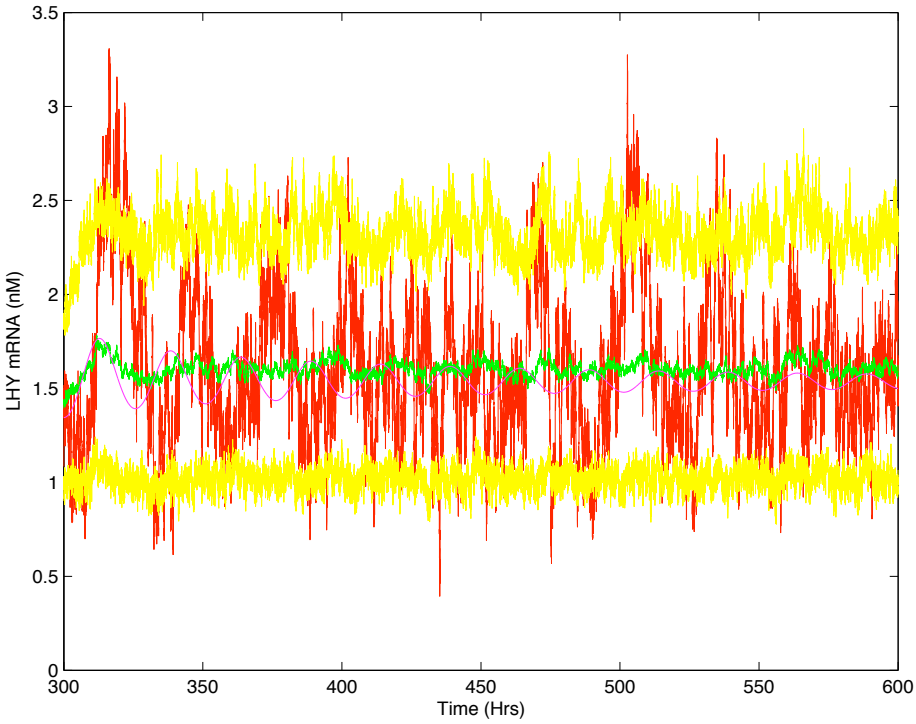

Figure 3.30: $\alpha = 0.01$, $\beta = 0.1$ LHY mRNA, constant darkness.

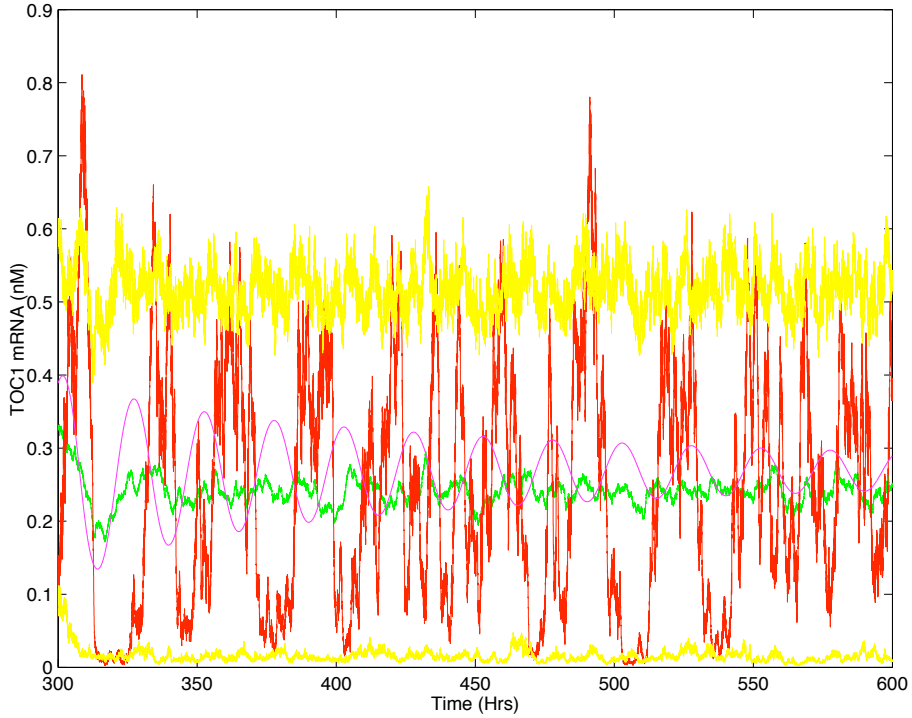

Figure 3.31: $\alpha = 0.01$, $\beta = 0.1$ TOC1 mRNA, constant darkness.

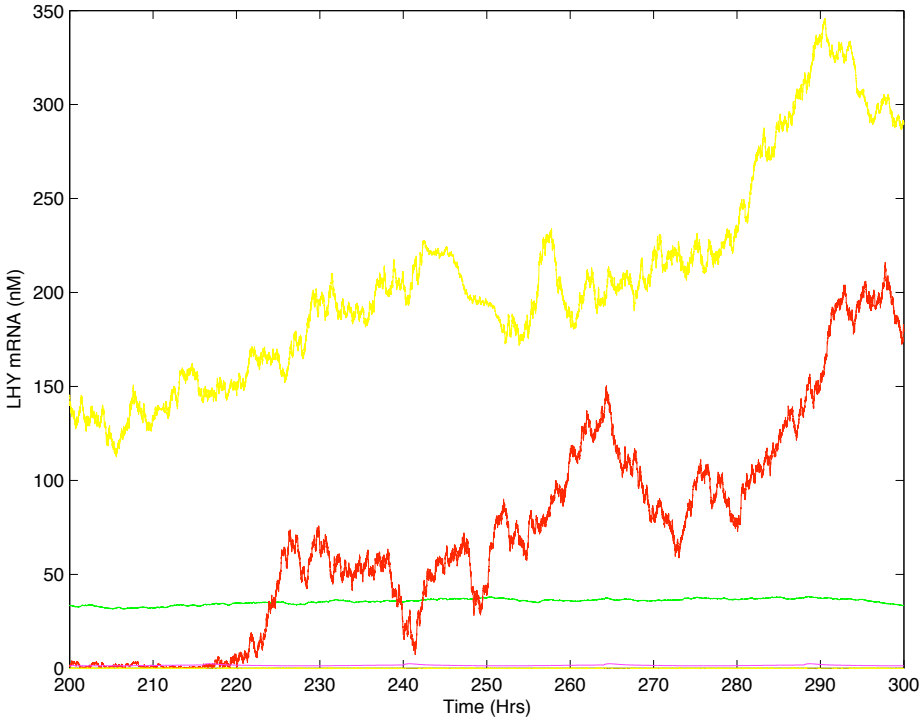

Figure 3.32: $\alpha = 0.1$, $\beta = 1$ LHY mRNA, 12 hours light/dark.

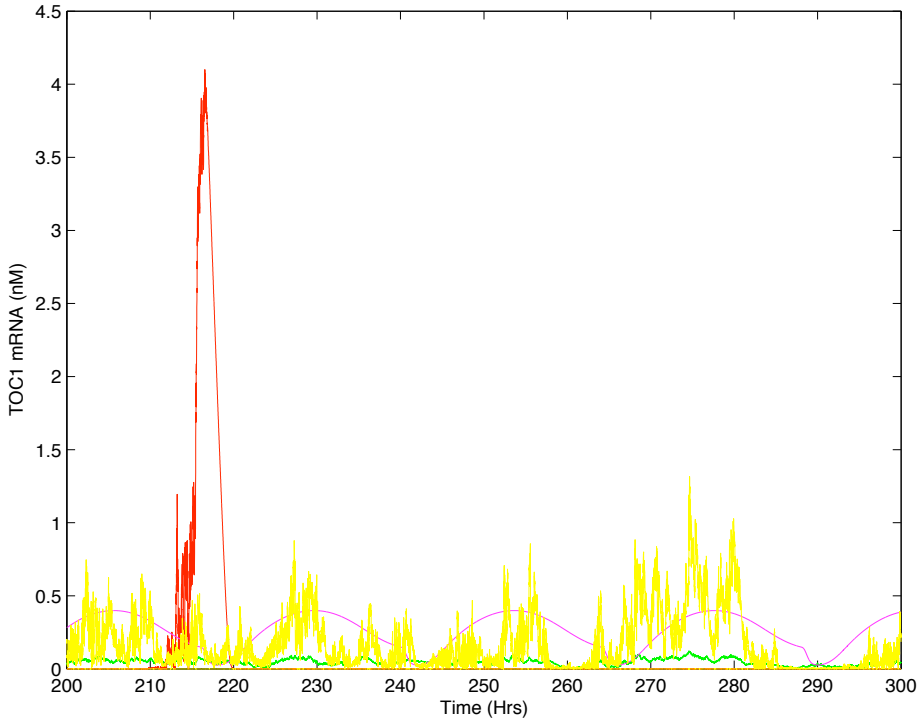

Figure 3.33: $\alpha = 0.1$, $\beta = 1$ TOC1 mRNA, 12 hours light/dark.

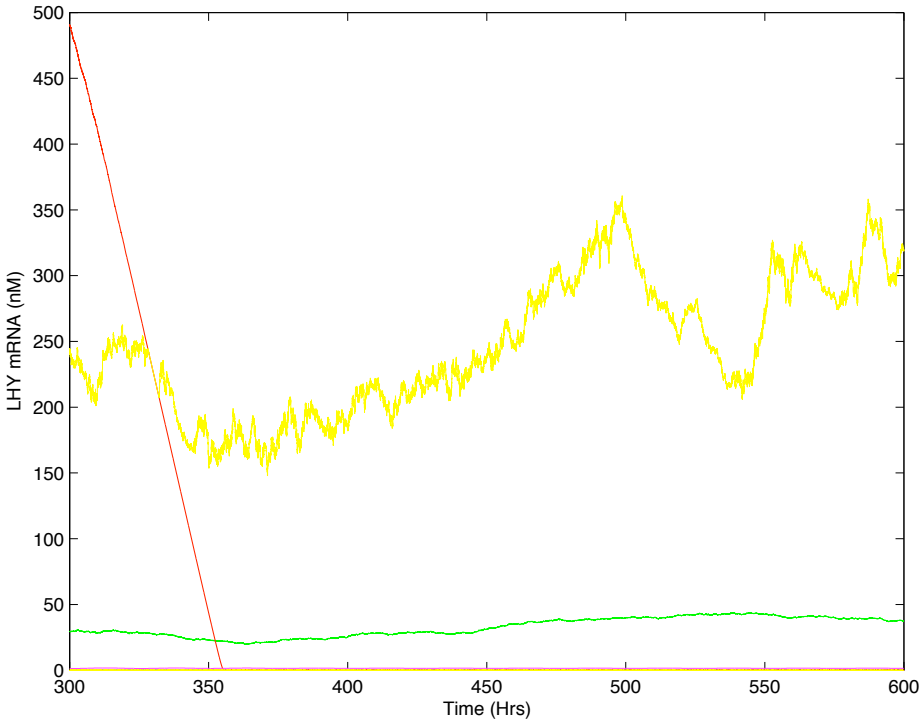

Figure 3.34: $\alpha = 0.1$, $\beta = 1$ LHY mRNA, constant darkness.

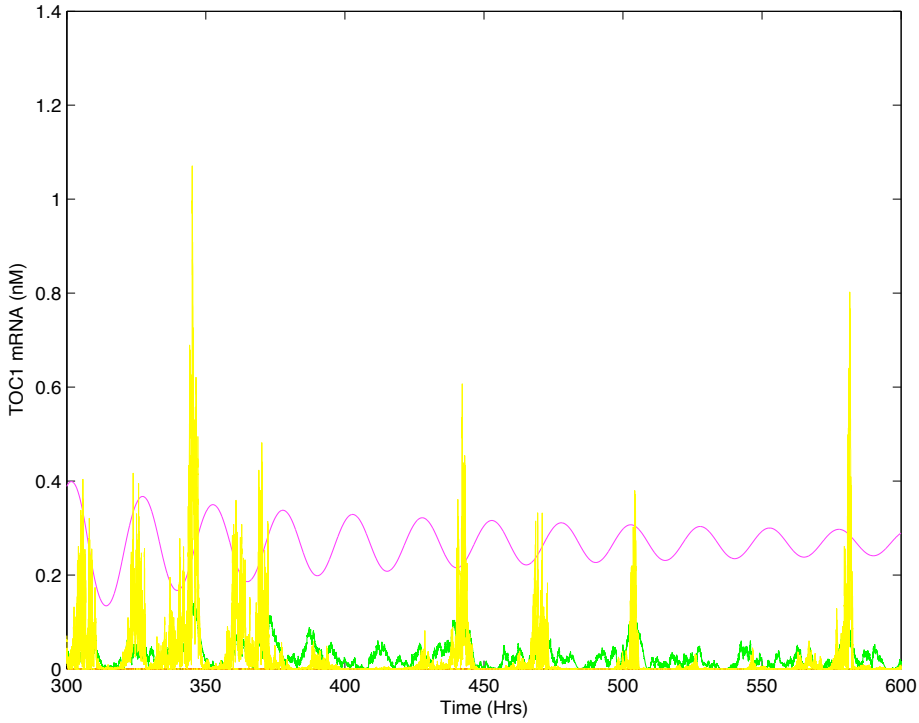

Figure 3.35: $\alpha = 0.1$, $\beta = 1$ TOC1 mRNA, constant darkness.

3.5 Summary

At the beginning of the chapter we introduced 3 constraints by which circadian clocks must abide as stated by Dunlap [11]. By introducing molecular noise by means of discrete and continuous stochastic models, we have been able to assess the robustness of a model of the *Arabidopsis* clock.

The discrete stochastic model showed the circadian rhythm in the light/dark cycle still persisted when the numbers of molecules in the system were no more than a few thousand. Reducing the number of molecules to even a few hundred renders circadian entrainment unfeasible, in contrast to models studied by Gonze et al. When there are only a few tens of molecules in the system, the circadian rhythm is completely destroyed by the noise. In fact we see no oscillations at all. We suggested that this poses interesting questions to both biologists and modelers. Biologists

should be asked to investigate how many molecules are needed to produce circadian rhythms. They could measure the numbers of molecules, at least approximately, by, say, quantitative flourescence: even this would be a fairly heroic measurement. Modelers may be interested in decomposing the reactions that are here summarized with Hill functions into simpler steps: it does not seem sensible to use Hill functions for such small numbers of molecules.

The behaviour in constant darkness proved much less robust to stochastic perturbation, with agreement between stochastic and deterministic models appearing only for very large system sizes. The model seems to be very sensitive to molecular noise in the constant darkness phase, much more so than when forced with a 12 hour light/dark cycle.

Dunlap also stated that circadian clocks must be robust to fluctuating transcription and translation rates. We considered white noise perturbations of transcription and translation rates, but without knowing how much they vary we are able to make only the most tentative of biologically relevant comments. All the points mentioned here suggest the *Arabidopsis* circadian clock is, at least under normal diurnal light variation, robust with respect to molecular noise and parameter variation. The observations seen in the dark/dark cycles may be overlooked since nature never intended plants to function in constant darkness.

CHAPTER IV

Stochastic Score Functions

The idea of robustness had been discussed in the previous chapter by examining the effects molecular noise and parameter variation had on the circadian clock. Recall that only for large system size did the desired circadian rhythm persist for both light/dark and dark/dark cycles. Where the light/dark phrase refers to the model being subjected to 12 hours of light then 12 hours of darkness for the first 300 hours and that dark/dark refers to the model being plunged into continuous darkness after 300 hours of light/dark entrainment. In this chapter we assess what affect molecular noise has on the score functions designed by Locke et al. which in essence is a comparison with experiments. The simulations under consideration in these score functions are the ones obtained from the SDE model of the *Arabidopsis* clock. This model was used since it is easier to analyze data with discrete time-steps, 0.0005 hours. We will first explain what principles were used in designing the score functions, before moving on to describe how we smoothed the stochastic realizations so that we could examine the periodic behavior of LHY/CCA1 and TOC1 mRNA levels that were observed in the previous chapter.

The score functions designed in Locke et al. [31] described how well a solution from their model matched experimental behavior in a qualitative way rather than a

quantitative way. The idea is that they wanted to ensure that the solutions from their model shared certain qualitative features with observations from the experiments and so they defined terms in the score function that make an order-one contribution for a prescribed size of deviation from "correct" behavior. They chose this approach due to a lack of time points in the experiments and also since the biological data obtained from these experiments is quite noisy it seems inappropriate to compare a model quantitatively to experimental data.

4.1 Extracting qualitative features from stochastic realizations

If we look back to the previous chapter and examine the single realizations of the stochastic simulation; notice how its not so easy to examine the periodic behavior with all the noise. To combat this we decided to employ a *simple moving window average filter* [36] to smooth the data.

If we were to have a series of equally spaced data points $f_i \equiv f(t_i)$ where $i = \ldots -2, -1, 0, 1, 2, \ldots$. In a simple moving window average we replace each f_i by a linear combination g_i of itself and its neighbors,

$$g_i = \sum_{n=-n_L}^{n_R} c_n f_{i+n}. \tag{4.1}$$

Where $c_n = 1/(n_L + n_R + 1)$, n_L is the number of points taken into account earlier then the data point i and n_R number of points used after it. Note that $n_L = n_R = 500$ was fixed for all sets of data points passed through this filter.

The smoothing filter was applied to each realization obtained from the stochastic simulation. With the smoothed data we were now able to calculate the *period*. To calculate the period of an expressed gene we first need to find the average concentration of that gene in the time interval under review i.e. in the light/dark phase

we found the average expressed concentration between 200 and 300 hours. We then noted the time points that the smoothed data crossed this average. The *period* was taken to be the time taken to cross the average three times. In this time window we could then search for the peak time of expression and its corresponding concentration. We can also search for the minimum concentration and corresponding time point. The necessity of calculating these values will become apparent in the next section when we consider the score functions. All the concentrations used in the score functions were averages taken 0.025 hours either side of the desired time points.

4.2 The modified cost functions

The original cost functions used in parameter searches for the deterministic model needed to be altered to account for the stochasticity generated by the SDE's in the new model. The original cost functions were too harsh on individual realizations and were softened accordingly.

The first cost function examines the summed error in the period, τ, over a light/dark cycle over a period of 100 hours, $200 < t < 300$. The notation $\langle\rangle_{ld}$ means "average over this period". In the original function the acceptable period difference was about 25 minutes from the desired period of 24 hours. Our cost function does not punish period differences over 25 minutes as harshly.

$$\delta_{\tau_{ld}} = \sum_{i=L,T} \langle (24 - \tau_i^{(m)})^2 / 0.75 \rangle_{ld} \tag{4.2}$$

The second function examines the summed error in the period over the constant darkness period, considering a time interval of 300 hours, $300 < t < 600$. Substantial biological studies, mentioned in [31], provides evidence that the free running clock has a period greater than 24 hours, closer to 25 hours.

$$\delta_{\tau_{dd}} = \sum_{i=L,T} \langle (25 - \tau_i^{(m)})^2 / f \rangle_{ld} \tag{4.3}$$

Where $f = 0.5$ if $\tau_i^{(m)} \leq 25$ and $f = 2$ if $\tau_i^{(m)} > 25$.

The third cost function, which is concerned with the timing of events near dawn,

$$\delta_\phi = \sum_{i=L,T} \left[\langle \Delta\Phi_i^2 \rangle_{ld} + \left(\frac{\sigma[c_i^{(m)}(t_p)]_{ld}}{0.1 \langle c_i^{(m)}(t_p) \rangle_{ld}} \right)^2 + \left(\frac{\sigma[\Delta\Phi_i]}{15/60} \right)^2 \right] \tag{4.4}$$

The first term looks at the mean difference between the time from dawn at which the RNA levels peak from dawn, in the light/dark cycles. $\Delta\Phi_i = \overline{\phi_i} - \phi_i$, where ϕ_i is the phase from dawn, $\overline{\phi_L} = 1h, \overline{\phi_T} = 11h$ are the target phases of LHY mRNA and TOC1 mRNA respectively. The next two terms look at the consistency of the oscillations. The first makes an order one contribution to the score when the peak heights are within 10 percent of their mean, while the second requires that variations in peak phases are no more than about 15 minutes. Here $\sigma[]_{ld}$ is the standard deviation for cycles in light/dark.

The fourth cost function,

$$\delta_{size} = \sum_{i=T} \left(\frac{\tau_0}{\tau_e} \right)^2 \tag{4.5}$$

checks to make sure that the oscillations do not decay too quickly after entering the dark:dark phase. Here

$$\tau_0 = -300/log((\Delta c_{T_{ld}}^{(m)} - \Delta c_{T_d}^{(m)})/\Delta c_{T_{ld}}^{(m)})$$

where $\Delta c_i^{(m)} = c_{i_{max}}^{(m)} - c_{i_{min}}^{(m)}$, gives an estimate of the rate of decay of the peak amplitudes of the oscillations and $\tau_e = -300/log(0.75)$ sets the scale for an acceptable drop in size of the TOC1 oscillations: down by a quarter over 300hs.

The last cost function aims to constrain the qualitative form of the time course of LHY mRNA. The first term checks to see if a sharp peak exists and requires that LHT mRNA drops by 2/3 of its peak amplitude within 2 hours before and after the peak. The second term checks to see if the expression has a broad minimum, producing an O(1) contribution if the expression has increased by only 5 percent of the level 2 hours before LHY's peak, both 2hrs before and 2hrs after the minimum point. The last term checks to see if the expression levels drop as we move from light/dark to constant darkness.

$$\delta_{c_L} = \sum_{i=2,-2} \left\langle \left(\frac{2/3 c_L^{(m)}(t_p)}{c_L^{(m)}(t_p) - c_L^{(m)}(t_p + i)} \right)^2 \right\rangle_{ld} + \ldots \left\langle \left(\frac{0.05(c_L^{(m)}(t_p - 2) - c_L^{(m)}(t_m))}{c_L^{(m)}(t_m) - c_L^{(m)}(t_m + i)} \right)^2 \right\rangle_{ld} + 10 \left(\frac{\langle c_L^{(m)}(t_{pd}) \rangle_{ld}}{\langle c_L^{(m)}(t_{pl}) \rangle_{ld}} \right)^4 \tag{4.6}$$

4.3 Comparing the best parameters

Locke et al. [31] solved the original set of ODEs, described in the previous chapter, for 10^6 random points generated using a variant of the Sobol Algorithm [36]. They then calculated the score functions for all these points. Of which they took the top 50 solutions and passed them through a simulated annealing routine. For information about the Sobol Algorithm and annealing routines we refer the reader to Locke et al. [31] and Numerical Recipes in C++ text [36].

Figure (4.1) shows the score of the 48 annealed parameter sets obtained with the modified score functions. Two of the 50 annealed sets of parameters were too stiff for our numerical method and have thus been left out of this study.

In the previous chapter we found certain sizes of ω that gave us the desired circadian behavior, $\omega = 1000$, $\omega = 10000$ and $\omega = 100000$. These values will be used

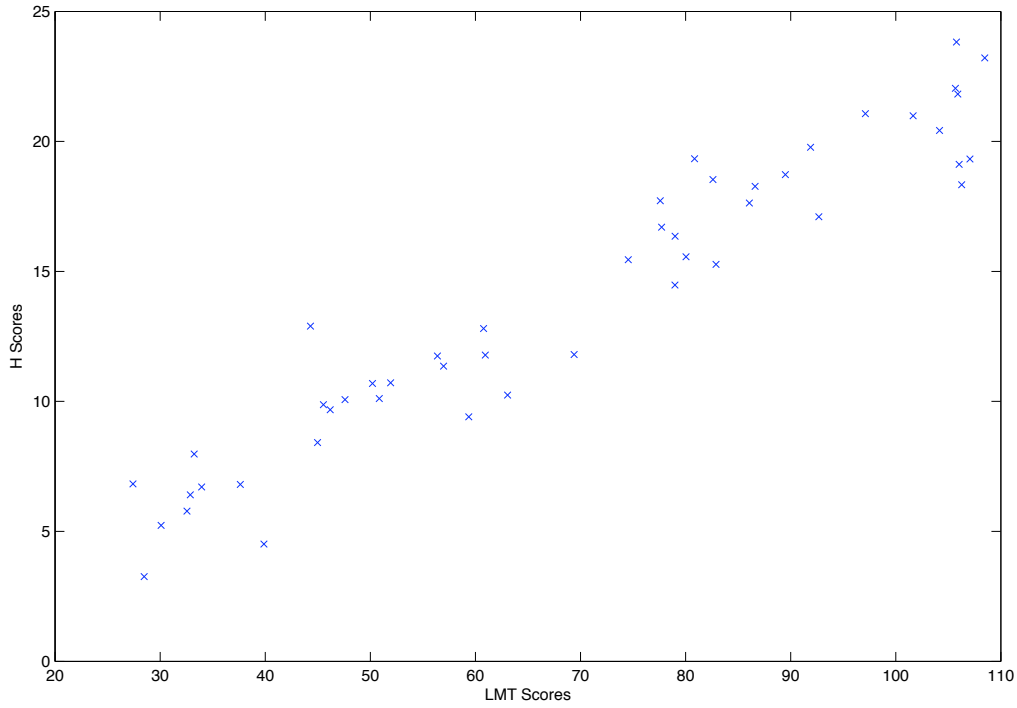

Figure 4.1: The result of comparing the scores obtained for trajectories of the deterministic model under both the original and modified scoring schemes.

to run stochastic simulations for the 48 parameter sets. The average score taken over 20 simulations for each parameter set will then be compared to the new deterministic scores. As we increase system size we expect the stochastic scores to converge to the deterministic scores in figure 4.1.

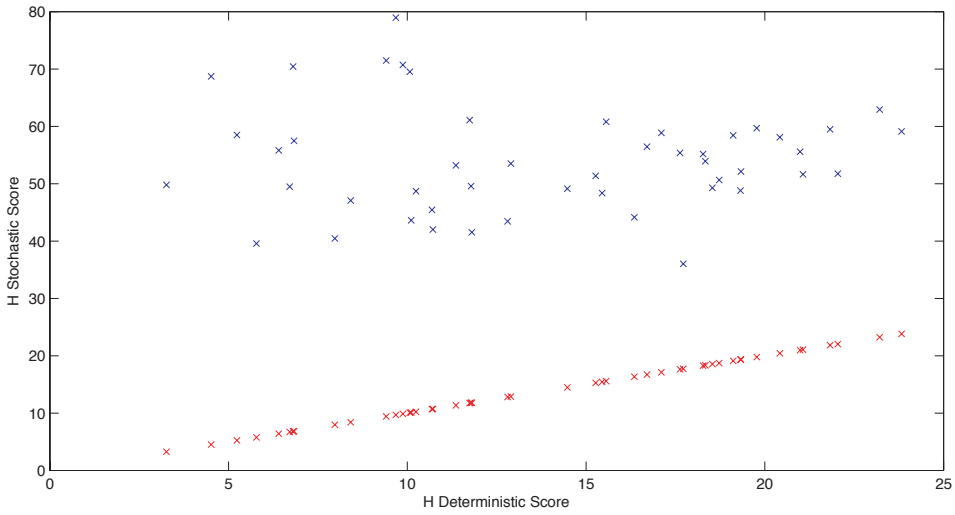

Figure 4.2: $\omega = 1000$, blue: stochastic scores, red: deterministic scores

Figure 4.2 shows the score values for $\omega = 1000$ which we deemed a sensible value in the previous section. As we expected the points are a reasonable distance from the deterministic scores. We feel that the main reason the scores are so high is due to the fact that in the dark/dark cycle the oscillations have not dropped sufficiently. Recall that the fifth score function checks to see if LHY mRNA expression has dropped sufficiently as we move from light/dark to dark/dark. Thus if expression levels have not dropped sufficiently then it is punished by the last term in the fifth score function.

As we increase ω further, figures 4.4 and 4.4, we see that the scores converge to the deterministic score values. This is no surprise if we recall how the numerical simulations behaved in the previous chapter, mainly that as we increased system size we approached the deterministic solution.

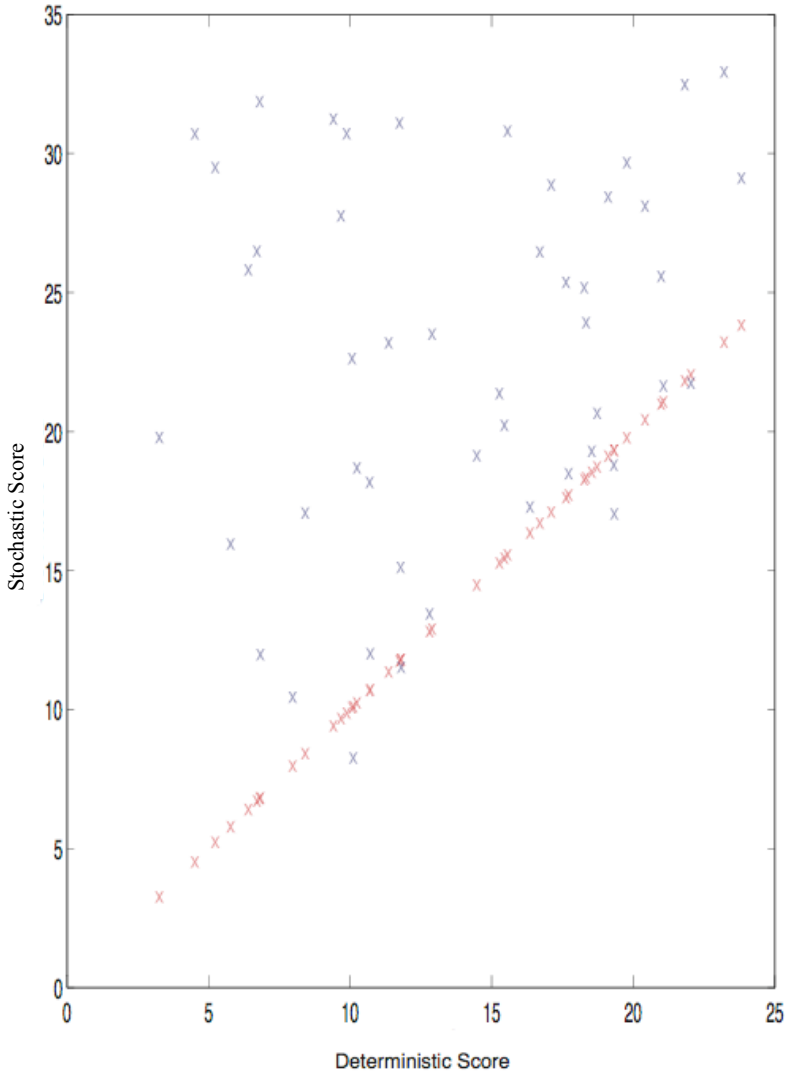

Figure 4.3: $\omega = 10000$, blue: stochastic scores, red: deterministic scores

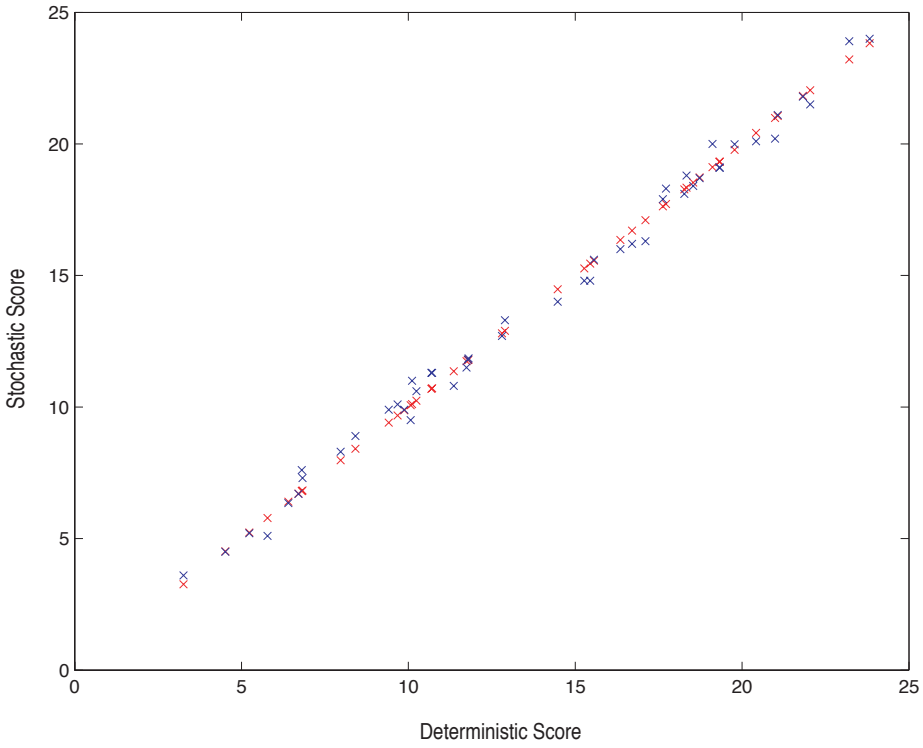

Figure 4.4: $\omega = 100000$, blue: stochastic scores, red: deterministic scores

The story in this chapter is not so exciting as the results show that the circadian rhythm is robust, but only for small molecular fluctuations. A further note could be made about the distribution of score values. Mainly that the introduction of biological noise did not significantly improve any of the score values in any significant way. This suggests that the noise in the system has not inadvertently provided us with a better set of parameter values.

CHAPTER V

Concluding Remarks

We have examined to what extent molecular noise affected the circadian rhythm of the *Arabidopsis thaliana* circadian clock. The results were not surprising but did pose some interesting questions to both biologists and modelers.

The initial circadian clock network was decomposed into elementary reaction steps and then modeled numerically using a Gillespie type algorithm. We chose not to breakdown the Hill functions into further steps as this would have increased the number of parameters in the model. In essence we kept the Michaelis-Menten kinetics in our reaction probabilities. From a modeling point of view this was deemed sensible, since the numbers of molecules we are dealing with are very low and Michaelis-Menten kinetics is more suited to dealing with much larger quantities. A further study may involve disposing of Michaelis-Menten kinetics and indeed breaking down the Hill functions into much more simpler reactions. Armed with this new model we may now be able to compare and contrast the results from both types of models. This may give us further insight into the unusual behavior seen in dark/dark cycles for reasonable system size values that we observed.

The score functions did not provide us with any more interesting insight into the robustness of the circadian clock with respect to biological noise. However it may be

interesting to see if the introduction of biological noise has opened up new areas of parameter space, in which we observe the desired circadian rhythm.

From both the discrete stochastic model and parameter variation model we can draw on two questions with which biologists may be able to help. Firstly it would be interesting to know the numbers of molecules involved in the chemical reactions in the circadian clock so that we have an actual limit to how low we can take the system size parameter in the model. Secondly some idea of how much the transcription and translation rates vary between cells as well as an idea of what these rates depend on, (i.e. the concentration of some nutrients within the cell), would provide us with more details about how to model such parameters. From a mathematical point of view it may be interesting to see how the circadian rhythm behaves when the transcription and translation rates follow a stochastic process themselves.

In this thesis we have ascertained that the circadian clock is robust to biological noise of various types; molecular noise due to small copy numbers and parameter variation, due to random changes of available nutrients for example. However we did notice that the rhythm was destroyed when the noise in both models was increased to large levels, which is of no real surprise. We see the stochastic methods used in this thesis as a tool to see how robust a gene regulatory network is to biological noise, from our results we can conclude the circadian clock network under scrutiny here is indeed robust. The next logical step would be to consider similar tests on later models of the circadian clock in *Arabidopsis thaliana*.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, fourth edition, 2002.

[2] A. Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of development pathway bifurcation in phage-λ-infected escherichia coli cells. *Genetics*, 149:1633–1648, 1998.

[3] N. Barkai and S. Leibler. Biological rhythms: Circadian clocks limited by noise. *Nature*, 403:267–268, 2000.

[4] W. J. Blake, M. Kaern, C. R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422:633–637, 2003.

[5] T. A. Brown. *Genomes 3*. Garland Science, New York, third edition, 2007.

[6] M. Carletti, K. Burrage, and P.M. Burrage. Numerical simulation of stochastic ordinary differential equations in biomathematical modelling. *Mathematics and Computers in Simulation*, 64:271–277, 2004.

[7] A. Colman-Lerner, A. Gordon, E. Serra, T. Chin, O. Resnekov, D. Endy, C. G. Pesce, and R. Brent. Regulated cell-to-cell variation in a cell-fate decision system. *Nature*, 437/29:699–706, 2005.

[8] A. Cornish-Bowden. *Fundementals of Enzyme Kinetics*. Portland Press, London, 1995.

[9] F. H. C. Crick. The central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[10] H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Jnl. Comp. Biology*, 9:67–103, 2002.

[11] J. C. Dunlap. Molecular bases for circadian clocks. *Cell*, 96:271–290, 1999.

[12] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297:1183–1186, 2002.

[13] C. W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer-Verlag, Berlin, 2004.

[14] D. T. Gillespie. A general method for numericaly simulating the stochastic time evolution of coupled chemical reactions. *Computational Physics*, 22:403–434, 1976.

[15] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Physical Chemistry*, 81:2340–2361, 1977.

[16] D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188:404–425, 1992.

[17] D. T. Gillespie. The chemical langevin equation. *Chemical Physics*, 113:297–306, 2000.

[18] J. H. Gillespie. *Stochastic Modelling for Systems Biology.* Oxford University Press, Oxford, 1991.

[19] D. Gonze, J. Halloy, and A. Goldbeter. Deterministic versus stochastic models for circadian rhythms. *Biological Physics*, 28:637–653, 2002.

[20] D. Gonze, J. Halloy, and A. Goldbeter. Robustness of circadian rhythms with respect to molecular noise. *PNAS*, 99:673–678, 2002.

[21] D. Gonze, J. Halloy, and A. Goldbeter. Emergence of coherent oscillations in stochastic models of circadian rhythms. *Physica A*, 342:221–233, 2004.

[22] D. Gonze, J. Halloy, and A. Goldbeter. Stochastic models for circadian oscillations: Emergence of a biological rhythm. *Quantum Chemistry*, 98:228–238, 2004.

[23] D. Gonze, J. Halloy, J.-C. Leloup, and A. Goldbeter. Stochastic models for circadian rhythms: Effect of molecular noise on periodic and chaotic behaviour. *C. R. Biologies*, 326:189–203, 2003.

[24] P. D. Gould, J. C. W. Locke, C. Larue, M. M. Southern, S. J. Davis, S. Hanano, R. Moyle, R. Milich, J. Putterill, A. J. Millar, and A. Hall. The molecular basis of temperature compensation in the arabidopsis circadian clock. *The Plant Cell*, 18:1177–1187, 2005.

[25] D. J. Higham and N. J. Higham. *MATLAB Guide.* SIAM, Philadelphia, 2000.

[26] J. R. Hubbard. *Programming with C++.* McGraw-Hill, 1996.

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

[28] T. B. Kepler and T. C. Elston. Stochasticity in transcriptional regulation: Origins, consequences and mathematical representations. *Biophysics*, 81:3116–3336, 2001.

[29] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations.* Springer-Verlag, Berlin, 1992.

[30] J.-C. Leloup, D. Gonze, and A. Goldbeter. Limit cycle models for circadian rhythms based on transcriptional regulation in drosophila and neurospora. *Biological Rhythms*, 14:433–448, 1999.

[31] J. C. W. Locke, A. J. Millar, and M. S. Turner. Modelling genetic networks with noisy and varied experimental data: the circadian clock in arabidopsis thaliana. *Theoretical Biology*, 234:383–393, 2005.

[32] J. C. W. Locke, A. J. Millar, M. S. Turner, M. M. Southern, L. Kozma-Bognr, V. Hibberd, and P. E. Brown. Extension of a genetic network model by iterative experimentation and mathematical analysis. *Molecular Systems Biology*, 2005 doi:10.1038/msb4100018.

[33] Berg O.G. A model for statistical fluctuations of protein numbers in a microbial-population. *Journal of Theoretical Biology*, 73:307, 1978.

[34] J. Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2:157–175, 2005.

[35] J. Paulsson and M. Ehrenberg. Noise in a minimal regulatory network: Plasmid copy number control. *Q. Rev. Biophysics*, 34:1–59, 2001.

[36] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C.* Cambridge University Press, Cambridge, 1988.

[37] D.R. Rigney. Note on the kinetics and stochasticity of induced protein synthesis as influenced by various models for messenger rna degradation. *Journal of Theoretical Biology*, 79:247–257, 1979.

[38] D.R. Rigney. Stochastic model of constituitive protein levels in growing and dividing bacterial cells. *Journal of Theoretical Biology*, 76:453–480, 1979.

[39] D.R. Rigney and Schieve W.C. Stochastic model of linear, continuous protein-synthesis in bacterial populations. *Journal of Theoretical Biology*, 69:761–766, 1977.

[40] A. Samach and P. A. Wigge. Ambient temperature perception in plants. *Current Opinion in Plant Biology*, 8:483–486, 2005.

[41] L. A. Segel, editor. *Biological Kinetics*. Cambridge University Press, Cambridge, 1991.

[42] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS*, 99:12795–12800, 2002.

[43] M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *PNAS*, 98:8614–8619, 2001.

[44] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam, 1983.

[45] G. von Dassow, E. Meir, E. M. Munro, and G. M. Odell. The segment polarity network is a robust developmental module. *Nature*, 406:188–192, 2000.

[46] G. von Dassow, E. Meir, E. M. Munro, and G. M. Odell. Robustness, flexibility, and the role of lateral inhibition in the neurogenic network. *Current Biology*, 12:778–786, 2002.

[47] G. von Dassow and G. M. Odell. Design and constraints of the drosophila segment polarity module: Robust spatial patterning emerges from intertwined cell state switches. *Jnl. Exp. Zoology*, 294:179–215, 2002.

[48] D. J. Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman and Hall, New York, 2006.