ANALYSIS OF A SIMPLE GENE EXPRESSION MODEL

SIMBARASHE CHIPINDIRWI

B.Sc. (Hons) Applied Mathematics, National University of Science and Technology, Zimbabwe, 2004

M.Sc. Mathematics, University of Zimbabwe, 2006

A Thesis

Submitted to the School of Graduate Studies

of the University of Lethbridge

in Partial Fulfillment of the

Requirements for the Degree

MASTER OF SCIENCE

Department of Chemistry and Biochemistry

University of Lethbridge

LETHBRIDGE, ALBERTA, CANADA

To my late parents.

Abstract

Gene expression is random owing to the low copy numbers of molecules in a living cell and the best way to study it is by use of a stochastic method, specifically the chemical master equation. The method is used here to derive analytically the invariant probability distributions, and expressions for the moments and noise strength for a simple gene model without feedback. Sensitivity analysis, emphasizing particularly the dependence of the probability distributions, the moments, and noise strength is carried out using Metabolic Control Analysis, which uses control coefficients that measure the response of observables when parameters change. Bifurcation analysis is also carried out. The results show that the number of mRNA molecules follows a hypergeometric probability distribution, and that noise decreases as the number of these molecules increases. Metabolic Control Analysis was successfully extended to genetic control mechanisms, with the obtained control coefficients satisfying a summation theorem. The system undergoes stochastic bifurcations as parameters change.

Preface

I should begin by telling you about the menacing shortcomings in me, when it comes to Biology. Before joining Professor Roussel's lab, I had some erratic encounters with Biology, but that was long back. If I remember well, the last time I was enrolled in a Biology class was in high school, and the only word I could remember when I came to Canada was "photosynthesis". I am not sure why that word kept lingering in my mind. I think that's the truth about life, the word "educated" can remain with us but, as is the case with most people, most of what we would have learnt never stays with us.

I appreciated Marc's work when I browsed through his website. Application of Mathematics to a variety of fields, among them Biology and Chemistry. My training previously centred in Mathematics, Applied Mathematics to be precise. Most Mathematicians would agree with me that in this field, most of the experts move around with void minds. Definitions and facts are all they need to prove, derive, or extend a theory, support some hypothesis or deduce a solution. The need to recall or memorize is minimal, in contrast to Biology. Most of the Biology textbooks are a pain to study for people who like proofs, let alone those who like playing around with figures since, except for the page numbers, they barely have numbers in them. Be that as it may, the interaction between Mathematics and Biology is inevitable, and that forced me to compromise and accept ambidexterity, so I studied the Biology necessary for this work. Apart from that, I am a long way behind.

My apologies to whoever is going to read this piece of work, and in particular, the Biologists. My Biology background is very shallow, but nonetheless, the biology facts engraved in this work suffice to set the base for my work. I also boast of a superb supervisory committee, a meticulously chosen concoction of experts, among them a Biochemist, hence it was safe to embark on this particular project.

Acknowledgments

Firstly, I would like to thank my supervisor, Professor Marc R. Roussel for a job well done in introducing and leading me throughout this work. He did this with the assistance of two other very important experts, Dr. Saurya Das and Dr. Ute Kothe, I also thank them very much. Dr. Kothe is the Biochemist I mentioned before, she started with me from those basic twenty amino acids and she knows very well my battles with biology. This trio was actually my supervisory committee and it was a great experience to work with them. I also thank the staff in the University of Lethbridge, and most importantly, all the people in the Department of Chemistry and Biochemistry for all the support and help during my studies. Lastly, I would like to thank my employer, the National University of Science and Technology (NUST), in Zimbabwe, for allowing me the chance to venture outside campus in search of knowledge.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cells perform crucial roles in living organisms. Amongst these are decision making, memory storage and integration of signals. What surprises is that cells manage to execute their duties in environments characterized by noise. They operate precisely within strict and safe parameters despite this noise. Noise causes unpredictability of events in cells, which then respond to changes to maintain order to behavior and development. The strength of the noise or randomness is termed stochasticity. Perhaps we henceforth need to brace ourselves for some of the challenging questions inevitable in biology: What is stochasticity? Where does it originate from? What is the evidence of its existence in biological systems? How does it affect the modeling of biochemical processes? Most importantly, can we control stochasticity?

I will attempt to answer the preceding questions one at a time and with reference to a central biochemical process, gene expression. Gene expression is not an arbitrary choice; this work focuses on it for two reasons. Firstly, it would be easier to comprehend stochasticity by employing some sort of a common case study of it, than just an abstract discussion. Secondly and most importantly, gene expression forms the basis for the control of many cellular processes [1].

Therefore, I will try and provide some answers to the above questions with all the attention concentrated on gene expression. In contrast, my work does not revolve around a particular gene, but is applicable to a multitude of genes.

1.1 Gene Expression

In simple terms, a gene is a piece of the DNA which encodes RNAs for different uses in the cell, including the manufacture of specific proteins. The DNA stores all the genetic information in a cell. The process by which this information is retrieved and converted to "gene products" is what is termed gene expression. Although genes are randomly expressed, the expression is triggered. First, signals have to be sent to the gene, then an appropriate response is generated due to the regulation of that particular gene. The signals are normally changes in levels of biomolecules, or changes in conditions in and around the cell. Gene regulation is carried out by proteins and/or small regulatory RNA's, e.g the microRNAs, which act in response to the physiochemical environment. This could be the absence of a particular biochemical species, changes in temperature, etc. Proteins are synthesized using instructions carried by messenger RNA (mRNA). The mRNA or messenger ribonucleic acid is a short, single stranded copy of the DNA which has the same sequence as the coding strand with the base uracil in place of the base thymine. Thus, information flows from the gene to the mRNA and eventually to the proteins—the so–called Central Dogma of Molecular Biology. The flow of information to the mRNA is called transcription, while the conversion of the mRNA to proteins is called translation.

1.2 Transcription

Information stored by the gene will only be released if the gene has received a signal and is ready. Some genes are always ready for expression. These normally have the role of directing a constant and incessant supply of crucial cell commodities, for example, energy. On the other hand, there are genes that are switched "off" until the need for them to be switched "on" arises [2–6]. The reader will notice in this work that it is this second type

of gene that we have put much focus on, although our model also covers a constitutive gene once the rate at which the gene switches off takes very small values. For these genes, transitions between the "on" and "off" states are random. This is caused by sudden unpredictable or random binding/unbinding activities between the gene and the transcription initiation factors. There are other processes that can activate or inactivate the gene, for example chromatin formation.

Upstream of the gene is a non–coding piece of the DNA called the promoter. It serves the purpose of providing sites to which biomolecules bind, amongst them, the Transcription Factors (TFs) and the RNA polymerase (RNAp) [2, 7, 8]. The former enhance or inhibit the chances of the latter, the enzyme that catalyzes transcription, to bind to the transcription start site. TFs bind to some special regions of the promoter called the operator sites. There are also other factors that bind to the promoter and block or inhibit transcription, thus deactivating the gene [1, 6]. It is only after a successful binding of the TFs to the operator sites that the gene becomes ready to be transcribed. The interesting part about the binding of the TFs to the operator sites is that it is reversible. The reader will realize in the following sections that our model puts these reversible binding/unbinding processes in the forefront of the study, and so do the authors of a number of other studies [2, 8–10]. In [2], the author shows that the TF binding/unbinding mechanism is at the helm of most of the stochasticity, while [10] defines the origin of the fluctuations in mRNA distributions as being the transitions in promoter states. The author of [10] takes time to illustrate how stochastic effects easily spread to other processes, after starting in one process. We have also emphasized gene activation/deactivation due to the same fact that the randomness emanating from these reversible processes is inherited throughout the subsequent processes of transcription and translation. In gene expression, the gene switches with bias and it is not surprising at all that at any given time, the gene is more likely to be in one state than in the

other.

1.3 Translation

After the gene has been transcribed, the product mRNA is later decorded and the information it carries is then used to form proteins, through the process of translation. Proteins are formed by concatenation of amino acids through peptide bonds. A special type of RNA called transfer RNA (tRNA) is responsible for bringing the residues to the ribosome which links these residues through condensation reactions. The tRNA is just a short RNA molecule. It binds to a specific codon (base triplet) on the mRNA, using a complementary sequence called the anticodon. The mRNA molecules serve as templates from which proteins are produced (translated) by the ribosome, which monitors the codon–anticodon base pairings, as well as consolidating the peptide linkages. Although this is an important process, translation is not covered in this Thesis.

1.4 Stochasticity

Stochasticity is an effect or "force" that has an untimely and random power, which prevents the exact prediction of the later state of a system based on its previous standing. The only possible way to try and foretell the evolution of the state of the system affected is by utilization of probabilities. In fact, a system that evolves stochastically is best monitored by observing the changes in the probability distributions describing it, as time or conditions change. This is not easy, hence often some quantitative measures (moments) corresponding to the distributions, for instance the mean and the variance, are used. In the literature [5], the word noise has been used. Noise is a measure of the strength of randomness (stochasticity). Therefore, the "bias" that we acknowledged before in the switching characteristics of the

gene is a bias of the stochastic process. One cannot tell whether the gene will be active ("on") or inactive ("off") as time evolves, although conditions in a cell may bias the state toward one or the other outcome.

1.5 Origins of stochasticity in biochemical systems

In a chemical system, reactants are in constant motion (diffusion) and collisions are constantly taking place, some leading to chemical reactions, and others not. Collisions can also result in changes in kinetic energies of reactants, leading to randomness in individual reaction events. This is worse if the number of molecules is limited [7–9, 11–13]. As an example, genes normally appear in at most two copies, whereas the TFs usually occur in tens in a single cell [14]. It is in such cases that stochasticity becomes dominant, for the participation of each and every molecule is crucial. Individual reactions change the number of molecules by small integer amounts. In small populations, the disappearance of a single reactant molecule represents a large relative change in the population and thus in the reaction probabilities.

1.6 Evidence of stochasticity in biological systems

Suppose one monitors the number of molecules of a particular gene product. It is true that with time, the number of the same molecule varies. Similarly, if one observes the distribution of the same type of a gene product in two different cells, one would also observe a difference in the number of molecules. Genetically identical (clonal) cells may have a different number of molecules for the same gene product [9, 15]. Stochasticity has been pointed to as the cause of variability in cells. Stochasticity is significant in signalling pathways [16], and is responsible for phenomena such as the synchronization of populations of

cells [17]. It affects regulation of cells resulting in the observed diversification in phenotypes [10,18–20]. Fluctuations in the levels of intermediary metabolites caused by stochasticity have been cited as the explanation for the retardation in cell growth and reduction in metabolic efficiency [18]. The observed loss of synchrony in circadian clocks [21, 22], is also caused by noise. On the good side of it, noise helps lambda phage in the stochastic lysis/lysogeny stage switches [24], while *E. coli* benefits by utilizing it to run and tumble as a strategy for finding nutrient sources [25]. Yeast cells transition among multiple phenotypes, stochastically, which keeps their fitness at an optimum [26]. Regulation of gene expression is instrumental in some of these processes.

1.7 Stochasticity and the modeling of biological systems

Random events in cells occur by chance, i.e., on the molecular scale, no event is certain or occurs almost surely. The result is that we can only study the affected systems through probability distributions. Involvement of stochasticity changes the face of the modeling strategy entirely. Deterministic approaches become inapplicable, given the small copy numbers of molecules. Instead, qualitative distinctions in outcomes arise due to randomly timed events, necessitating the use of stochastic models. The study of these models could focus on particular aspects, for instance, the qualitative behavioral changes (stochastic bifurcations) or the response (sensitivity) in levels of biomolecules of the system as conditions in and around it change. Events are treated as memoryless, that is, they are not dependent on their previous standing, but only the current state will determine the next state. They thus belong to a class of Markov processes [31]. In these processes, transitions from one state to the other are considered, allowing one event (reaction) to occur at a time and changing the molecule numbers by discrete stoichiometric quantities. The numbers of molecules are normally used to keep track of the evolution of the system.

Pertaining to the gene expression process, we consider two different probabilities: the probability that the system is in a given state at time t, and the transition probability that the system moves from one state to another in time dt. The system thus has a probability of exiting a state and reverting back to the state and the difference between these probabilities would give us the rate of change of the probability of being in a certain state. This probability rate equation is called the Chemical Master Equation (CME) [4, 9, 13, 27–31]. Focusing on the gene and its products, the mRNA and proteins, we can derive an infinite number of differential equations. These would stand for the two–state phenomenon on the gene itself and the number of molecules of each product in each state of the gene. The full set of these joint probability rate equations gives us the CME.

1.8 Controlling stochasticity

The first step in alleviating or utilizing a phenomenon in a system is having some command over it as it unfolds. Despite the randomness of events inside them, cells function robustly. How can we explain this robustness? This insight is crucial to understand real organisms. It also assists in the manufacture of gene networks that are applicable in biotechnology and therapy.

Now every reaction rate depends on a rate constant whose values can change as conditions are altered. The key to controlling a reaction rate is therefore knowing the rate constants with the greatest effect on some particular system output, for example noise. A system may not change much (insensitivity or robustness) when certain parameters change or may change significantly (sensitive). Ultimately, controlling a system is all about manipulating system parameters and assessing its response after some perturbation. The study of the response of a system as parameters are changed is called sensitivity analysis [32]. We def-

initely need to know which parameters govern which distribution to carry out sensitivity analysis, considering that many biochemical processes and hence many parameters are involved.

Biological systems and processes governing them are intricate; the outcome is dependent on many biochemical processes or reactions. In fact, the control of the production or regulation of metabolites is shared amongst many reactions, with many different rate constants. This is the biggest challenge in the endeavour to study them. Analytically, one faces the difficulty of not only many but also long expressions which require the use of advanced mathematical principles to obtain them. In simulations many trials are required to get reasonable results.

To this end, I have introduced the gist of the problem in everyday language, but the real analysis will be backed by an exact mathematical framework. I will therefore take some time and space in this work to introduce the reader to some of the fundamental mathematical principles at the backbone of this work in sections below.

1.9 Probability

Since we have adopted a probabilistic description of all the events occurring in our system, it is necessary that we give some background about probability. In this context, it is more convenient that we refer to some experiment with many likely outcomes or possibilities. Any attempt of an outcome is called a trial. The outcome of a trial can be any out of a pool of possible outcomes, or the sample space. This means that its choice is random, hence we can call the outcome of a trial, X, a random variable. X can come up a number of times in a random manner. The fraction of times it comes up, or the frequency, should give an idea of how often to expect it. This is called the probability distribution of X, $P(X)$. Probability

is bounded, meaning that there is a maximum value for the likelihood of an event, and at this maximum, the event occurs with certainty or almost surely. There is also a minimum likelihood, in which case the event should not be expected at all. The values standing for these likelihoods are 1 and 0 respectively. All the probabilities will be sandwiched between these values. In mathematical notation, we can write,

$$0 \leq P(X) \leq 1.$$

The sum of all probabilities for outcomes in a sample space is one, i.e.,

$$\sum_{\text{all x}} P(X) = 1.$$

Section 2 of Gillespie's paper [33], treats the concept of probability so well that it is worth reading if one wants some more detail. There are two quantities that occur randomly in our case, the state of the gene, s, and the number of molecules of the gene product, m, both observed at time t. The state space therefore is the set of all possible pairs, $\Omega = \{s, m\}$, where s takes the values 0 or 1, corresponding to the "off" and "on" states of the gene, and m can take any value from 0 to ∞, i.e., $m \in [0, \infty)$. If X assumes only integer values, i.e is discrete, the resulting distribution is called a discrete probability distribution. The state space we have just described is hence a discrete space. Otherwise X would have a continuous probability distribution. An example of a discrete probability distribution is the hypergeometric distribution, which defines the number of successes obtained when one conducts a sequence, n, of draws, without replacement, from a population with a finite number of elements. The random variable X is itself called the discrete random variable, and is the function X defined on the probability space $(\Omega, 2^{\Omega}, P)$. X has the purpose of selecting elements of Ω, which are in the form of ordered pairs (s, m). The set of all

subsets of the state space Ω, is given as 2^{Ω}.

1.10 Joint probability distribution

Suppose on the same probability space, $(\Omega, 2^{\Omega}, P)$, we pick two random variables s and m. A joint probability for these two events can be described as a probability of events including both quantities s and m.

1.11 Conditional probability

The conditional probability between two events, say s and m, is defined by

$$P(m|s) = \frac{P(m \text{ and } s)}{P(s)},$$

where $P(m \text{ and } s)$ stands for the probability of all the outcomes that include both events.

1.12 Measures of central tendency and variability

One of the crucial roles played by statistical analyses is the characterization of **location** and **variability** of a data set. As time progresses, different numbers of mRNA molecules are observed owing to stochasticity. A collection of these counts will form data sets, each set having different measures of **central tendency** (mean) and **dispersion** (variance) which respectively measure the **location** and **variability** of the given data set. The shape of a particular distribution is characterized by the skewness and the kurtosis. The skewness measures the asymmetry in the distribution of the values of m whereas the kurtosis measures the peakedness of the data. In statistics, the most convenient way to characterize the

shape of a distribution is by use of a histogram, a graphical technique that can display both the skewness and the kurtosis, and the mean and the standard deviation.

1.13 The Chemical Master Equation (CME)

In physics, a master equation is a set of differential equations which describes the evolution of the probability distribution at a given time as the balance between transitions leading to each state and transitions removing the system from those states. A central question we should ask is how does $P(s,m)$ change with time? In other words, how does it evolve?

We begin discussing the CME by acknowledging the work done by Daniel T. Gillespie. The reader is hereby encouraged to see [33–36] for a detailed coverage of the CME. Mathematically, we take into account a system with M species, and N reactions occurring independently in this system. The system is considered spatially homogeneous, i.e. well-stirred, and is held at fixed temperature and volume. Let $P(\underline{x},t)$ be the probability of being in a state $\underline{x}$ at time t, where $\underline{x} \in \Omega$. If we consider a particular reaction $R_\mu : \Omega \to \Omega$, where $\mu \in \{1,2,3,...,N\}$, we can define a corresponding propensity function $a_\mu : \Omega \to \mathbb{R}^+$ and a corresponding stoichiometric transition vector ν_μ, a fixed constant vector. Now the CME gives the change in the probability of being in a state $\underline{x} \in \Omega$ as being a sum of all the probabilities of coming to $\underline{x}$ through R_μ in time dt, which is $a_\mu(\underline{x}-\nu_\mu)P(\underline{x}-\nu_\mu,t)$, less the sum of all the probabilities of going out of X through R_μ, given by $a_\mu(\underline{x})P(\underline{x},t)$. Thus mathematically, we can write the CME as

$$\frac{\partial P(\underline{x},t)}{\partial t} = \sum_{\mu=1}^{N} a_\mu(\underline{x}-\nu_\mu)P(\underline{x}-\nu_\mu,t) - \sum_{\mu=1}^{N} a_\mu(\underline{x})P(\underline{x},t). \tag{1.1}$$

1.14 Metabolic Control Analysis

The word control refers to adjustments made on a system's output. Verbally, it means the power or ability to give some direction, to stop or to start something. A distinction should be made between this word and the word regulation. Regulation refers to the system's means or mechanism to preserve variables at some constant levels with time, regardless of the settings of the external environment. These variables can, for example be temperature, concentrations, pH, etc. Metabolic control is hence the ability to inflict change(s) to a state of metabolism when the external environment changes. One can thus measure the change by utilizing the response by the metabolism to changes in external factors. In a quest to explain how cells can operate within strict and safe parameters in the presence of noise and how cells can use noise, for instance to amplify certain responses, it is best to use sensitivity analysis. One form of sensitivity analysis already in use in analysing biochemical processes is Metabolic Control Analysis (MCA) [37]. It is based on control coefficients which enable assessment of relative changes in any observable when parameters change.

My work follows a path that has been cleared by a number of researchers [38]. What I only do in this study is to extend, using already existing tools that were separately developed, the application of these tools to gene expression. As we will realize in the following section, studies have been going on, including research on gene expression itself and MCA, although not even one of those that I know has taken the route I took–applying MCA to gene expression.

1.15 Background knowledge

I mentioned before that biochemical systems in which participant molecule numbers are limited are subject to stochasticity, and that gene expression was no exception. Events in

these systems occur at random and unpredictably. Maintenance of the behavior of individual cells, and the system at large, plays a big role in making sure that the system remains robust to changes within itself and in its environment, as well as even being sensitive to these changes. So how do cells work normally given the unpredictability of reactions taking place inside them? Questions such as this have triggered studies in the regulation and control of biochemical systems [45–49].

Regulation refers to the ability of the system to maintain itself, whereas control involves external interference with the cellular activities or chemical reactions within the system to effect changes on the system's performance. Studies in control theory have been going on for over 40 years now [50–52]. In 1973 Henrik Kacser and Jim Burns [53] in Edinburgh proposed theoretical frameworks meant to deal with the regulation of metabolic systems. A year later, Reinhart Heinrich and Tom Rapoport in Europe embarked on similar studies [54, 55]. Despite the fact that these studies were independent and named differently, they both worked on the system's sensitivity to parameters, applied at steady state. Heinrich and Rapoport called their study Biochemical Systems Theory, while Kacser and Burns decided to name theirs Metabolic Control Analysis (MCA) [54–56].

Traditionally, a rate-limiting step is defined as the slowest step in a pathway. However, the meaning of "rate-limiting step" should actually be that step in a pathway which produces the greatest change in the overall flux, once its rate is increased [51]. The main tools employed by MCA are the control coefficients, which measure quantitatively the response in observables when parameters change. A greater absolute value for a control coefficient would then mean greater control [37]. Such quantification enables assessment of the effectiveness of a particular step, relative to another. Thus, if control coefficients are the same for several steps, one cannot speak of a single "rate-limiting step", instead one should conclude that the control over the overall flux is shared or distributed and existence of a

"rate-limiting step" becomes some sort of a hypothesis to be verified. It seems there is some slow reform evolving amongst biologists to a better definition though. For example a discussion about control coefficients and MCA was presented in a book by David Metzler [58].

My work is an application of MCA to the stochastic gene expression process. The simple stochastic model I use for this purpose, although it is just a caricature which lacks crucial features exhibited by a real gene expression system, makes the analysis easy. The results, presented in the form of mathematical expressions as well as their graphical representations, should form a convenient and reusable toolbox for the analysis of more complex and more realistic models. Extension of MCA to genetic control mechanisms, and most importantly, the development of stochastic control coefficients for the moments of the distributions, is a unique piece of work. The analytical approach has been used before to study similar models describing the gene expression process [9, 28–30, 63, 68], but these researches have ended on developing the expressions for the probability distributions and some of the first four moments. Usually it is only the first two moments that are derived, the mean and the variance. Paolo Visco and coworkers [4], used a similar analytical approach to study a DNA–inversion genetic switch model. MCA has been extended to the general stochastic regime (for e.g. in [59]), where an outline of the development of stochastic control coefficients for variances and covariances of concentrations and fluxes is given. These statistics measure the noise or strength of the stochasticity in a system, and are derived from the moments of the distribution. To date, no one has ever worked on developing stochastic control coefficients for the moments of the distributions in gene expression. As mentioned earlier, moments serve to capture the key qualities of a distribution in numerical form. Knowing how these important quantities respond to changes in parameters allows us to have a closer look at how the distributions themselves, and the system as a whole, respond

also. Moreover, the summation theorems governing the control coefficients [37, 60, 61], enable one to link the behavior of the system variables to the kinetics of the individual enzyme catalysed reactions.

However, the use of parametric sensitivity to study the dynamical behavior of systems is only applicable when small changes in parameters are effected. In principle, the system is allowed to settle to a steady state before a small disturbance is inflicted to see if it reverts back to the original steady state. Large perturbations are not covered by MCA, although research has started which aims to extend the scope of MCA to large parameter changes. Vassily Hatzimanikatis [64] proposed a nonlinear MCA model which allows determination of control coefficients as functions of the metabolic parameters and aimed at enabling the use of MCA in large parameter changes. The work of Small and Kacser [65,66], which considers linear kinetics is another attempt to extend MCA to cover large parameter changes. We can also study the behavior of a system in response to large changes in its parameters by using other methods that are not related to MCA, for example, bifurcation analysis. I use both of these methods, MCA and bifurcation analysis, to study gene expression so as to give the reader a complete picture of how the system behaves in the cases of both small and large parameter changes. The outline of my work is as follows.

In Chapter 2, I present a description of the simple gene expression model which I used, and its analysis. Chapter 3 contains the development of the stochastic control coefficients, their interpretation and an illustration of the fulfilment of one of the summation theorems. Bifurcation analysis together with a study of the same model using simulation is done in Chapter 4. The simulation is done to enable a comparison of my results with other researchers' work on a similar model. Lastly, in Chapter 5, I discuss my findings, present some concluding remarks and also give some future direction of this work.

Chapter 2

The Simple Gene Expression Model

2.1 Determining the distributions of biomolecules

It would be interesting to obtain expressions for the probabilities, in terms of the parameters. However, this involves solving the CME, which is an infinite set of differential equations. Real systems are represented by large and more complex models but it is hard to analytically solve the resultant CMEs. This is the reason why the model used in this study is simple, and represents those genes which switch on and off. Larger and more realistic models are studied through simulations but, unfortunately, one may not obtain expressions for the probability functions from this method. Analytical methods tend to enable researchers to focus on the content of a model, as opposed to its implementation through a simulation.

Thus, we consider and develop analytic solutions to the simple constitutive gene expression model depicted in a diagram below. Compared to simulation, there are a number of advantages that we get by using an exact approach. For instance, knowledge about model parameters, i.e. the biochemical rate constants of the underlying biochemical reactions, is something quite rare. In analytical methods, expressions for the probability distributions are obtained in terms of the parameters. Thus, an analytical approach leads to equations and allows one to be able to select parameters and try them easily and quickly. Unlike simulations, equations are reusable. Other researchers can use the already available equations by merely putting in the parameters appropriate to their situation of interest. Also, it is difficult to get derivatives which are required to obtain control coefficients by using simulation results.

Using the analytical approach, we calculate the stationary probability distributions of the biomolecules (reactions) constituting gene expression. This means that the probability distributions do not depend on time. Studying the stationary distributions enables us to only study the changes on these distributions due to changes in parameters, thereby eliminating the need to consider that these distributions change as time changes. The aim is to understand the complications of gene expression profoundly but through a simple model. A simple model provides one with insight into larger and more realistic models. This build-up process allows assessment of the effects of changing biochemical rates on the distributions of crucial (basic) biomolecules without having to worry about complications arising as more biochemical players add in.

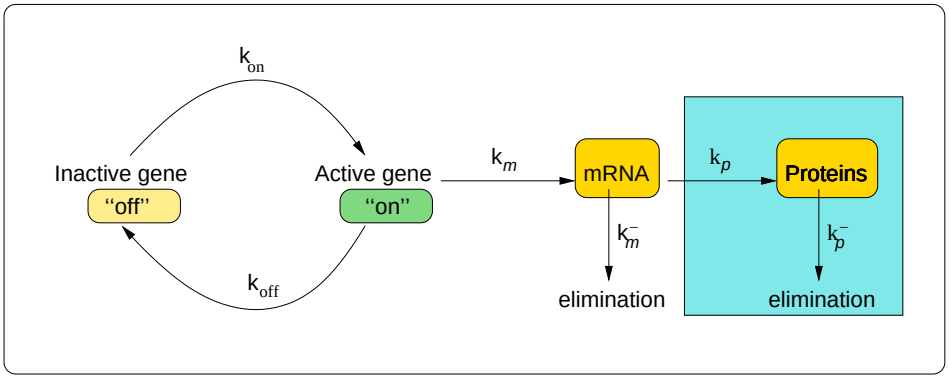

Figure 2.1: The model diagram

In this diagram, the constants k_{on}, k_{off}, k_m and k_p denote rate constants for gene activation, gene deactivation, mRNA and protein production respectively, whereas k_m^- and k_p^- respectively indicate rate constants for mRNA and protein elimination by degradation into constituents, covalent modification or use to form more complex molecules. We begin the analysis by outlining some assumptions of the model.

2.2 Assumptions of the model

1. mRNA is produced only when the gene is active.

2. mRNA production is not regulated, that is there is no feedback. This means that proteins produced in the translation process will not affect gene expression. Also, all the rates mentioned above are constant, i.e. they do not depend on any time-varying or concentration-dependent quantities.

3. The probability of two reactions occurring simultaneously is infinitesimal, a basic assumption of the master equation modeling [33].

2.3 Objectives of the study

The objectives of this study are:

1. To derive the joint probability distribution for the state of the gene and the number of mRNA molecules, i.e. $P(s,m)$ and hence the marginal probability distribution

$$P(m) = \sum_{s=0}^{1} P(s,m).$$

2. To derive the conditional probability distributions $P(m|s)$.

3. To derive the expressions for the moments for both the marginal and the conditional probabilities.

4. To derive the expressions for the control coefficients for the moments mentioned above. We also determine if these coefficients fulfill some summation theorems.

5. To study the stochastic bifurcations that the system undergoes when parameters are changed.

6. To study the same model using numerical simulations and compare the results.

2.4 Analysis of the Model

We get two chemical master equations for the evolution of mRNA molecules, depending on the state of the gene as shown below:

$$\frac{dP(1,m,t)}{dt} = k_{\text{on}}P(0,m,t) - (k_{\text{off}} + k_m + k_m^- m)P(1,m,t)$$

$$+k_m^-(m+1)P(1,m+1,t) + k_m P(1,m-1,t), \;\; m \geq 1 \tag{2.1}$$

and

$$\frac{dP(0,m,t)}{dt} = k_{\text{off}}P(1,m,t) + k_m^-(m+1)P(0,m+1,t)$$

$$-(k_{\text{on}} + k_m^- m)P(0,m,t), \;\; m \geq 0. \tag{2.2}$$

$P(1,m,t)$ and $P(0,m,t)$ are the probability distributions of the mRNA molecules when the gene is "on" and "off" respectively. Eqs. (2.1) and (2.2) are infinite sets of ordinary differential equations since m can take any nonnegative value. A special case of Eq. (2.1) arises when the gene is "on" and there are no mRNA molecules. This is tantamount to

taking $m = 0$ in Eq. (2.1), which gives

$$\frac{dP(1,0,t)}{dt} = k_{\text{on}}P(0,0,t) + k_m^- P(1,1,t) - (k_{\text{off}} + k_m)P(1,0,t),$$

considering the boundary conditions $P(s,-1,t) = 0$. Assuming stationarity, i.e. probabilities not dependent on time, we obtain

$$P(0,m) = \frac{1}{k_{\text{on}}}[(k_{\text{off}} + k_m + k_m^- m)P(1,m) - k_m^-(m+1)P(1,m+1) - k_m P(1,m-1)], \quad (2.3)$$

from (2.1), where we have dropped the variable t since we aim to obtain the stationary probability distributions. Similarly we obtain the special case

$$P(0,0) = \frac{1}{k_{\text{on}}}[(k_{\text{off}} + k_m)P(1,0) - k_m^- P(1,1)]. \quad (2.4)$$

Eq. (2.3) gives $P(0,m)$, and hence $P(0,m+1)$ in terms of $P(1,\cdot)$, which enables us to eliminate $P(0,\cdot)$ from the problem of solving for the stationary states. Substituting our $P(0,m)$, and $P(0,m+1)$ in Eq. (2.2) gives

$$[k_m(k_{\text{on}} + k_m^-) + k_m^-(2k_m + k_{\text{on}} + k_{\text{off}})m + (k_m^-)^2 m^2]P(1,m)$$

$$-[k_m^-(k_{\text{on}} + k_m + k_{\text{off}} + k_m^-) + 2(k_m^-)^2 m](m+1)P(1,m+1)$$

$$+(k_m^-)^2(m+1)(m+2)P(1,m+2) - k_m(k_{\text{on}} + k_m^- m)P(1,m-1) = 0, \quad m \geq 1. \quad (2.5)$$

Equation (2.5) can be solved for $P(1,m)$ by first defining, respectively for the "on" and "off" states, the moment generating functions

$$F(s,z) = \sum_{m=0}^{\infty} z^m P(s,m), |z| \leq 1. \tag{2.6}$$

We then multiply equation Eq. (2.5) throughout by z^m and sum over m. The result is the second-order differential equation

$$\gamma(1-z)F''(z) + (\beta z - 1)F'(z) + \eta F(z) = 0, \tag{2.7}$$

where,

$$\gamma = \frac{k_m^-}{k_{\text{on}} + k_{\text{off}} + k_m + k_m^-},\ \beta = \gamma\alpha \text{ and } \eta = \frac{\gamma\alpha(k_{\text{on}} + k_m^-)}{k_m^-},\ \text{where } \alpha = \frac{k_m}{k_m^-}.$$

Note that in Eq. (2.7) we have just dropped the first argument of $F(1,z)$ for ease of presentation. It is easy to show that Eq. (2.7), after a change of variables, transforms to Kummer's differential equation

$$u_z F''(u_z) + (\psi - u_z)F'(u_z) - \varepsilon F(u_z) = 0, \tag{2.8}$$

where

$$u_z = \alpha(z-1),\ \psi = 1 + \frac{k_{\text{on}} + k_{\text{off}}}{k_m^-} \text{ and } \varepsilon = 1 + \frac{k_{\text{on}}}{k_m^-}.$$

2.5 Solutions of Kummer's differential equation

Eq. (2.8) has the form of Kummer's differential equation and has two independent solutions [31], which we denote as ${}_1F_1(\varepsilon;\psi;u_z)$ and $U(\varepsilon;\psi;u_z)$ with

$$ {}_1F_1(\varepsilon;\psi;u_z) = \sum_{n=0}^{\infty} \frac{1}{n!}\frac{(\varepsilon)_n}{(\psi)_n} u_z^n $$

and

$$ U(\varepsilon;\psi;u_z) = \frac{\pi}{\sin(\pi\psi)}\left(\frac{{}_1F_1(\varepsilon;\psi;u_z)}{\Gamma(1+\varepsilon-\psi)\Gamma(\psi)} - u_z^{1-\psi}\frac{{}_1F_1(1+\varepsilon-\psi;2-\psi;u_z)}{\Gamma(\varepsilon)\Gamma(2-\psi)}\right). $$

${}_1F_1(\varepsilon;\psi;u_z)$ is the confluent hypergeometric function. $(\zeta)_n$ denotes the Pochhammer symbol or the rising factorial and is given by

$$ (\zeta)_n = \zeta(\zeta+1)...(\zeta+n-1) = \frac{\Gamma(\zeta+n)}{\Gamma(\zeta)}, (\zeta)_0 = 1. \tag{2.9} $$

Therefore, the general solution of Eq. (2.8) is $F(u_z) = \mu\ {}_1F_1(\varepsilon;\psi;u_z) + k\ U(\varepsilon;\psi;u_z)$, where μ and k are constant coefficients. However, $U(\varepsilon;\psi;\alpha(z-1))$ is unbounded near $z=1$, yet $F(1,z)$ should satisfy $\lim_{z\to 1} F(1,z) \leq 1$ according to Eq. (2.6). These conditions ensure that $k=0$. Hence we find that

$$ F(1,z) = \mu \sum_{n=0}^{\infty} \frac{1}{n!}\frac{(\varepsilon)_n}{(\psi)_n} u_z^n = \mu\ {}_1F_1(\varepsilon;\psi;u_z). \tag{2.10} $$

Using Eqs. (2.3), (2.6) and taking, as noted in [39],

$$ \frac{d}{dz}\ {}_1F_1(\varepsilon;\psi;u_z) = \alpha\ \frac{\varepsilon}{\psi}\ {}_1F_1(\varepsilon+1;\psi+1;u_z), \tag{2.11} $$

we have

$$F(0,z) = \frac{\mu\,\alpha}{\varepsilon-1}\left[\left(\frac{\psi-\varepsilon}{\alpha}+(1-z)\right)\,{}_1F_1(\varepsilon;\psi;u_z)-\frac{\varepsilon}{\psi}(1-z)\,{}_1F_1(\varepsilon+1;\psi+1;u_z)\right]$$

$$= \frac{\alpha}{\psi-1}\left[\left(\frac{\psi-\varepsilon}{\alpha}+1-z\right)\,{}_1F_1(\varepsilon;\psi;u_z)-\frac{\varepsilon}{\psi}(1-z)\,{}_1F_1(\varepsilon+1;\psi+1;u_z)\right], \quad (2.12)$$

after substituting the value of μ calculated below. The constant μ can easily be obtained from the normalization condition, $F(1,1)+F(0,1)=1$. Following our steps, we conclude from Eqs. (2.10), (2.12) and the normalization condition that

$$\mu = \frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}} = \frac{\varepsilon-1}{\psi-1}.$$

Considering the state $S(t)$ of the promoter, the master equation involving the probability rate for the evolution of the "on" state, $\frac{dP(1,\,t)}{dt}$, can be written as

$$\frac{dP(1,t)}{dt}+(k_{\text{on}}+k_{\text{off}})P(1,t)=k_{\text{on}},$$

which, when solved, gives

$$P(1,t)=\frac{\varepsilon-1}{\psi-1}[1-\exp(-(k_{\text{on}}+k_{\text{off}})t)].$$

Hence, we observe that

$$\frac{\varepsilon-1}{\psi-1}=\lim_{t\to\infty}P(1,t). \quad (2.13)$$

2.6 Expressions for the probability distributions

One can obtain the probabilities as stated in [29], since from Eq. (2.6) one obtains that

$$\begin{aligned} P(1,1) &= \frac{d}{dz}F(1,z)\bigg|_{z=0}, \\ P(1,2) &= \frac{1}{2!}\frac{d^2}{dz^2}F(1,z)\bigg|_{z=0}, \\ P(1,3) &= \frac{1}{3!}\frac{d^3}{dz^3}F(1,z)\bigg|_{z=0}. \end{aligned}$$

One can therefore easily show by induction that

$$P(1,m) = \frac{1}{m!}\frac{d^m}{dz^m}F(1,z)\bigg|_{z=0}.$$

Therefore,

$$P(1,0) = \frac{1}{0!}F(1,z)\bigg|_{z=0} = \frac{\varepsilon-1}{\psi-1}\,{}_1F_1(\varepsilon;\psi;u_0),$$

$$\begin{aligned} P(1,1) = \frac{1}{1!}\frac{d}{dz}F(1,z)\bigg|_{z=0} &= \alpha\,\frac{\varepsilon}{\psi}\frac{\varepsilon-1}{\psi-1}\,{}_1F_1(\varepsilon+1;\psi+1;\alpha(z-1))\bigg|_{z=0} \\ &= \alpha\,\frac{\varepsilon}{\psi}\frac{\varepsilon-1}{\psi-1}\,{}_1F_1(\varepsilon+1;\psi+1;u_0), \end{aligned}$$

while,

$$P(1,2) = \frac{1}{2!}\frac{d^2}{dz^2}F(1,z)\bigg|_{z=0} = \frac{1}{2!}\,\alpha^2\,\frac{\varepsilon}{\psi}\,\frac{\varepsilon-1}{\psi-1}\frac{\varepsilon+1}{\psi+1}\,{}_1F_1(\varepsilon+2;\psi+2;u_0).$$

As seen below, we can prove by induction that

$$P(1,m) = \frac{\alpha^m}{m!}\ \frac{\varepsilon-1}{\psi-1}\frac{(\varepsilon)_m}{(\psi)_m}\ {}_1F_1(\varepsilon+m;\psi+m;u_0). \tag{2.14}$$

Theorem 1 *The stationary probability distribution for the number of mRNA molecules transcribed from an active unregulated gene follows a confluent hypergeometric probability distribution. Specifically, the distribution is of the form presented in Eq. (2.14).*

Proof:

We have shown above that Eq. (2.14) is true for the probabilities of having 1 or 2 mRNA molecules. Suppose it is true up to some particular value m. We want to prove that

$$P(1,m+1) = \frac{\alpha^{m+1}}{(m+1)!}\ \frac{(\varepsilon)_m}{(\psi)_m}\frac{\varepsilon-1}{\psi-1}\ \frac{(\varepsilon+m)}{(\psi+m)}\ {}_1F_1(\varepsilon+m+1;\psi+m+1;u_0).$$

Now, from Eq. (2.14).

$$\begin{aligned} P(1,m+1) &= \frac{1}{(m+1)!}\ \frac{d^{m+1}}{dz^{m+1}}F(1,z)\bigg|_{z=0} = \frac{1}{(m+1)!}\ \frac{d}{dz}\left[\frac{d^m}{dz^m}F(1,z)\right]\bigg|_{z=0} \\ &= \frac{1}{(m+1)!}\ \frac{d}{dz}\left[\alpha^m\frac{\varepsilon-1}{\psi-1}\frac{(\varepsilon)_m}{(\psi)_m}\ {}_1F_1(\varepsilon+m;\psi+m;\alpha(z-1))\right]\bigg|_{z=0} \\ &= \frac{\alpha^{m+1}}{(m+1)!}\frac{(\varepsilon)_m}{(\psi)_m}\frac{\varepsilon-1}{\psi-1}\frac{(\varepsilon+m)}{(\psi+m)}\ {}_1F_1(\varepsilon+m+1;\psi+m+1;u_0). \end{aligned}$$

To derive an expression for $P(0,m)$, we first replace m by $m+1$ and $m-1$ in the expression for $P(1,m)$ to obtain, respectively

$$P(1,m+1) = \frac{\alpha^{m+1}}{(m+1)!}\frac{\varepsilon-1}{\psi-1}\frac{(\varepsilon)_{m+1}}{(\psi)_{m+1}}\ {}_1F_1(\varepsilon+m+1;\psi+m+1;u_0), \tag{2.15}$$

$$P(1,m-1) = \frac{\alpha^{m-1}}{(m-1)!}\frac{\varepsilon-1}{\psi-1}\frac{(\varepsilon)_{m-1}}{(\psi)_{m-1}}\ {}_1F_1(\varepsilon+m-1;\psi+m-1;u_0), \tag{2.16}$$

and then substitute these in Eq. (2.3). Using the definition of the rising factorial given above, it follows that $(\varepsilon)_{m+1} = \varepsilon(\varepsilon+1)...(\varepsilon+m) = (\varepsilon)_m(\varepsilon+m)$, and similarly, one should easily deduce that $(\varepsilon)_{m+n} = (\varepsilon)_m(\varepsilon+m)_n$. We therefore make the following simplifications.

$$\frac{(\varepsilon)_{m+1}}{(\psi)_{m+1}} = \frac{(\varepsilon)_m}{(\psi)_m}\frac{(\varepsilon+m)}{(\psi+m)},$$

while

$$\frac{(\varepsilon)_{m-1}}{(\psi)_{m-1}} = \frac{(\varepsilon)_m}{(\psi)_m}\frac{(\psi+m-1)}{(\varepsilon+m-1)}.$$

Substituting these in Eq. (2.3), we deduce the expression for $P(0,m)$ as

$$P(0,m) = \frac{\alpha^m}{(\psi-1)m!}\frac{(\varepsilon)_m}{(\psi)_m}\Bigg[(\alpha+\psi+m-\varepsilon)\ {}_1F_1(\varepsilon+m;\psi+m;u_0)$$

$$-\alpha\frac{\varepsilon+m}{\psi+m}\ {}_1F_1(\varepsilon+m+1;\psi+m+1;u_0)$$

$$-m\frac{\psi+m-1}{\varepsilon+m-1}\ {}_1F_1(\varepsilon+m-1;\psi+m-1;u_0)\Bigg]. \tag{2.17}$$

The marginal probability distribution function can easily be recovered by summing the joint probabilities over all possible values of m. This assists in cases where the same copy of a particular gene occurs in separate cells. Even if the probability that this particular gene is active may be known, it may not be clear in which cells this gene is active. It would be interesting to be able to predict the distribution of the number of mRNA molecules at any

time that can be transcribed from this gene, i.e. the probability $P(m)$ of having m mRNA molecules irrespective of the state of the gene. The marginal distribution of m is

$$P(m) = \sum_{s=0}^{1} P(s,m) = P(0,m) + P(1,m).$$

The above expressions for $P(0,m)$ and $P(1,m)$ can be described as the probabilities of being in a state s and obtaining m molecules. In other words, they are simply the joint probability distributions. Mathematically, they are best interpreted as intersections of two sets of probabilities, $P(s) \cap P(m)$. Dividing these joint probability distributions by the probability that the promoter is in a state s leaves one with the conditional probabilities. We derive and present these in the following section.

2.7 Conditional probabilities

As described above, the joint probability, $P(1,m) = P(1) \cap P(m)$, is an intersection between the probability that the gene is active and the probability of obtaining m mRNA molecules. Generally, the conditional probability $P(m|s)$, which is the probability of having a certain number m of molecules of mRNA given that the gene is in a state s, is found by using the relationship

$$P(m|s) = \frac{P(m) \cap P(s)}{P(s)} = \frac{P(s,m)}{P(s)}.$$

Using the above relationship, we find the following expressions for the conditional probabilities $P(m|1)$ and $P(m|0)$,

$$P(m|1) = \frac{P(1,m)}{\mu} = \frac{\alpha^m}{m!} \frac{(\varepsilon)_m}{(\psi)_m} {}_1F_1(\varepsilon + m; \psi + m; u_0) \tag{2.18}$$

and

$$P(m|0) = \frac{P(0,m)}{1-\mu} = \frac{\alpha^m}{(\psi-\varepsilon)m!}\frac{(\varepsilon)_m}{(\psi)_m}\Bigg[(\alpha+\psi+m-\varepsilon)\ {}_1F_1(\varepsilon+m;\psi+m;u_0)$$

$$-\alpha\frac{\varepsilon+m}{\psi+m}\ {}_1F_1(\varepsilon+m+1;\psi+m+1;u_0)$$

$$-m\frac{\psi+m-1}{\varepsilon+m-1}\ {}_1F_1(\varepsilon+m-1;\psi+m-1;u_0)\Bigg]. \tag{2.19}$$

The probabilities relating to the state of the promoter are $P(1) = \mu = \frac{\varepsilon-1}{\psi-1}$ and $P(0) = 1-\mu = \frac{\psi-\varepsilon}{\psi-1}$. We will derive the marginal probability distributions in a section below. In the meantime, we take a look at the moments related to the conditional probability distributions we have just obtained.

2.8 Moments of the distribution

In a stochastic environment like the one which we are analysing, the number of molecules continuously fluctuates. Instead of focusing on the instantaneous number of molecules, one should speak of average numbers and deviations from these averages. These are expressed as functions of the model parameters and are expected to change as these parameters are changed. One can therefore make inferences about the effect of a change of any parameter appearing in these expressions. The key qualities of a distribution are captured in numerical form by measures known as moments. Presented below are some derivations of the expressions for the most important moments, the first (or mean) and the second (van Kampen, 1992). Our aim is to derive the moments for the marginal distributions, a task we cannot

do directly without traversing through two other kinds of moments, the factorial and conditional moments. The following two sections contain detailed derivations of these two sets of moments.

2.9 Factorial moments

A factorial moment is the expectation value of the rising factorial defined in Eq. (2.9). It is found by differentiating the moment generating functions in Eqs. (2.10), (2.12), and evaluating the results at $z = 1$. The first and second derivatives give the first and second factorial moments respectively. The ordinary moments of the distributions can be recovered from the factorial moments. We first define E_s as a partial expectation value, given by

$$E_s(f(m)) = \sum_{m=0}^{\infty} f(m)P(s,m).$$

Next, we ascertain that if $\nu_{s,i}$ is the i^{th} factorial moment when the gene is in a state s, then

$$\nu_{s,i} := E_s\left[\frac{m!}{(m-i)!}\right] = \left(\frac{d^i F(s,z)}{dz^i}\right)\Bigg|_{z=1}. \tag{2.20}$$

This can be proved easily, since mathematically

$$E_s\left[\frac{m!}{(m-i)!}\right] = E_s[m(m-1)...(m-i+1)]$$

$$= \sum_{m=0}^{\infty}[m(m-1)...(m-i+1)]P(s,m) = \sum_{m=0}^{\infty}[m(m-1)...(m-i+1)]P(s,m)z^{m-i}\Big|_{z=1}$$

$$= \sum_{m=0}^{\infty}P(s,m)\frac{d^i}{dz^i}z^m\Big|_{z=1} = \frac{d^i}{dz^i}\left[\sum_{m=0}^{\infty}z^m P(s,m)\right]\Bigg|_{z=1} = \frac{d^i F(s,z)}{dz^i}\Big|_{z=1}.$$

The probabilities in Eqs. (2.6) converge absolutely to zero since the moment generating functions converge, which allows us to commute the summation and differential operators in the above steps. In fact, from Eqs. (2.10), (2.12), we have

$$\forall z_1, z_2 \in \{z : |z| \leq 1\}, z_1 \neq z_2 \Rightarrow F(s,z_1) \neq F(s,z_2).$$

Thus $F(s,z)$ is one-to-one. However, it is possible that $P(s,m_1) = P(s,m_2)$ when $m_1 \neq m_2$, where $m_1, m_2 \in [0,\infty)$ meaning that for some parameters, $P(s,m)$ vs m has a maximum. Since

$$|F(s,z)| = \sum_{m=0}^{\infty} |z^m P(s,m)| = \sum_{m=0}^{\infty} |z^m||P(s,m)| \leq \sum_{m=0}^{\infty} |P(s,m)|,$$

and $\sum_{m=0}^{\infty} P(s,m) \leq 1$, then $\sum_{m=0}^{\infty} P(s,m)$ is absolutely convergent. Using our moment generating functions in Eqs. (2.10), (2.12), the first two factorial moments can be derived as

$$\begin{aligned}
\nu_{s,1} = \left.\frac{dF(s,z)}{dz}\right|_{z=1} &= \left[\sum_{m=0}^{\infty} m z^{m-1} P(s,m)\right]\Bigg|_{z=1} = \sum_{m=0}^{\infty} m P(s,m). \\
\left.\frac{d^2F(s,z)}{dz^2}\right|_{z=1} &= \left[\sum_{m=0}^{\infty} m(m-1) z^{m-2} P(s,m)\right]\Bigg|_{z=1} \\
= \sum_{m=0}^{\infty} m(m-1) P(s,m) &= \sum_{m=0}^{\infty} m^2 P(s,m) - \sum_{m=0}^{\infty} m P(s,m) \\
= E_s(m^2) - E_s(m) &= E_s(m^2) - \nu_{s,1}, \\
\Rightarrow \nu_{s,2} = \left.\frac{d^2F(s,z)}{dz^2}\right|_{z=1} &= E_s(m^2) - E_s(m).
\end{aligned}$$

Therefore, the second moment in a state s is the sum of the first and second factorial moments in that state, i.e.

$$E_s(m^2) = \nu_{s,2} + \nu_{s,1} = \left.\frac{d^2F(s,z)}{dz^2}\right|_{z=1} + \nu_{s,1}. \tag{2.21}$$

From Eq. (2.11)

$$\frac{d^iF(1,z)}{dz^i} = \alpha^i \, \frac{\varepsilon-1}{\psi-1}\frac{(\varepsilon)_i}{(\psi)_i} {}_1F_1(\varepsilon+i;\psi+i;u_z),$$

such that

$$\nu_{1,i} = \left[\alpha^i \, \frac{\varepsilon-1}{\psi-1}\frac{(\varepsilon)_i}{(\psi)_i} \, {}_1F_1(\varepsilon+i;\psi+i;u_z)\right]\Bigg|_{z=1} = \alpha^i \, \frac{\varepsilon-1}{\psi-1}\frac{(\varepsilon)_i}{(\psi)_i}.$$

If we consider, for instance $i=1$ and $i=2$, then we get

$$E_1(m) = \alpha \, \frac{\varepsilon-1}{\psi-1}\frac{\varepsilon}{\psi}, \tag{2.22}$$

and

$$\nu_{1,2} = \alpha \, \frac{\varepsilon+1}{\psi+1} E_1(m). \tag{2.23}$$

In a similar fashion, we derive the first and second factorial moments, for the "off" state. Differentiating once the expression for $F(0,z)$ presented in Eq. (2.12) above, followed by the substitution $z=1$, we find the first "off" state (factorial) moment to be

$$E_0(m) = \frac{\alpha}{\psi}\frac{\varepsilon-1}{\psi-1}(\psi-\varepsilon), \tag{2.24}$$

while the related second factorial moment is found as

$$\nu_{0,2} = \frac{\alpha\varepsilon}{\psi+1} E_0(m). \tag{2.25}$$

2.10 The conditional moments

The partial moments that we derived above are useful in giving us the conditional moments. These will assist in capturing the qualities of the statistics of a distribution, given a certain state of the gene. We can thus define, quantitatively, the mean, the variance and the level of stochasticity, given that the gene is in a particular state. The moments already derived above represent an intersection of the set of values of a given statistic and the state of the gene. To obtain the conditional moments, we divide each moment by the respective probability of the state of the gene. In this respect, we take $\langle m^j \rangle_s$ as the j^{th} conditional moment when the gene is in a state s, such that

$$\langle m^j \rangle_s = \sum_{m=0}^{\infty} m^j P(m|s) = \sum_{m=0}^{\infty} m^j \frac{P(s,m)}{P(s)} = \frac{1}{P(s)} \sum_{m=0}^{\infty} m^j P(s,m) = \frac{E_s(m^j)}{P(s)}. \tag{2.26}$$

For instance, we find that the first conditional moment given that the gene is active would be found as

$$\langle m^1 \rangle_1 = \sum_{m=0}^{\infty} mP(m|1) = \sum_{m=0}^{\infty} m \frac{P(1,m)}{P(1)} = \frac{\psi-1}{\varepsilon-1} \sum_{m=0}^{\infty} mP(1,m), \tag{2.27}$$

that is,

$$\langle m^1 \rangle_1 = \frac{E_1(m^1)}{P(1)} = \frac{\alpha \frac{\varepsilon-1}{\psi-1} \frac{\varepsilon}{\psi}}{\frac{\varepsilon-1}{\psi-1}} = \alpha \, \frac{\varepsilon}{\psi}. \tag{2.28}$$

Using Eq. (2.26), we get that

$$\langle m^2 \rangle_1 = \frac{1}{\mu} E_1(m^2) = \frac{1}{\mu}[\nu_{1,2} + E_1(m)],$$

simplifying to

$$\left[\alpha \frac{\varepsilon+1}{\psi+1} + 1\right] \frac{E_1(m)}{\mu} = \left[\alpha \frac{\varepsilon+1}{\psi+1} + 1\right] \langle m^1 \rangle_1. \tag{2.29}$$

When the gene is inactive, we find that

$$\langle m^1 \rangle_0 = \frac{\alpha}{\psi}(\varepsilon - 1), \tag{2.30}$$

while

$$\langle m^2 \rangle_0 = \left[\frac{\alpha\varepsilon}{\psi+1} + 1\right] \langle m^1 \rangle_0. \tag{2.31}$$

2.11 Marginal probabilities and their moments

The marginal distribution described in section 2.6 above is given by

$$P(m) = P(0,m) + P(1,m) = \frac{\alpha^m}{(\psi-1)m!} \frac{(\varepsilon)_m}{(\psi)_m} \Bigg[(\alpha+\psi+m-1)\ {}_1F_1(\varepsilon+m;\psi+m;u_0)$$

$$-\alpha \frac{\varepsilon+m}{\psi+m}\ {}_1F_1(\varepsilon+m+1;\psi+m+1;u_0)$$

$$-m \frac{\psi+m-1}{\varepsilon+m-1}\ {}_1F_1(\varepsilon+m-1;\psi+m-1;u_0) \Bigg]. \tag{2.32}$$

Suppose $\langle m^j \rangle$ represents the j^{th} moment for the marginal distribution, where $j \in [0, \infty)$, then

$$\langle m^j \rangle = \sum_{m=0}^{\infty} m^j [P(0,m) + P(1,m)] = E_0(m^j) + E_1(m^j). \tag{2.33}$$

The case $j = 0$ simply gives $\sum_{m=0}^{\infty} [P(0,m) + P(1,m)] = 1$, the normalization condition for the joint distribution. For $j = 1$ and $j = 2$, we find that

$$\begin{aligned} \langle m \rangle &= E_0(m) + E_1(m), \\ \langle m^2 \rangle &= E_0(m^2) + E_1(m^2) = \nu_{0,2} + \nu_{1,2} + E_0(m) + E_1(m), \end{aligned}$$

giving us

$$\langle m \rangle = \alpha \frac{\varepsilon - 1}{\psi - 1} \tag{2.34}$$

$$\langle m^2 \rangle = \left(1 + \frac{\alpha\varepsilon}{\psi}\right) \langle m \rangle. \tag{2.35}$$

The first moment signifies the mean, $\langle m \rangle$, or the average number of mRNA molecules. Using the two moments above, we can deduce an expression for the variance, σ^2, which measures the spread of the numbers of the mRNA molecules, m, from their mean. We find that,

$$\sigma^2 = \langle m^2 \rangle - \langle m \rangle^2 = \left[\frac{\alpha\varepsilon}{\psi} + 1 - \langle m \rangle\right] \langle m \rangle. \tag{2.36}$$

Using the variance, we can get a second useful statistic, the standard deviation, by simply computing its square root. The standard deviation tells us how much each of the observed number of mRNA molecules, on the average, varies from the mean, $\langle m \rangle$. A larger value for the standard deviation implies greater variability around the mean. Since a mean allows an

entire distribution of the numbers of mRNA molecules to be characterized by an individual score, its reliability diminishes as the standard deviation becomes larger.

Considering the dynamics on the promoter and the mRNA dynamics we can define parameters α, K_{eq} and ε. The first parameter is a ratio of rate constants for mRNA synthesis/degradation and is hence related to mRNA dynamics. K_{eq} is an equilibrium constant for the state of the promoter and is thus a ratio of the rate constants for gene activation/deactivation. The third parameter, ε, links the time scales of these two subsystems, i.e. $\alpha = \frac{k_m}{k_{\bar{m}}}, \varepsilon = 1 + \frac{k_{\text{on}}}{k_{\bar{m}}}$ and $K_{eq} = \frac{k_{\text{on}}}{k_{\text{off}}}$. Now, in terms of these parameters, the standard deviation can be expressed as

$$\sigma = \sqrt{\frac{\alpha K_{eq}[K_{eq}(\varepsilon(2+K_{eq})+\alpha-1)+\varepsilon-1]}{(1+K_{eq})^2(\varepsilon(1+K_{eq})-1)}}, \tag{2.37}$$

and we find that as $K_{eq} \to \infty$, $\sigma \to \sqrt{\alpha}$. Thus as K_{eq} becomes large, the standard deviation is only affected by the mRNA dynamics. In Fig. 2.2 below we plot the standard deviation, σ, for the marginal probability distribution, against one of the parameters, K_{eq}. This figure shows that variability increases as the parameter K_{eq} increases. Increasing the parameter K_{eq} is similar to keeping the gene more into the active state and hence more molecules are produced and the variability thus increases. The observable however reaches a maximum value of 2.59 as the parameter $K_{eq} \to \infty$, and is zero when K_{eq} is zero. A similar behavior was observed for the relationship between the observables $\sigma_0, \sigma_1, \langle m \rangle$, $\langle m^1 \rangle_0$, $\langle m^1 \rangle_1$ and the parameter K_{eq}.

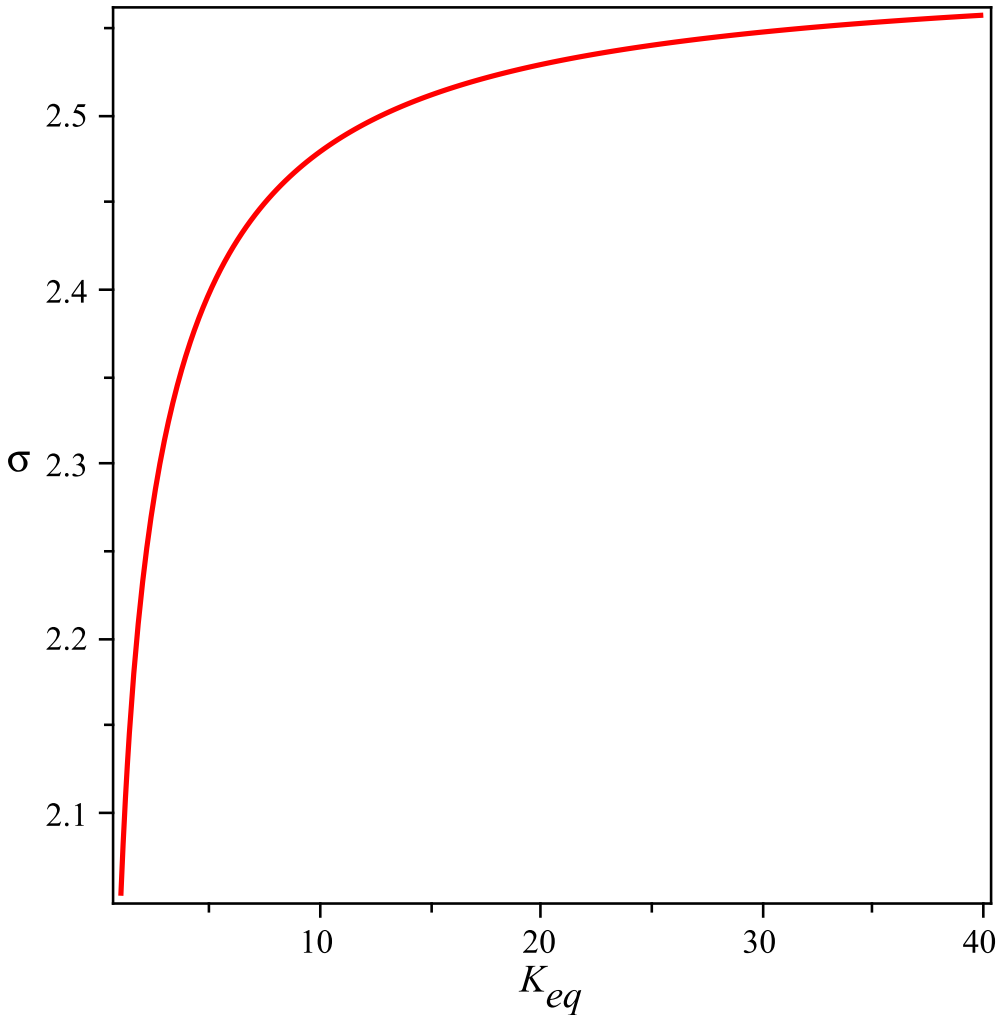

Figure 2.2: Behavior of the standard deviation for the marginal distribution, σ, as K_{eq} is varied, with $\alpha = 6.7$ and $\varepsilon = 3$ (as in [40]).

2.12 Noise and its measures

One way to study genetic control mechanisms is by analysing expressions that relate to the strength of the noise. Mathematically, the strength of the noise is measured by using the coefficient of variation, CV, which is defined as the ratio of the standard deviation to the mean of the distribution [42]. It is dimensionless and serves the purpose of revealing the magnitude of the ratio of a typical fluctuation to the mean. The mean itself is accompanied by some disadvantages which make it unsuitable for use in the interpretation of data when sampling. For example, it is affected by extreme values in a data set, the larger the values sitting at the end points of the set, the larger the mean of the whole data set. When we say the standard deviation helps in telling us about the variability, it does not really mean we can depend on it. More so, a lower standard deviation does not necessarily mean less variable data. As an example, we can assume that we have two data sets, one with a mean value of 60 and standard deviation 10, and a second one with 0.10 and 0.40 for the mean and standard deviation respectively. Thus we could say the first data set has a greater variability as compared to the second set based on the standard deviations. However, upon computing the CV, we find that these two data sets have CVs of values 0.17 and 4 respectively, which contradicts our initial conclusion in favor of a new one, that there is greater variability in the second data set. The CV measures variability relative to the mean and is hence preferable for many purposes.

TF binding/unbinding is one source of noise which should not be neglected in all models designed to model mRNA and protein distributions [28]. Investigating the behaviour of the CV as K_{eq} changes assists us in corroborating this fact. Our results reveal a decaying plot of the CV against K_{eq} (Fig. 2.3). Noise is high in all the parameter (k_{on} and k_{off}) regimes favouring small values of K_{eq}. Thus the efficiency of the binding of transcription factors to enable transcription plays a significant role in determining the magnitude of the fluctuations

originating from the activation and deactivation of the gene. As a rule of thumb, noise, η, is inversely proportional to the square root of the number of molecules, $\eta \propto \frac{1}{\sqrt{m}}$ [63, 69].

Increasing K_{eq} is tantamount to keeping the gene in the active state more often than it is kept inactive. This would ensure an increase in the number of mRNA molecules, thereby lessening the noise levels. The reverse is also true. Figure 2.3 illustrates that, unless the gene is assumed to be active always [68], as can be the case with other genes, for instance the so-called housekeeping genes, binding and unbinding of transcriptional factors on operator sites should not be ignored as a major source of intrinsic noise. Noise is inherent in stochastic systems and it would not be accurate to consider fluctuations only at the transcriptional and translational levels as important in models. Gene activation and deactivation processes are never equiprobable, stochasticity causes bias towards one of these processes more than to the other, at any time. The effect of such bias is observed in the uneven distribution or bursts of the gene products, e.g mRNA, with time [40, 68].

Using our analytic expressions for the mean and the standard deviation above, Eqs. 2.34 and 2.38, we find that noise is inversely proportional to the parameter α. Moreover, as

$$K_{eq} \to \infty, \mathrm{CV} \to \frac{1}{\sqrt{\alpha}}.$$

Fig. 2.3 shows that the CV reduces to a minimum, found to be 0.082, as $K_{eq} \to \infty$, implying that the strength of the noise pegs at a constant low level as the number of molecules keeps increasing. The noise strength correspondingly increases as K_{eq} decreases.

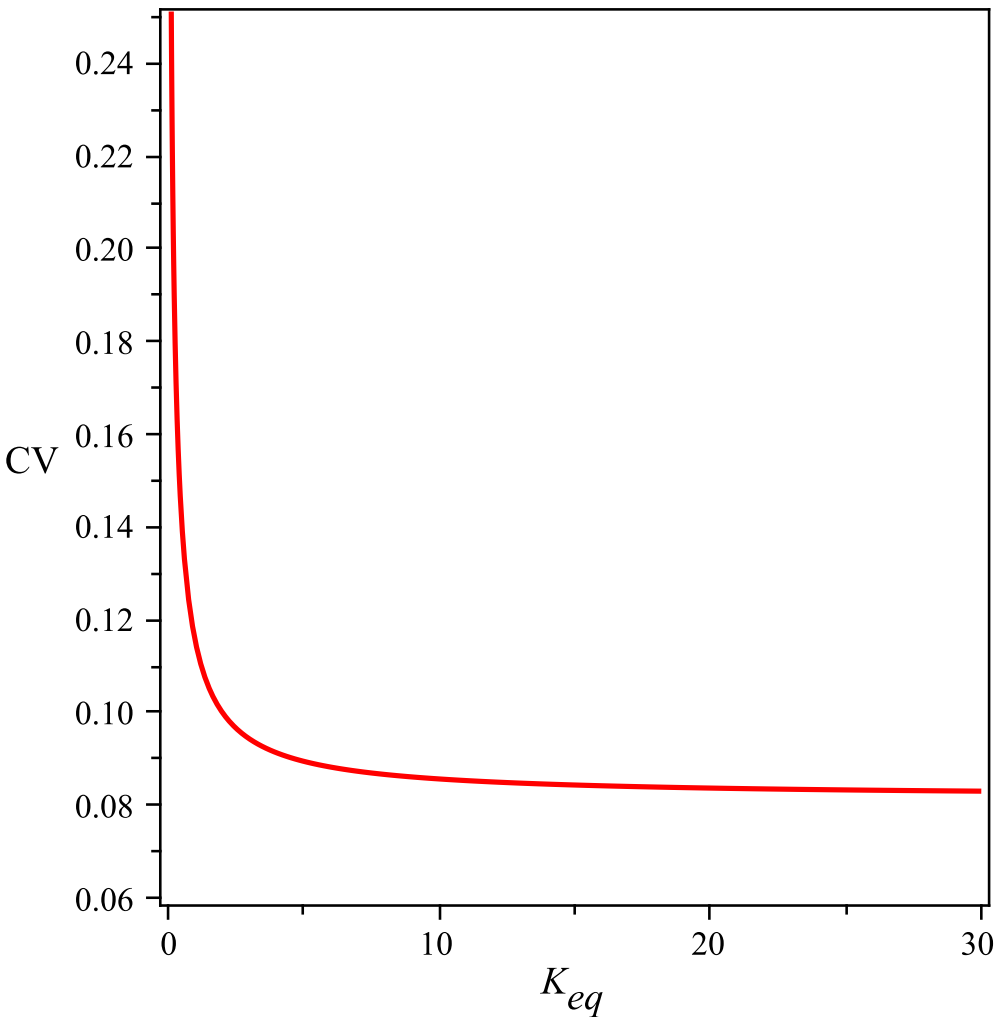

Figure 2.3: A plot of the coefficient of variation against the equilibrium constant, K_{eq}, while $\alpha = 150$, $\varepsilon = 72000$ (as in [30]).

2.13 The skewness and the kurtosis

We considered the number of mRNA molecules, m, as a random variable with mean $\langle m \rangle$ and variance σ^2 calculated above. Two more important moments for the marginal distribution of m can still be calculated. These are the third and fourth moments, called the skewness and the kurtosis respectively.

The skewness measures the asymmetry in the distribution of the values of m. We can easily get a picture of whether the distribution has a longer tail or a shorter tail, to the left or to the right of the mean value by observing the sign and magnitude of the third moment. A positive value indicates a longer tail to the right, in contrast to a negative skewness, marked by a longer tail to the left.

The excess kurtosis measures the peakedness of the data compared to the normal distribution, which has an excess kurtosis of zero. A distribution with a negative excess kurtosis is described as platykurtic, and the opposite is described as leptokurtic. A platykurtic distribution has a broad, slowly decaying maximum. When the kurtosis is greater than zero or positive, the curve is strongly peaked and we find a large area squashed into a small range of the distribution. We have to be aware that although two distributions may have the same variance and approximately the same value for the skewness, they may differ markedly when it comes to the kurtosis.

Mathematically, the two are standardized third and fourth central moments, where the r^{th} standardized central moment is given by

$$E\left[\left(\frac{m-\langle m \rangle}{\sigma}\right)^r\right]. \tag{2.38}$$

Thus, if $r = 3$, we can calculate the skewness, γ_1. Using our previously derived results and a binomial expansion of the term in the numerator, we find this as

$$\gamma_1 = \frac{1}{\sigma^3}[E(m^3) - 3\langle m\rangle E(m^2) + 2\langle m\rangle^3]$$

$$= \frac{1}{\sigma^3}[E_0(m^3) + E_1(m^3) - 3\langle m\rangle E(m^2) + 2\langle m\rangle^3], \tag{2.39}$$

which, upon substitution, gives

$$\gamma_1 = \frac{1}{\sigma^3}\left[\frac{\alpha^3}{\psi-1}\frac{(\varepsilon)_2}{(\psi)_2}\left(\frac{\varepsilon+2}{\psi+2}(3+\psi-\varepsilon)-3\right) + \left(1+\frac{3\alpha\varepsilon}{\psi+1}(1-\langle m\rangle) - 3\langle m\rangle\right)E_0(m)\right.$$

$$\left. + \left(1 + (3+\alpha\frac{\varepsilon+2}{\psi+2} - 3\langle m\rangle)\alpha\frac{\varepsilon+1}{\psi+1} - 3\langle m\rangle\right)E_1(m) + 2\langle m\rangle^3\right]. \tag{2.40}$$

Similarly, taking $r = 4$, we get the fourth central moment, or the kurtosis (β_2). However, the excess kurtosis $\gamma_2 = \beta_2 - 3$ is more convenient since $\gamma_2 = 0$ for a Gaussian distribution.

$$\gamma_2 = \frac{1}{\sigma^4}[E_0(m^4) + E_1(m^4) - 4\langle m\rangle(E_0(m^3) + E_1(m^3))$$

$$+6\langle m\rangle^2(E_0(m^2) + E_1(m^2)) - 3\langle m\rangle^4] - 3. \tag{2.41}$$

The partial expectation values required are

$$E_1(m^3) = \left[1 + \frac{\alpha(\varepsilon+1)}{\psi+1}\left(3 + \frac{\alpha(\varepsilon+2)}{\psi+2}\right)\right]E_1(m),$$

$$E_0(m^3) = \frac{\alpha^3}{\psi-1}\frac{(\varepsilon)_2}{(\psi)_2}\left[\frac{\varepsilon+2}{\psi+2}(3+\psi-\varepsilon)-3\right] + \left(1+\frac{3\alpha\varepsilon}{\psi+1}\right)E_0(m),$$

$$E_1(m^4) = \alpha^4\frac{\varepsilon-1}{\psi-1}\frac{(\varepsilon)_4}{(\psi)_4} + \left[6\left(1+\frac{\alpha(\varepsilon+1)}{\psi+1}(3+\frac{\alpha(\varepsilon+2)}{\psi+2})\right) - 11\frac{\alpha(\varepsilon+1)}{\psi+1} - 5\right]E_1(m),$$

and

$$E_0(m^4) = \frac{\alpha^4}{\psi-1}\frac{(\varepsilon)_3}{(\psi)_3}\left[\frac{\varepsilon+3}{\psi+3}(4+\psi-\varepsilon)-4\right] + \frac{6\alpha^3}{\psi-1}\frac{(\varepsilon)_2}{(\psi)_2}\left[\frac{\varepsilon+2}{\psi+2}(3+\psi-\varepsilon)-3\right]$$

$$+\left[1+\frac{7\alpha\varepsilon}{\psi+1}\right]E_0(m).$$

Similarly, one can obtain expressions for the skewness and the kurtosis corresponding to the conditional probability distribution functions derived above, i.e. $P(m|0)$ and $P(m|1)$. This can be done by computing the "conditional" expectations,

$$\langle f(m)\rangle_s = \sum_{m=0}^{\infty} f(m)P(m|s),$$

$s = 0, 1$, of the bracketed term in Eq. (2.37). This leads to a general expression for the skewness (catering for the two states of the promoter), denoted as

$$\gamma_1^s = \left\langle \left(\frac{m-\langle m^1\rangle_s}{\sigma_s}\right)^3 \right\rangle_s = \frac{1}{\sigma_s^3}\left[\langle m^3\rangle_s - 3\langle m^1\rangle_s\langle m^2\rangle_s + 2(\langle m^1\rangle_s)^3\right].$$

Hence, when $s = 0$ and 1, respectively for the "off" and "on" states, we get

$$\gamma_1^0 = \frac{\psi-1}{\sigma_0^3(\psi-\varepsilon)}\left[E_0(m^3) - 3\langle m^1\rangle_0\, E_0(m^2) + 2(\langle m^1\rangle_0)^3\frac{\psi-\varepsilon}{\psi-1}\right], \tag{2.42}$$

$$\gamma_1^1 = \frac{\psi - 1}{\sigma_1^3(\varepsilon - 1)} \left[E_1(m^3) - 3\langle m^1 \rangle_1 \, E_1(m^2) + 2(\langle m^1 \rangle_1)^3 \frac{\varepsilon - 1}{\psi - 1} \right]. \tag{2.43}$$

The corresponding expressions for the excess kurtosis can be obtained from,

$$\gamma_2^s = \left\langle \left(\frac{m - \langle m^1 \rangle_s}{\sigma_s} \right)^4 \right\rangle_s = \frac{1}{\sigma_s^4} \left[\langle m^4 \rangle_s - 4\langle m^1 \rangle_s \langle m^3 \rangle_s + 6(\langle m^1 \rangle_s)^2 \langle m^2 \rangle_s - 3(\langle m^1 \rangle_s)^4 \right] - 3,$$

implying that,

$$\gamma_2^0 = \frac{\psi - 1}{\sigma_0^4(\psi - \varepsilon)} \left[E_0(m^4) - 4\langle m^1 \rangle_0 \, E_0(m^3) \right.$$

$$\left. + 6(\langle m^1 \rangle_0)^2 \, E_0(m^2) - 3(\langle m^1 \rangle_0)^4 \frac{\psi - \varepsilon}{\psi - 1} \right] - 3, \tag{2.44}$$

$$\gamma_2^1 = \frac{\psi - 1}{\sigma_1^4(\varepsilon - 1)} \left[E_1(m^4) - 4\langle m^1 \rangle_1 \, E_1(m^3) \right.$$

$$\left. + 6(\langle m^1 \rangle_1)^2 \, E_1(m^2) - 3(\langle m^1 \rangle_1)^4 \frac{\varepsilon - 1}{\psi - 1} \right] - 3, \tag{2.45}$$

where, in each case above,

$$\sigma_s^2 = \langle m^2 \rangle_s - (\langle m^1 \rangle_s)^2$$

and $E_s(\cdot)$ denotes the partial expectation. In the next Chapter, we will use these analysis results to obtain the control coefficients. We end this chapter by investigating the behavior of the skewness and the kurtosis for both the conditional and the marginal distributions, as one of the parameters, K_{eq}, varies. Our results are summarized in Fig. 2.4 below.

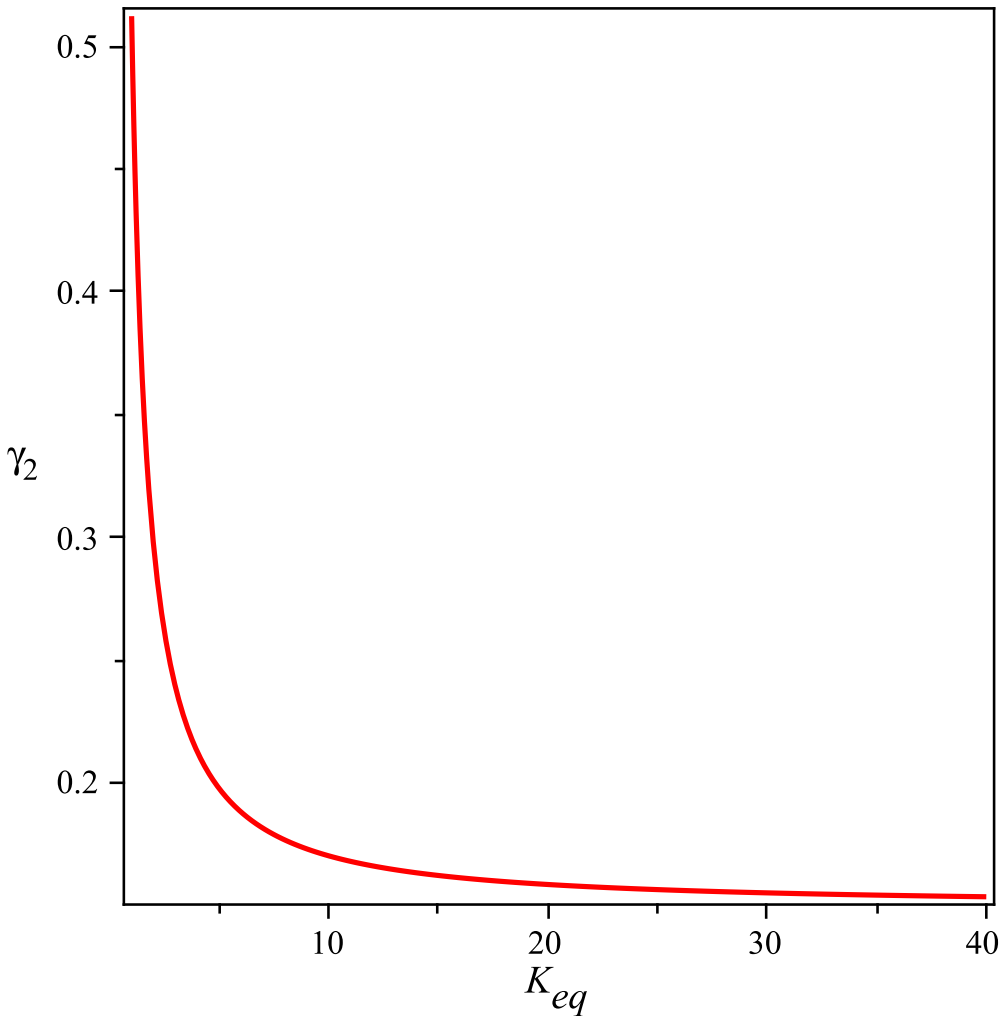

Figure 2.4: The relationship between the excess kurtosis for the marginal distribution, γ_2 and the parameter K_{eq}, while $\alpha = 6.7$ and $\varepsilon = 3$ (as in [40]).

Fig. 2.4 illustrates that the observable, γ_2 in this case, decreases as K_{eq} increases. This curve is asymptotic to the line $K_{eq} = 0$, but $\gamma_2 \to 0.150$ as $K_{eq} \to \infty$. The relationship between the observables $\gamma_1, \gamma_1^1, \gamma_1^0$, γ_2^1, γ_2^0 and the parameter K_{eq} is similar to the one depicted in this figure. Increasing the parameter K_{eq} forces the distributions to become less leptokurtic, i.e. less frequencies remain trapped in small ranges of the distribution.

2.14 Transcriptional factor–operator binding and unbinding equilibrium

In this section, we look at how changes in K_{eq} affect the shapes of the distributions. Transcriptional factors bind to the operator sites of the promoter thereby enabling transcription. This binding process is reversible, and so at any given time, an equilibrium can be attained between the binding and unbinding processes. As mentioned before, we can define an effective equilibrium constant $K_{eq} = \frac{k_{\text{on}}}{k_{\text{off}}}$, for the reversible processes of activation/deactivation of the gene.

Results displayed in Figs. 2.5 and 2.6 below show that increasing K_{eq} results in a shift of the non-zero probability ranges towards higher values of m, stretching the curves along the m-axis and reducing their peaks. A decrease in K_{eq} generates the reverse transformations. Therefore, increasing K_{eq} reduces the excess kurtosis but increases the standard deviations, yet decreasing K_{eq} increases the excess kurtosis and reduces the standard deviation. Correspondingly, Figs. 2.5 is flatter and wider as compared to Fig. 2.6 which is peaked and narrow. The same plots or observations could have been obtained had $P(m|0)$ been used instead.

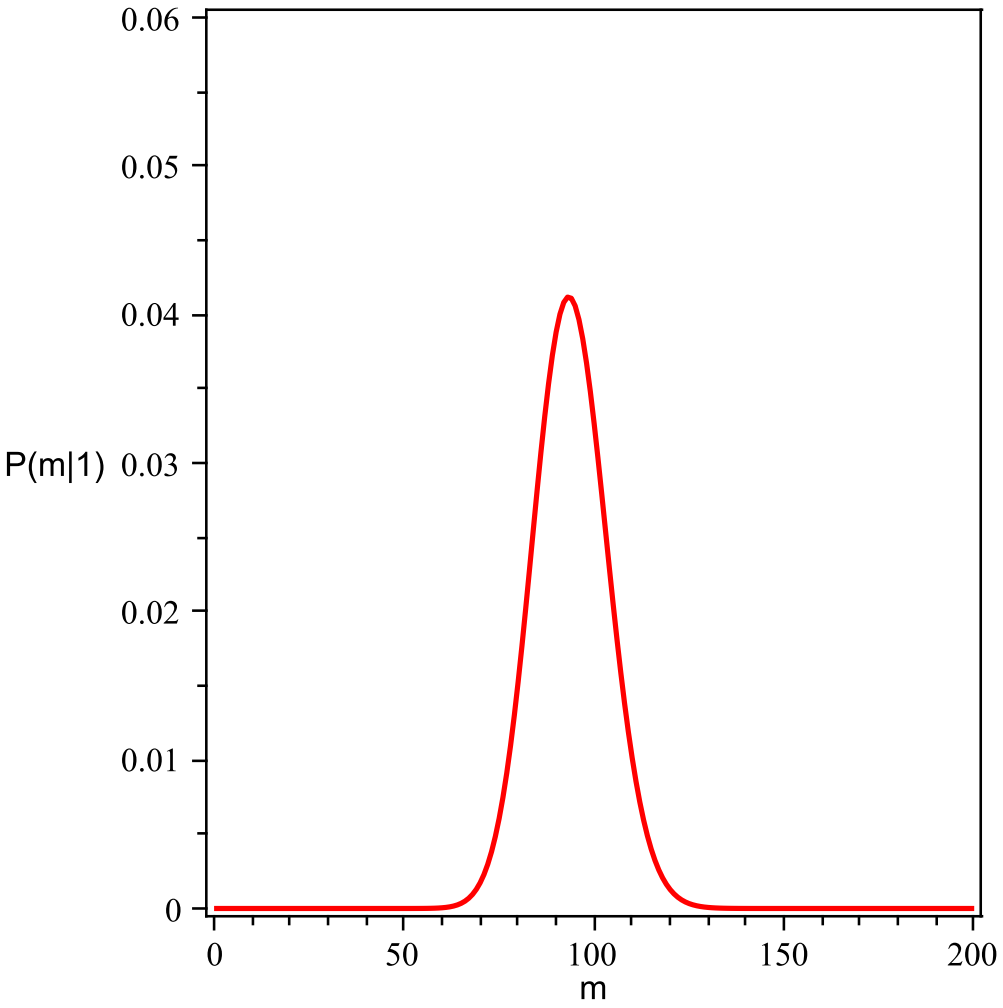

Figure 2.5: The conditional probability $P(m|1)$, with parameters $K_{eq} = 1.67, \alpha = 150$, and $\varepsilon = 72000$ (as in [30]).

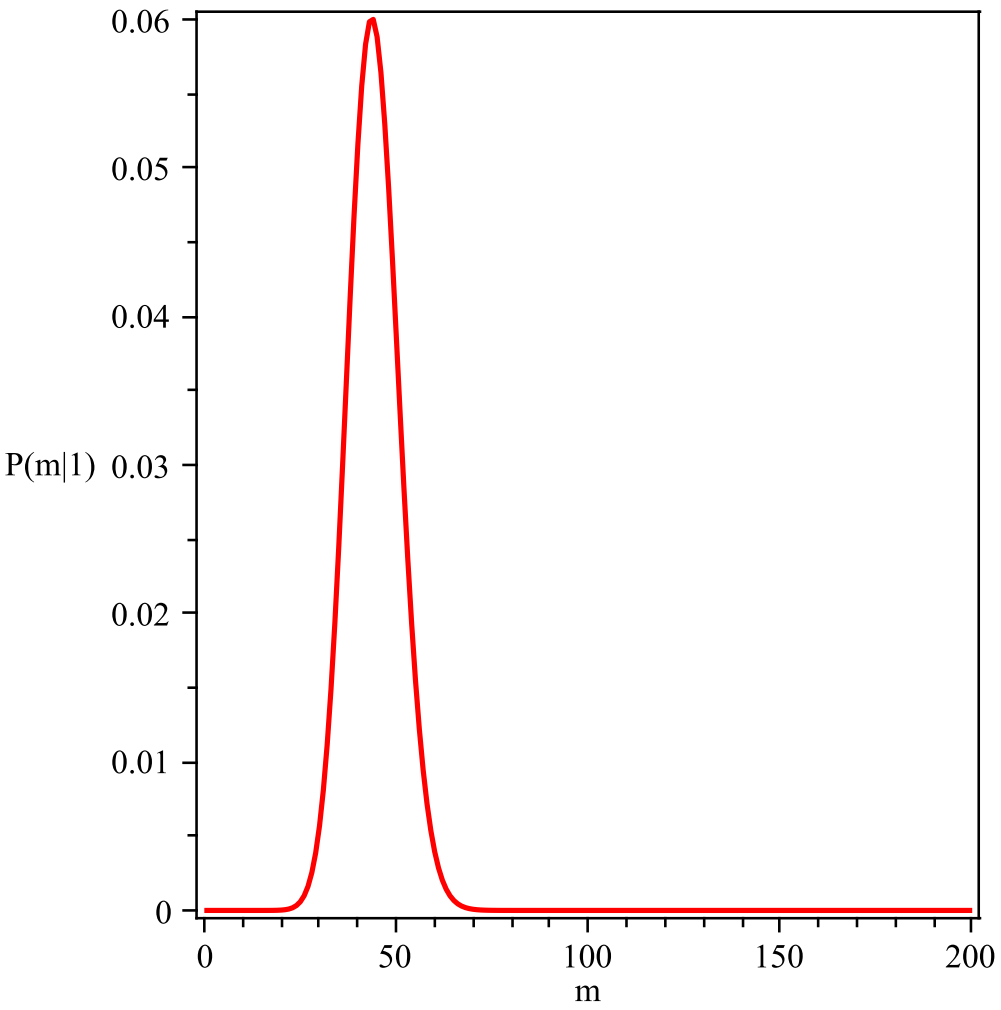

Figure 2.6: The effect of decreasing K_{eq} on the conditional probability $P(m|1)$, parameters $K_{eq} = 0.42, \alpha = 150$, and $\varepsilon = 72000$ (as in [30]).

Chapter 3

Extension of Metabolic Control Analysis to genetic systems

3.1 The stochastic control coefficients

Explanations on how cells operate within strict and safe parameters in the presence of noise and how cells can use noise, for instance to amplify certain responses can be forged once one has established the expressions for the moments of the distributions for the biomolecules. These issues can be addressed by using sensitivity analysis. Specifically, Metabolic Control Analysis (MCA), is a form of sensitivity analysis adapted to biochemical processes [37]. It uses control coefficients to assess relative changes in any observable when parameters change. These are mathematically defined by

$$C_{\rho}^{X} = \frac{\partial \ln X}{\partial \ln \rho} \approx \frac{\Delta \ln X}{\Delta \ln \rho}, \tag{3.1}$$

where X is an observable and ρ, a parameter. In our model diagram, the observables include the mean and standard deviation for the marginal and conditional probability distributions of the mRNA, the skewness and the kurtosis, and the parameters, ρ, could be either of α, ε or K_{eq}. The analytical expressions for the moments are in terms of these parameters and the control coefficients are also calculated for these three parameters, α, ε and K_{eq}. K_{eq} changes when, for instance, the strength of the promoter is changed by manipulating its DNA sequence, a central technique in genetic engineering. Experimentally, we can measure the change in X for a given change in ρ, which allows us to relate theory to experiment, through the approximation of C_{ρ}^{X} using finite differences shown above.

Using the moments of the marginal and conditional distributions, we get the expressions for the control coefficients for the respective averages and standard deviations. Some of

the expressions for the control coefficients are quite unmanageable and too long to be presentable. Instead numerical values for all the control coefficients are obtained using Maple, see Table 3.2, and a few selected expressions are tabulated in Table 3.3. The values of the parameters were obtained by substituting the values of the rate constants in Table 3.1 below.

Table 3.1: Rate constants in the model and their values

Parameter	Value	Units	Source
k_{on}	0.002	s^{-1}	[40]
k_{off}	0.006	s^{-1}	[40]
k_m	0.0067	s^{-1}	[40]
k_m^-	0.001	s^{-1}	[40]

In Table 3.2 below, we see positive and negative control coefficients. A positive control coefficient tells us that the observable increases as the parameter increases, while a negative control coefficient tells us that the observable decreases as the parameter increases. A value of zero for a control coefficient shows that the observable will not respond in any way to changes in the corresponding parameter.

The values of the control coefficients in Table 3.2 below give us an idea or help us to explain how cells maintain behaviour or control stochasticity. The cell can adjust or "tune" a certain observable without interfering with another observable. As an example, since the

Table 3.2: The observables and the respective numerical values for the control coefficients, for $\alpha = 150, K_{eq} = 1.2, \varepsilon = 72000$ (as in [30]).

Observable (X)	C^X_α	C^X_ε	$C^X_{K_{eq}}$
γ_1	-0.243	-0.327	-0.337
γ_1^1	-0.288	-0.009	-0.353
γ_1^0	-0.236	-0.410	-0.273
γ_2	-0.528	0.579	-0.821
γ_2^1	-0.648	0.139	-0.880
γ_2^0	-0.487	-0.806	-0.641
$\langle m \rangle$	1	0	0.750
$\langle m^1 \rangle_1$	1	-0.333	0.667
$\langle m^1 \rangle_0$	1	0.167	0.667
σ	0.679	-0.239	0.450
σ_1	0.654	-0.326	0.375
σ_0	0.671	-0.130	0.403

value of $C_{\varepsilon}^{\langle m \rangle}$ is zero while $C_{\varepsilon}^{\sigma} = -0.239$, the cell can adjust the standard deviation of the number of mRNA molecules without interfering with the mean by changing ε. In fact, an increase in the parameter ε is followed by a decrease in the standard deviation, σ. Such a response has an effect on the CV, which is given by the ratio $\frac{\sigma}{\langle m \rangle}$. Keeping $\langle m \rangle$ fixed and decreasing σ means a decrease in the CV or noise. Thus, noise can be regulated in the cell by concentrating on certain parameters like ε which affect the observables directly linked to noise.

Results in Table 3.2 also confirm the changes displayed in Figs. 2.5 and 2.6 since $C_{K_{eq}}^{\langle m^1 \rangle_1} = 0.667$, which tells us that the mean number of mRNA increases as K_{eq} increases, hence the observed shift to the right by the distribution. In contrast, $C_{K_{eq}}^{\gamma_2} = -0.880$, implying a decrease in the corresponding excess kurtosis as the same parameter increases. The distribution thus becomes flatter and wider at a higher value of K_{eq}. Therefore, increasing K_{eq} reduces the excess kurtosis but increases the standard deviations, yet decreasing K_{eq} increases the excess kurtosis and reduces the standard deviation. Correspondingly, Figs 2.5 is flatter and wider as compared to Fig 2.6 which is peaked and narrow.

In Figs 3.1 and 3.2 below, we plot the control coefficients $C_{K_{eq}}^{\sigma}$ and $C_{K_{eq}}^{\gamma_2}$, respectively, against K_{eq}, to have a general idea of how the response posed by these variables changes as the parameter K_{eq} changes. These two control coefficients were arbitrarily chosen after observing that the rest of the control coefficients defining the responses of the other remaining observables to changes in the parameter K_{eq} were similar to either of these two.

Table 3.3: Expressions for some of the control coefficients for the average and the standard deviations of the marginal and conditional probabilities

C_ρ^X	Expression
$C_\varepsilon^{\langle m\rangle}$	0
$C_\alpha^{\langle m\rangle}$, $\forall s\, C_\alpha^{\langle m^1\rangle_s}$	1
$\forall s\, C_{K_{eq}}^{\langle m^1\rangle_s}$	$\frac{\varepsilon-1}{\varepsilon K_{eq}+\varepsilon-1}$
$C_{K_{eq}}^{\langle m\rangle}$	$\frac{1}{1+K_{eq}}$
$C_\varepsilon^{\langle m^1\rangle_1}$	$\frac{-1}{\varepsilon K_{eq}+\varepsilon-1}$
$C_\varepsilon^{\langle m^1\rangle_0}$	$\frac{\varepsilon K_{eq}}{(\varepsilon-1)(\varepsilon K_{eq}+\varepsilon-1)}$
C_ε^σ	$\frac{-\alpha\varepsilon K_{eq}(1+K_{eq})}{2(\varepsilon K_{eq}+\varepsilon-1)(\varepsilon(1+K_{eq})^2+K_{eq}(\alpha-1)-1)}$
C_α^σ	$\frac{1}{2}\left[1+\frac{\alpha K_{eq}}{\varepsilon(K_{eq}+1)^2-K_{eq}(1-\alpha)-1}\right]$
$C_{K_{eq}}^\sigma$	$\frac{K_{eq}(3\varepsilon^2(1+K_{eq}+\varepsilon K_{eq}^2(\varepsilon-\alpha)-\varepsilon(2-\alpha)(2+K_{eq})-(2\alpha-1)))+(\varepsilon-1)^2}{2(1+K_{eq})(\varepsilon(1+K_{eq})^2)+K_{eq}(\alpha-1)-1}$

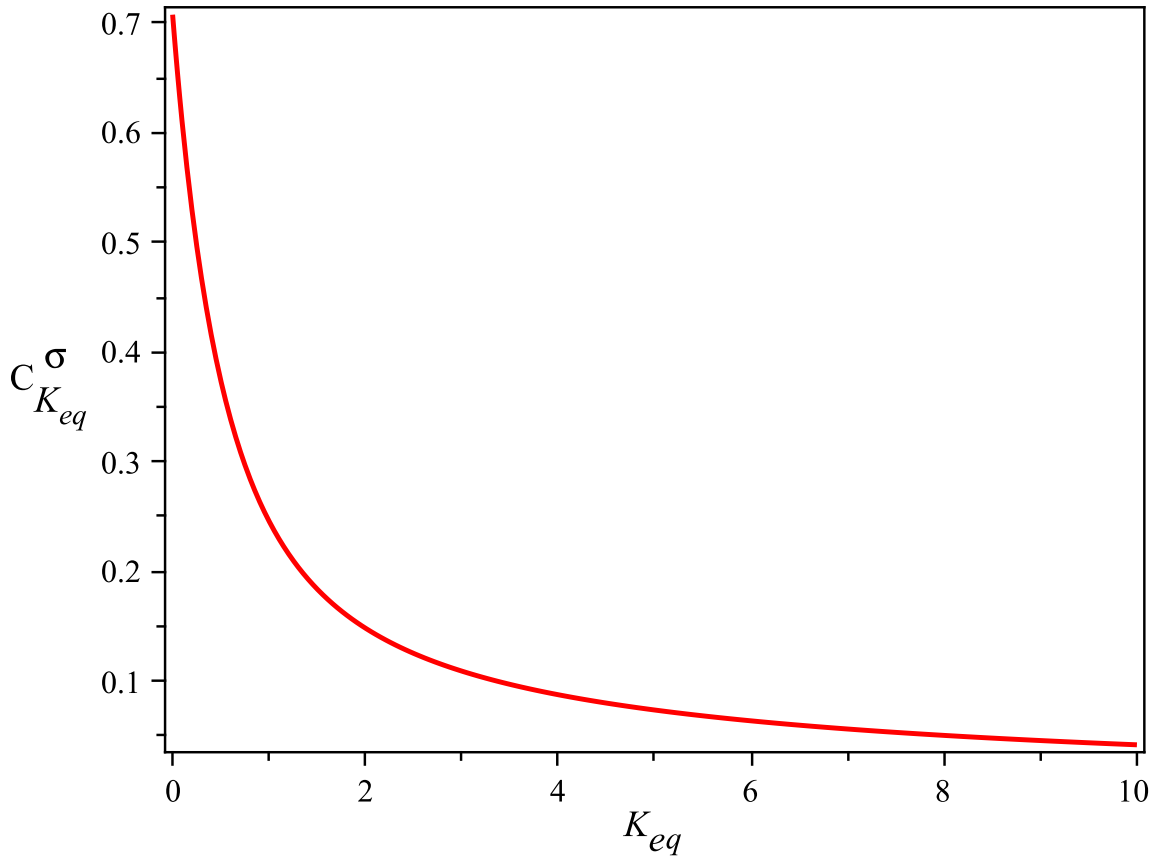

Figure 3.1: Behavior of the control coefficient, $C^{\sigma}_{K_{eq}}$ as K_{eq} varies, while $\alpha = 6.7$ and $\varepsilon = 3$ (as in [40]).

Fig. 3.1 shows that the dependence of the observable σ on the parameter K_{eq} reduces to negligible values as K_{eq} increases. In fact as $K_{eq} \to \infty$, $C^{\sigma}_{K_{eq}} \to 0$, while as $K_{eq} \to 0$, $C^{\sigma}_{K_{eq}} \to 0.710$. A similar behavior was observed for the relationship between the control coefficients $C^{\sigma_0}_{K_{eq}}, C^{\sigma_1}_{K_{eq}}, C^{\langle m^1 \rangle_0}_{K_{eq}}, C^{\langle m^1 \rangle_1}_{K_{eq}}$ and the parameter K_{eq}. In relation to Fig. 2.2 (a plot of σ vs K_{eq}), this graph indicates exactly what one would expect to observe. As σ gets large enough we will not see a change by making it bigger. As more and more mRNA molecules are produced, the less responsive the observable(s) become. Fluctuations are minimized whenever the numbers of molecules are large.

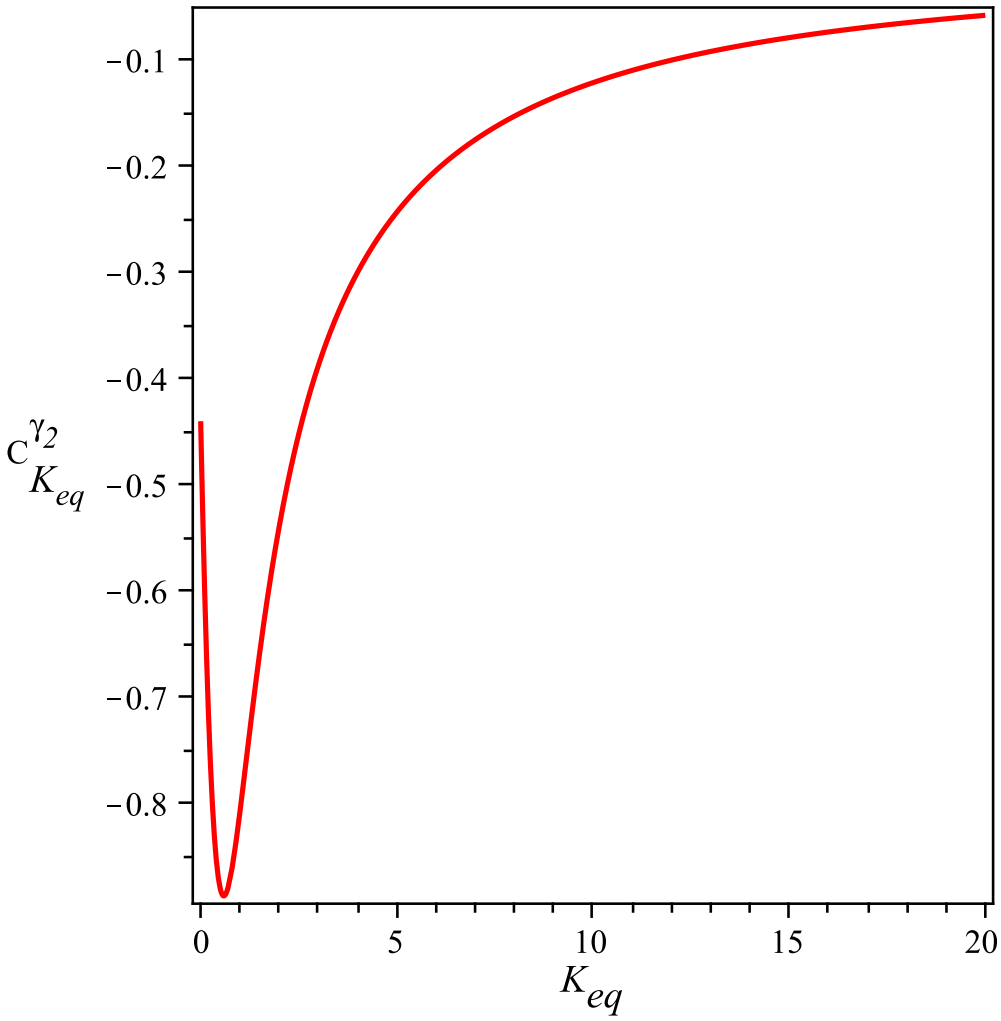

Figure 3.2: Relationship between $C^{\gamma_2}_{K_{eq}}$ and K_{eq}, for $\alpha = 6.7$ and $\varepsilon = 3$ (as in [40]).

The control coefficient $C^{\gamma_2}_{K_{eq}}$ decreases to a minimum value of -0.885, which occurs when $K_{eq} = 0.539$ (Fig. 3.2). At this value of the parameter K_{eq}, while the sensitivity stays negative, the observable shifts from increasingly sensitive as K_{eq} increases to being less sensitive. A scrutiny of the end-behaviors of this graph shows that as $K_{eq} \to \infty$, $C^{\gamma_2}_{K_{eq}} \to 0$, while $C^{\gamma_2}_{K_{eq}} \to -0.434$ as $K_{eq} \to 0$. A similar behavior as depicted on the graph was observed when the control coefficients $C^{\gamma_1}_{K_{eq}}, C^{\gamma_1^s}_{K_{eq}}, C^{\gamma_2^s}_{K_{eq}}$ $(s = 0, 1)$, were plotted against K_{eq}.

All the observables with positive control coefficients for the parameter K_{eq} increase as K_{eq} increases. Table 3.2 shows these observables to be the measures of central tendency and the measures of dispersion for both the marginal and conditional probabilities. The kurtosis and the skewness have negative control coefficients for the parameter K_{eq}, and, as we would expect, decrease as the parameter increases (Fig. 3.1). Figs. 3.1 and 3.2 above help us partition the set of all the control coefficients seen in Table 3.2 into two sets, in accordance with whether they increase or decrease as K_{eq} changes. The set comprising the measures of central tendency and the measures of dispersion of the marginal and conditional distributions obeys the behavior depicted in Fig. 3.1. The second set, consisting of the skewness and the kurtosis of the marginal and conditional distributions behaves as shown in Fig. 3.2.

Perhaps, henceforth, the reader has to be made aware of one of the reasons for doing such a project. This work aims at developing tools that can be used by other researchers, mostly biologists. Synthetic biologists, for example, produce various copies of genes for various purposes or targets. Some of these targets include different performance levels measured by the amount of output (gene products). The ultimate aspiration of all such scientists is to be able to inflict changes to the natural phenomenon exhibited by systems, i.e. controlling the system. This work therefore serves the purpose of highlighting the key points leading to a special type of behavior and how it may be achieved. The partitioning of the responses according to Figs. 3.1 and 3.2 is one form of help one gets from this study. We have partitioned a set of twelve observables into two sets depending on how they respond to a particular parameter, K_{eq} in this case. This makes it easy to foretell certain kinds of behavior before experiments are carried out, thereby saving time and resources.

3.2 Summation theorems

In my brief discussion of the Theory of Metabolic Control Analysis (MCA), last section of Chapter 1, I mentioned that this theory is based on control coefficients. A particular metabolism can be influenced by a number of enzymes meaning that the overall output, i.e. the distribution of an observable, in a metabolic pathway is affected by activities of at least one enzyme. In simple words, we say that control over the distribution of the metabolite is shared. One of the summation theorems says that the sum of all the concentration control coefficients for the enzymes affecting a particular metabolism should add up, and the sum is 0 [44]. Thus, there exists a negative response of equal magnitude somewhere along the pathway by a certain enzyme meant to cancel or negate a positive effect on the metabolite distribution by another enzyme, so that the sum of zero is maintained.

Let C_{ω}^{X} be the control coefficient measuring the response of the observable X to a change in a rate constant $\omega = \{k_m, k_m^-, k_{\text{on}}, k_{\text{off}}\}$. Using the chain rule one can easily establish that

$$C_{\omega}^{X} = \frac{\omega}{X}\sum_{\rho}\frac{\partial X}{\partial \rho}\frac{\partial \rho}{\partial \omega} = \sum_{\rho} C_{\rho}^{X}\frac{\omega}{\rho}\frac{\partial \rho}{\partial \omega}$$

where, as before, $\rho = \{\alpha = \frac{k_m}{k_m^-}, \varepsilon = 1 + \frac{k_{\text{on}}}{k_m^-}, K_{eq} = \frac{k_{\text{on}}}{k_{\text{off}}}\}$. Hence,

$$\sum_{\omega} C_{\omega}^{X} = \sum_{\omega}\left[\sum_{\rho} C_{\rho}^{X}\frac{\omega}{\rho}\frac{\partial \rho}{\partial \omega}\right]$$

$$= \sum_{\omega}\left[C_{\alpha}^{X}\frac{\omega}{\alpha}\frac{\partial \alpha}{\partial \omega} + C_{\varepsilon}^{X}\frac{\omega}{\varepsilon}\frac{\partial \varepsilon}{\partial \omega} + C_{K_{eq}}^{X}\frac{\omega}{K_{eq}}\frac{\partial K_{eq}}{\partial \omega}\right]$$

$$
\begin{aligned}
&= C_{\alpha}^{X}\left[\frac{k_m}{\alpha}\frac{\partial\alpha}{\partial k_m}+\frac{k_m^-}{\alpha}\frac{\partial\alpha}{\partial k_m^-}+\frac{k_{\text{on}}}{\alpha}\frac{\partial\alpha}{\partial k_{\text{on}}}+\frac{k_{\text{off}}}{\alpha}\frac{\partial\alpha}{\partial k_{\text{off}}}\right]\\
&+ C_{\varepsilon}^{X}\left[\frac{k_m}{\varepsilon}\frac{\partial\varepsilon}{\partial k_m}+\frac{k_m^-}{\varepsilon}\frac{\partial\varepsilon}{\partial k_m^-}+\frac{k_{\text{on}}}{\varepsilon}\frac{\partial\varepsilon}{\partial k_{\text{on}}}+\frac{k_{\text{off}}}{\varepsilon}\frac{\partial\varepsilon}{\partial k_{\text{off}}}\right]\\
&+ C_{K_{eq}}^{X}\left[\frac{k_m}{K_{eq}}\frac{\partial K_{eq}}{\partial k_m}+\frac{k_m^-}{K_{eq}}\frac{\partial K_{eq}}{\partial k_m^-}+\frac{k_{\text{on}}}{K_{eq}}\frac{\partial K_{eq}}{\partial k_{\text{on}}}+\frac{k_{\text{off}}}{K_{eq}}\frac{\partial K_{eq}}{\partial k_{\text{off}}}\right]\\
&= C_{\alpha}^{X}\left[1-1+0+0\right]+C_{\varepsilon}^{X}\left[0+\frac{1}{\varepsilon}-1+1-\frac{1}{\varepsilon}+0\right]\\
&+ C_{K_{eq}}^{X}\left[0+0+1-1\right]\\
&\Rightarrow \sum_{\omega} C_{\omega}^{X}=0.
\end{aligned}
$$

Hence, regardless of the observable, X, chosen, the sum (inside the square brackets) remains zero since the parametric sensitivities, $\frac{\omega}{\rho}\frac{\partial\rho}{\partial\omega}$, will sum to zero, therefore a summation theorem holds for any observable one may choose. This means that the Metabolic Control Analysis theory can be extended to genetic control mechanisms. Gene expression is stochastic, owing to the number of genes themselves, as well as some other factors involved in the expression process, for example the transcription factors, but still there is control over the process. As an example, an increase in the rate of one reaction, be it transcription or mRNA degradation, is matched by an increase or decrease of another reaction meant to negate the effect so that the cell continues to function normally without having an excess of a certain type of molecule relative to the others.

As illustrated before when we discussed the control coefficients, $C_{\varepsilon}^{\langle m\rangle}=0$ and $C_{\varepsilon}^{\sigma}=-0.239$, gene activation and mRNA dynamics can be utilized antagonistically in the control of noise. The parameter $\varepsilon=1+\frac{k_{\text{on}}}{k_m^-}$ links mRNA dynamics to the activities on the promoter. Thus, for instance, by increasing the promoter activity, which increases this parameter, the gene has

to keep the rate at which mRNA degrades constant so that noise is decreased. Conversely, if mRNA degradation is high, then noise is high since the parameter ε becomes small. To lessen the noise, the gene has to increase the activity of the promoter, which increases the number of mRNA molecules.

This chapter concentrated on the response by the key qualities (moments), that describe the distributions, to changes in parameters. The results, mainly the control coefficients, have been given both in numerical form, and, for some, in terms of expressions. We can also explore the dependence of the distributions on the parameters. This is done in the following chapter.

Chapter 4

The Stochastic Bifurcations

4.1 Dependence of the conditional probabilities on the parameter ε

In this chapter, we look at how the distributions themselves, not particular aspects or qualities in them, would respond to changes in parameters. In particular, and for interest's sake, we will study how the distributions behave when ε, which links the dynamics of the mRNA to the dynamics of the promoter, changes.

We begin by seeing what the conditional probability distributions, as well as the marginal distributions, look like when plotted using selected values of the parameters, and then how they look like as we change one of the parameters. The results are presented in graphical form. The last section of this chapter contains results of a numerical simulation of the same model, which are also presented in graphical form. This does not only enable a comparison of the results obtained from the model using the two different approaches, but also presents us with the opportunity to compare our results with simulation results obtained from a similar model studied by other researchers [68].

Using the parameters from [30], we first plot the conditional probability distribution function, $P(m|1)$, which we derived in Chapter 2 Fig. 4.1. We arbitrarily chose to display $P(m|1)$; $P(m|0)$ is identical and symmetrical for these parameters. Our aim here is to spot the effects of altering the values of the parameters, e.g. ε, on these distributions. Given the three parameters ε, α and K_{eq}, one may be interested in the selection of particular parameters that influence a particular observable most. Qualitative changes in probability

distributions as parameters change, called stochastic bifurcations, can be investigated by examining the changes in the probability distributions as parameters change.

To get a rough idea about how (or even whether) the distributions respond to changes in parameters, we fix values of two of the parameters, namely α and K_{eq}, and then vary ε. As discussed in Chapter 2, the parameter ε links the time scales of the two subsystems, one regarding mRNA dynamics and represented by α, and the other, standing for the dynamics on the promoter and represented by K_{eq}. Results from the preceding chapter show the effect that an increase in the parameter ε had on the standard deviation, i.e., an increase in this parameter results in a decrease in the standard deviation. As can be observed upon reducing ε (from Fig. 4.1 through Fig. 4.3 below), the distributions shift from being symmetrical to being non-symmetrical. $P(m|0)$ eventually ceases having a maximum and becomes monotonic while $P(m|1)$ still has a maximum but is more flattened and negatively skewed. This indicates a qualitative behavioral change by the system.

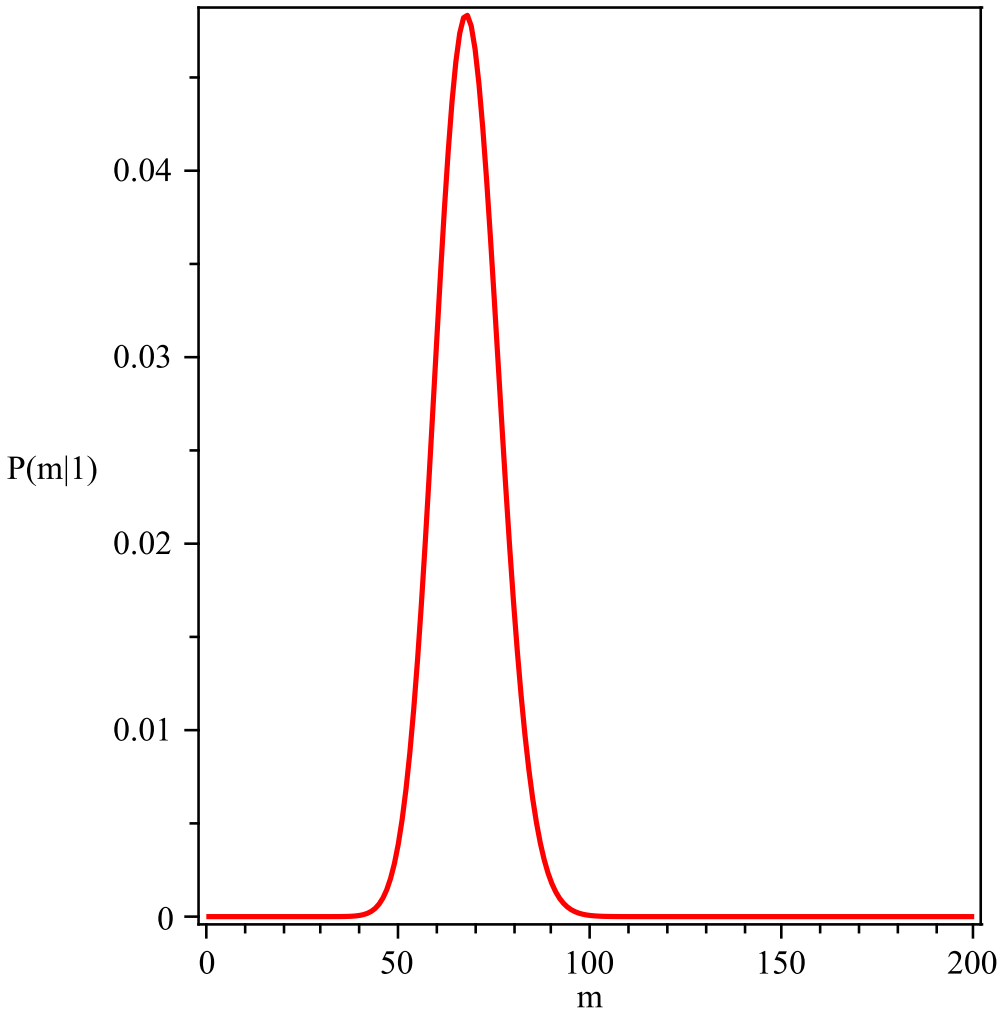

Figure 4.1: The conditional probability, $P(m|1)$, of having m mRNA molecules given that the gene is in an active state, for $\alpha = 150, \varepsilon = 72000, K_{eq} = 1.2$ (as given in [30]).

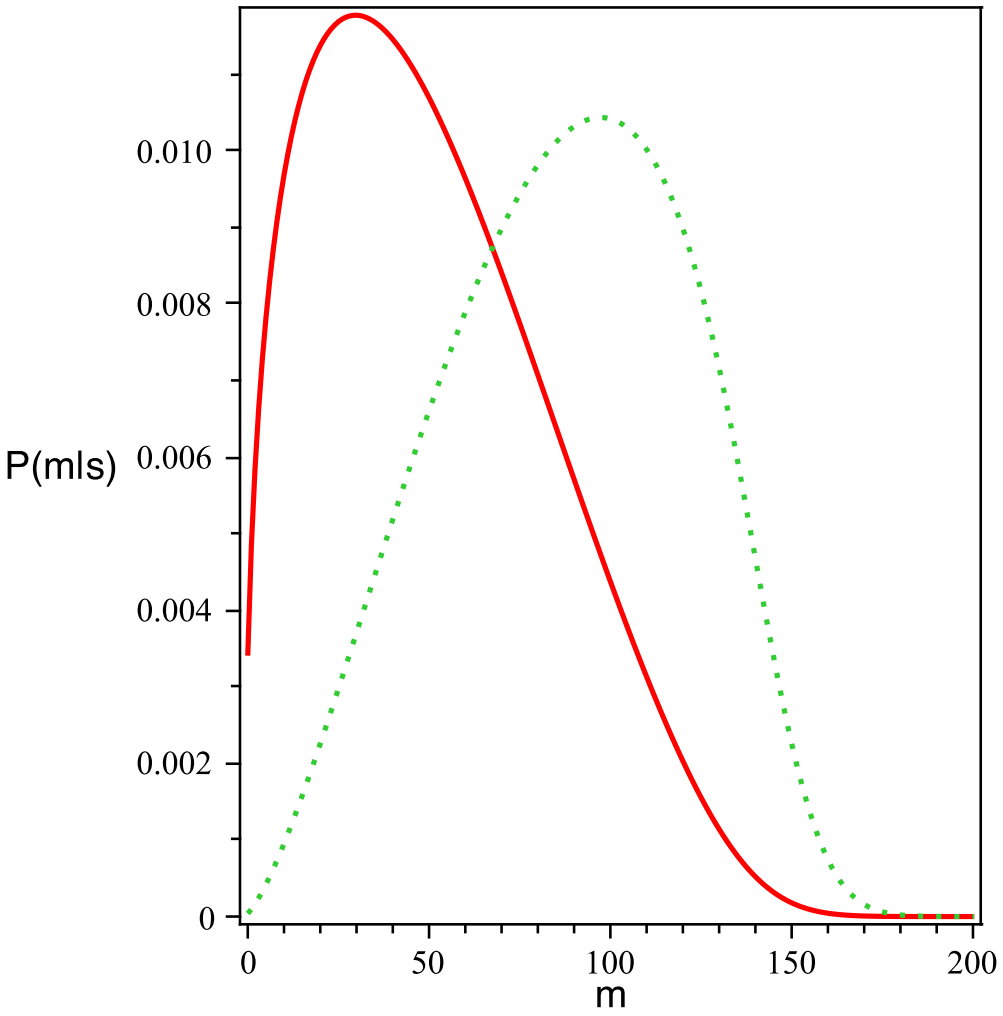

Figure 4.2: A change in the shape of $P(m|s)$ [$P(m|0)$ continuous and $P(m|1)$ dashed] relative to Fig. 4.1 when the value of ε is reduced from 72000 to 2.44, with $\alpha = 150$ and $K_{eq} = 1.2$ (values of α and K_{eq} are as given in [30]).

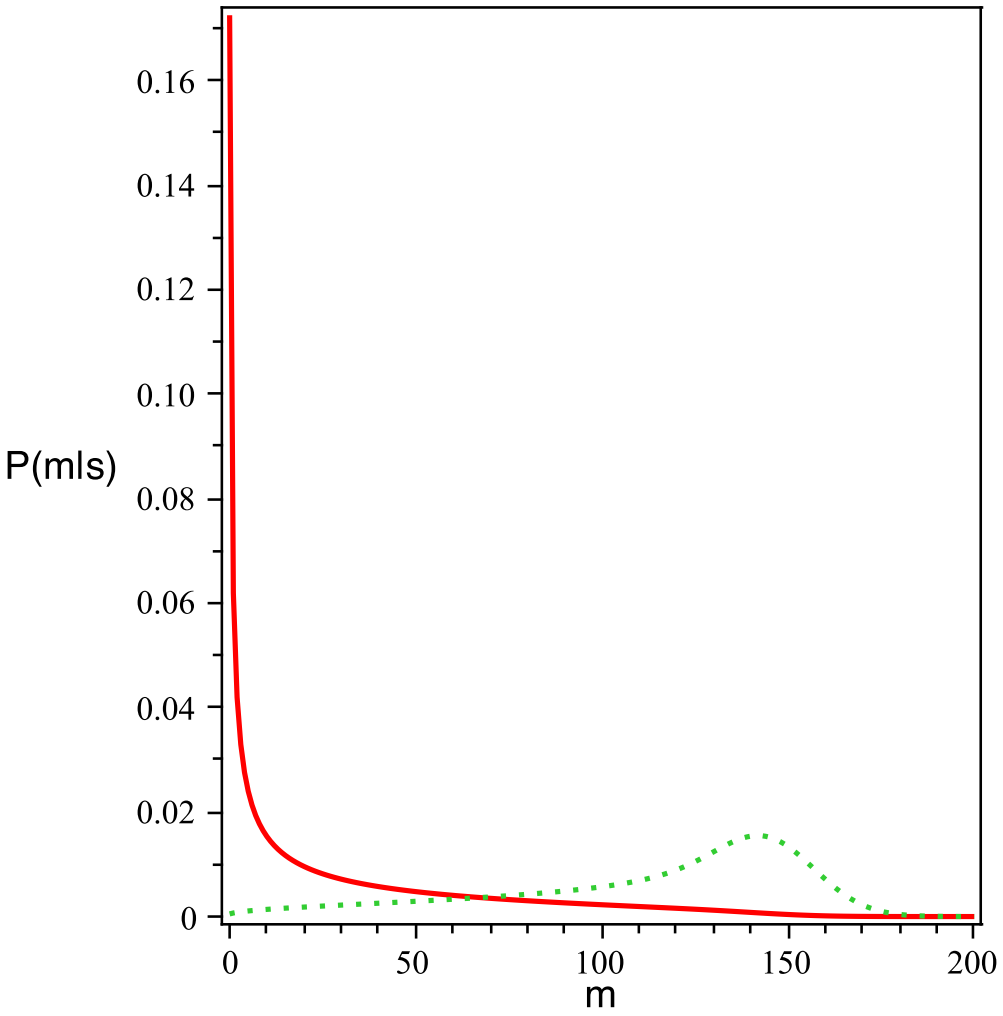

Figure 4.3: An illustration of the changes in behavior relative to Figs. 4.1 and of the conditional probability distributions, $P(m|s)$ [$P(m|0)$ continuous and $P(m|1)$ dashed] when $\varepsilon = 1.36$, $\alpha = 150$, $K_{eq} = 1.2$ (values of α and K_{eq} are as given in [30]).

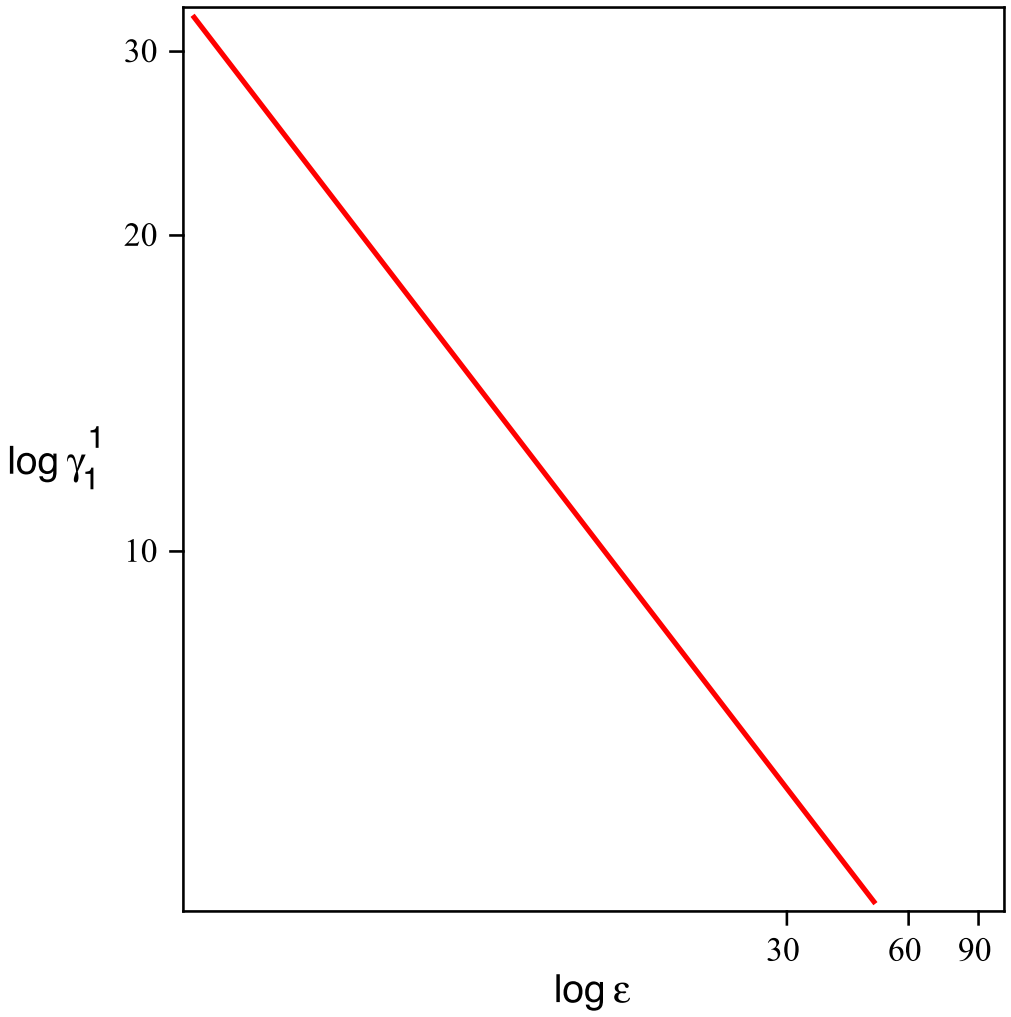

Figure 4.4: Logarithmic plot of the skewness, for the active distribution, γ_1^1, against ε, with $\alpha = 150$ and $K_{eq} = 1.2$ (as given in [30]).

Fig. 4.4 shows that as the parameter ε increases, the skewness of $P(m|1)$ decreases. This can also be predicted from Table 3.2, where we have $C_\varepsilon^{\gamma_1^1} = -0.009$, implying that the distribution $P(m|1)$ becomes less skewed (more symmetrical) as ε increases, as seen in Fig. 4.1.

4.2 The marginal distribution and its response to changes in the parameter ε

In Figs. 4.1 through Fig. 4.3 above, we have seen the graphical representations of the conditional probabilities plotted using different values for the parameter ε, while the other two parameters, α and K_{eq} remain fixed. In this section we explore, in a similar manner, the changes that the marginal distribution undergoes. This should help us reveal the overall effect on the distribution of the state-to-state transitions by the promoter. The marginal distribution can be easily obtained experimentally by sampling from a population of non-synchronized cells.

The marginal probability is a sum of the two joint probabilities, $P(0,m)$ and $P(1,m)$, $m \in [0,\infty)$, and hence as one would anticipate, should respond to changes in parameters. Fig. 4.5 below shows, for some selected values of the parameters α, ε and K_{eq}, what $P(m)$ looks like, whereas the subsequent figures, Fig. 4.6 through Fig. 4.8, illustrate the changes that this same distribution undergoes when ε is changed (reduced). As was the case with the conditional distribution $P(m|0)$, the marginal distribution is observed to shift from being symmetrical to non-symmetrical. As we noted in Chapter 3, decreasing ε leads to an increase in the standard deviation, meaning that the distributions become broader: variability increases. In Fig. 4.8 we observe the marginal probability distribution also changing from having a global maximum at some value of $m > 0$, as observed in Fig. 4.1 through Fig. 4.3, to having a local maximum at some $m > 0$, and the global maximum at $m = 0$. This is a clear indication of the fact that a change in parameters causes the system to undergo stochastic bifurcations. Unlike in the deterministic case where bifurcations are recognized through the degeneration of stable fixed points to either unstable or double stable points, bifurcations in stochastic systems are noted differently. The notion of a stable fixed point

is not well-defined in stochastic systems, instead a bifurcation occurs in stochastic systems whenever a qualitative change occurs in the shape of the distribution. For instance, the appearance of additional peaks or a change in the sign of the skewness both indicate a bifurcation.

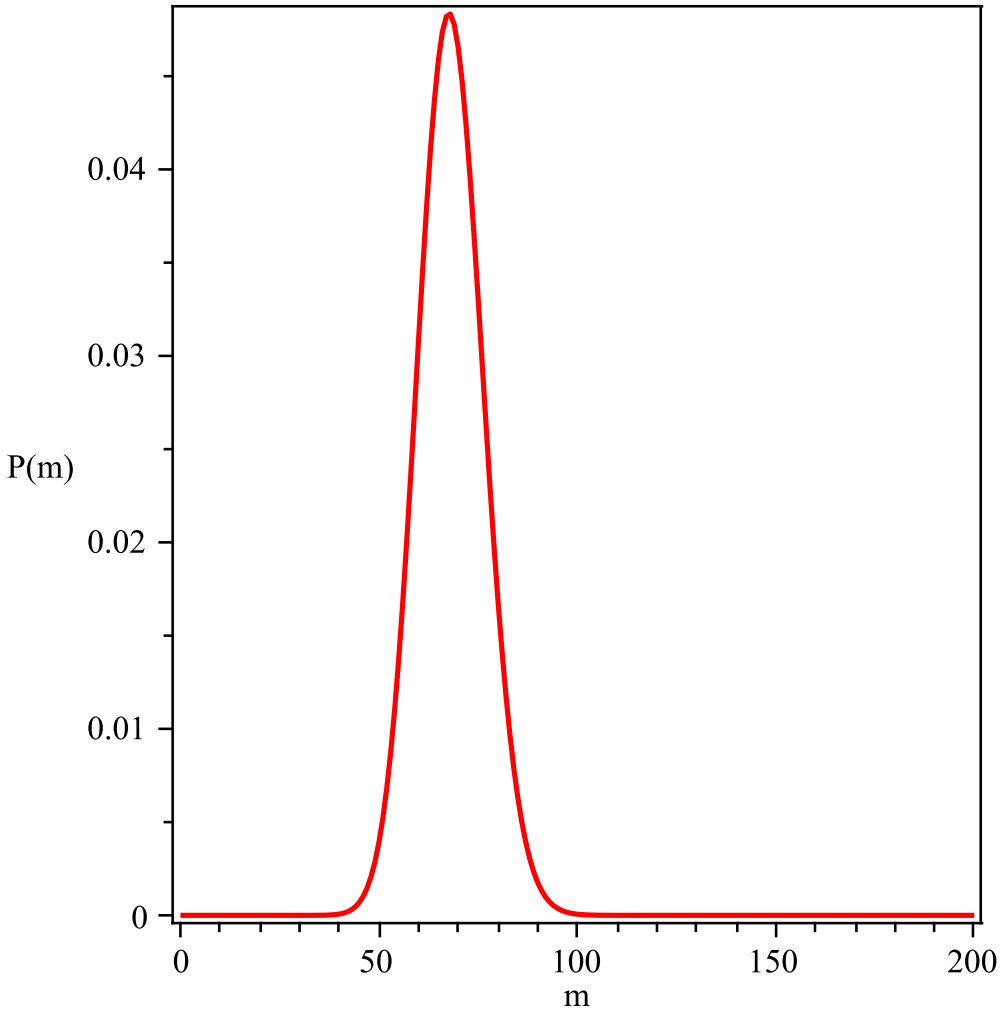

Figure 4.5: A plot of the marginal probability distribution for the values of the parameters $\alpha = 150$, $\varepsilon = 72000$ and $K_{eq} = 1.2$ (as in [30]).

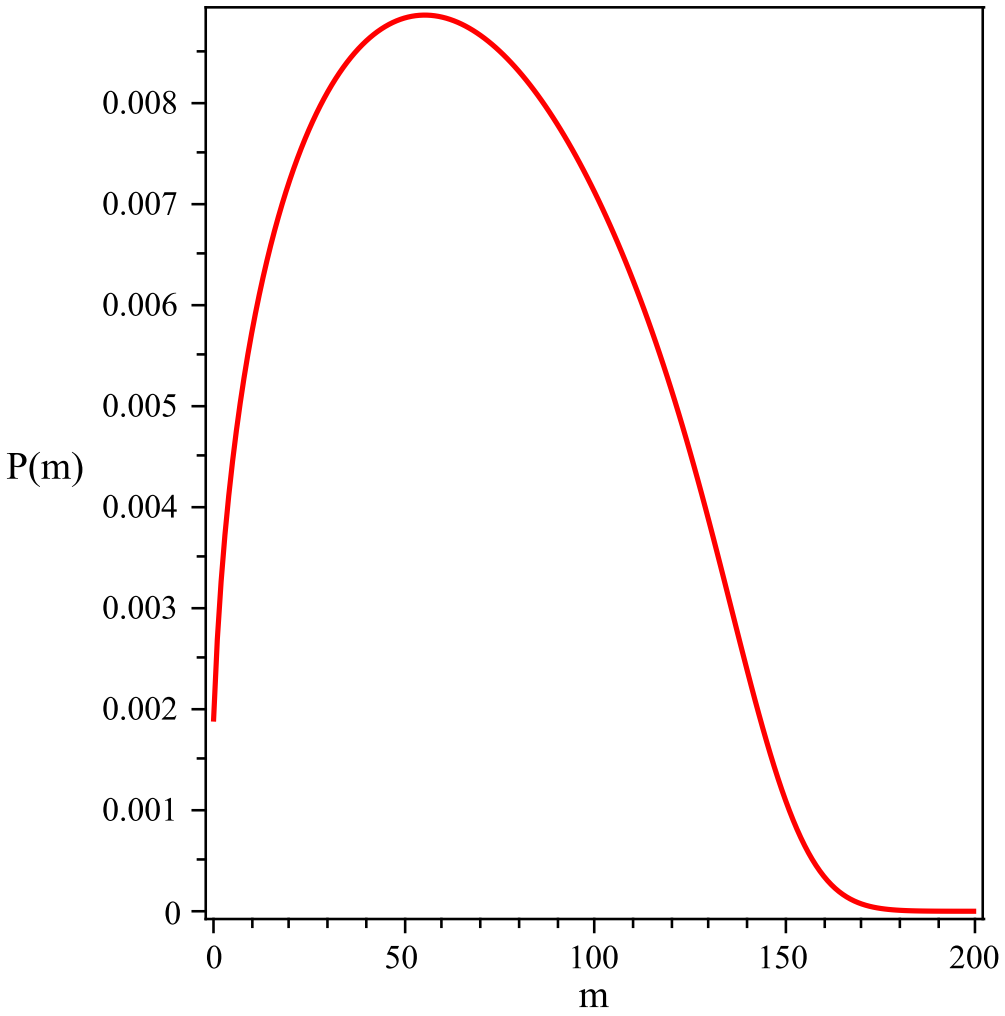

Figure 4.6: A plot of the marginal probability distribution for the values of the parameters $\varepsilon = 2.44$, $\alpha = 150$ and $K_{eq} = 1.2$ (values of α and K_{eq} are as given in [30]).

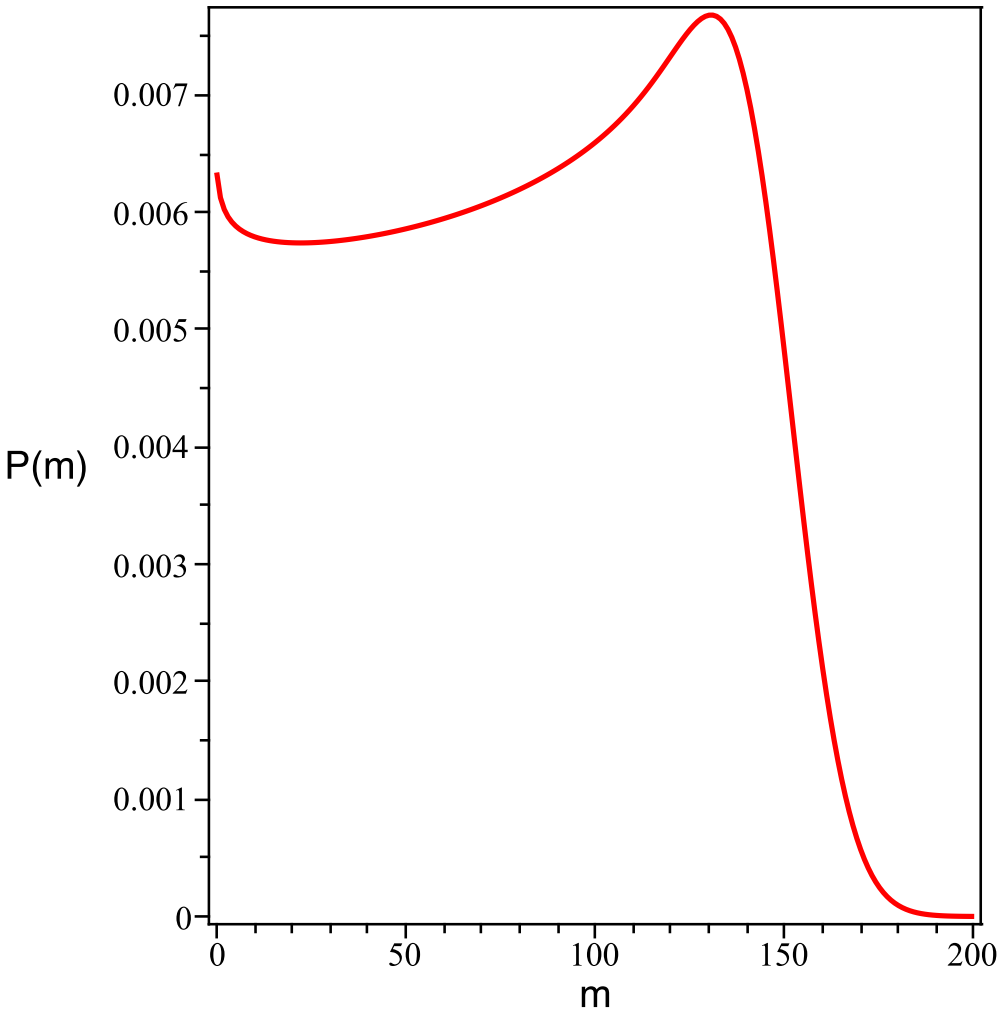

Figure 4.7: A plot of the marginal probability distribution for the values of the parameters $\varepsilon = 1.96$, $\alpha = 150$ and $K_{eq} = 1.2$ (values of α and K_{eq} are as given in [30]).

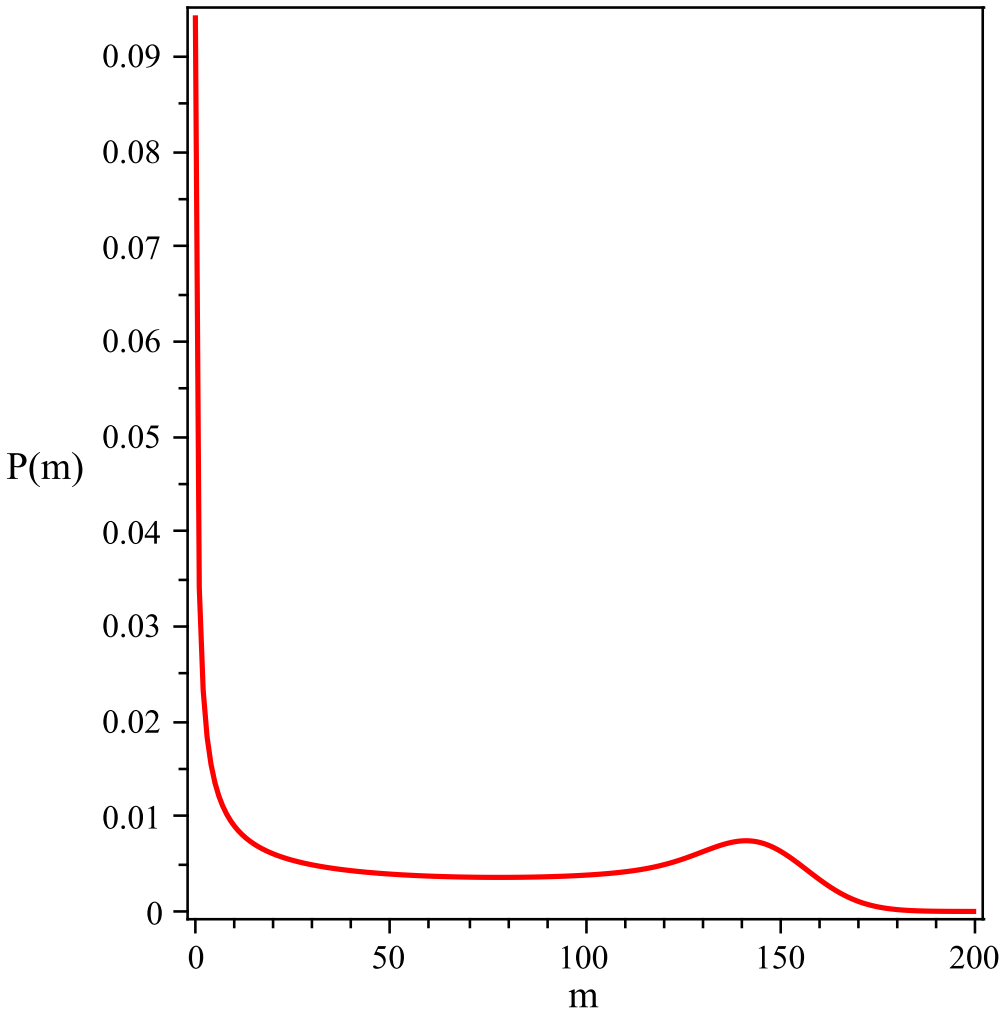

Figure 4.8: A plot of the marginal probability distribution for the values of the parameters $\varepsilon = 1.36$, $\alpha = 150$ and $K_{eq} = 1.2$ (values of α and K_{eq} are as given in [30]).

4.3 Derivation of the bifurcation analysis equation for the conditional probability $P(m|0)$

The bifurcation we observed above is mainly due to a change of the probability distribution of $P(m|0)$ from having a maximum at some value of $m > 0$, to having a maximum at the point $m = 0$. This stochastic bifurcation occurs when the rate of change of $P(m|0)$ with respect to m, evaluated at $m = 0$, equals zero. Expressed, mathematically, this is

$$\left.\frac{dP(m|0)}{dm}\right|_{m=0} = 0.$$

This equation defines a surface, in parameter space, on one side of which the distribution has a maximum, while on the other side, it is monotonic. After differentiation of $P(m|0)$, we find that

$$\frac{dP(m|0)}{dm} = \beta\left[\left(\ln\alpha - \Psi(m+1) + \Psi(\varepsilon+m) - \Psi(\psi+m)\right)\Omega + \frac{d\Omega}{dm}\right],$$

where,

$$\beta = \frac{1}{\psi-\varepsilon}\frac{\alpha^m}{m!}\frac{(\varepsilon)_m}{(\psi)_m},$$

$$\Omega = (\alpha+\psi+m-\varepsilon)\ {}_1F_1(\varepsilon+m;\psi+m;u_0) - \alpha\frac{\varepsilon+m}{\psi+m}\ {}_1F_1(\varepsilon+m+1;\psi+m+1;u_0)$$

$$-m\frac{\psi+m-1}{\varepsilon+m-1}\ {}_1F_1(\varepsilon+m-1;\psi+m-1;u_0),$$

$$\Rightarrow \frac{d\Omega}{dm} = {}_1F_1(\varepsilon+m;\psi+m;u_0)$$

$$+(\alpha+\psi+m-\varepsilon)\sum_{n=0}^{\infty}\frac{(\varepsilon+m)_n}{(\psi+m)_n}\frac{u_0^n}{n!}\sum_{p=0}^{n-1}\left(\frac{1}{p+\varepsilon+m}-\frac{1}{p+\psi+m}\right)$$

$$+\alpha\frac{\varepsilon+m}{\psi+m}\sum_{n=0}^{\infty}\frac{(\varepsilon+m+1)_n}{(\psi+m+1)_n}\frac{u_0^n}{n!}\sum_{p=0}^{n-1}\left(\frac{1}{p+\psi+m+1}-\frac{1}{p+\varepsilon+m+1}\right)$$

$$+m\frac{\psi+m-1}{\varepsilon+m-1}\sum_{n=0}^{\infty}\frac{(\varepsilon+m-1)_n}{(\psi+m-1)_n}\frac{u_0^n}{n!}\sum_{p=0}^{n-1}\left(\frac{1}{p+\psi+m-1}-\frac{1}{p+\varepsilon+m-1}\right)$$

$$-\alpha\frac{\psi-\varepsilon}{(\psi+m)^2}\ {}_1F_1(\varepsilon+m+1;\psi+m+1;u_0)$$

$$-\left[\frac{\psi+m-1}{\varepsilon+m-1}-m\frac{\varepsilon-\psi}{(\varepsilon+m-1)^2}\right]\ {}_1F_1(\varepsilon+m-1;\psi+m-1;u_0).$$

$\Gamma(m)=\int_0^\infty t^{m-1}e^{-t}dt=(m-1)!$ defines the Gamma function, $\Psi(\cdot)$ is the digamma function, and the derivative of the confluent hypergeometric function with respect to any of its parameters is as defined in [41]. Inserting $m=0$ and equating the result to zero, we get

$$\left.\frac{dP(m|0)}{dm}\right|_{m=0}=\frac{\Omega}{\psi-\varepsilon}(\ln\alpha-\Psi(1)+\Psi(\varepsilon)-\Psi(\psi))+\frac{1}{\psi-\varepsilon}\left.\frac{d\Omega}{dm}\right|_{m=0}=0 \qquad (4.1)$$

where,

$$\frac{1}{\psi-\varepsilon}\frac{d\Omega}{dm}\bigg|_{m=0} = \frac{1}{\psi-\varepsilon}\,{}_1F_1(\varepsilon;\psi;u_0) - \frac{\alpha}{\psi^2}\,{}_1F_1(\varepsilon+1;\psi+1;u_0)$$

$$-\frac{\psi-1}{(\psi-\varepsilon)(\varepsilon-1)}\,{}_1F_1(\varepsilon-1;\psi-1;u_0)$$

$$+\sum_{n=0}^{\infty}\left[\frac{(\varepsilon)_n}{(\psi)_n}\frac{u_0^n}{n!}\sum_{p=0}^{n-1}\left(\frac{\alpha+\psi-\varepsilon}{(p+\varepsilon)(p+\psi)} - \frac{\varepsilon+n}{\psi+n}\frac{\alpha}{(p+\varepsilon+1)(p+\psi+1)}\right)\right].$$

We will carry on with the investigation of this bifurcations condition at a later time. This involves obtaining, using the equation derived above, values of ε at which Eq. 4.1 is satisfied given values of the other parameters, α and K_{eq}. It is unfortunate that at this moment, the numerics are proving to be difficult.

4.4 A simulation of the simple gene expression model

In this section, we explore the same model using a different approach, simulation. The reason for doing this is to see if we can get the same graphical representation of the marginal distribution as that obtained from the analytical approach. This comparison is necessary because there is a difference in our theoretical results and the results presented in [68]. We substitute the values of the parameters given in [68] in our analytical expression for the marginal distribution as well as in a simulation to produce two graphs, Fig. 4.9 and Fig. 4.10, which we compare. We also compare the results displayed on these two graphs to those obtained through simulation in [68].

Simulations help us get the overall picture of the outcome from a model, or simply what to expect. A simple model like ours can be easily coded, for instance using Matlab, and simulated. Hence, a simulation would be faster if one is only concerned with simply getting an idea about how a model performs. It is, however, difficult to explore the performance of the model, for example it would be hard and time consuming to carry out sensitivity or bifurcation analysis through a simulation. The difficulty is normally two-fold, since the values of the parameters are often not precisely known, yet even if they are known, one has to do a lot of sampling.

Using the Gillespie stochastic simulation algorithm [35, 43], we obtain the simulation results, displayed in Fig. 4.10, for our model. Compared to the result for the marginal distribution from the analytical approach, Fig. 4.9, the results are similar—an asymmetrical and uni-modal curve. The parameters used to produce these curves were obtained from [68], who reported different simulation results—an asymmetrical bimodal mRNA distribution with peaks near $m = 0$ and $m = 5$. There are no analytic expressions for the distributions given in [68] and there is no explanation as to whether a single simulation or multiple in-

dependent simulations were performed to come up with the result. We go along with the observation that the mRNA distribution is asymmetrical but disagree on it being bimodal. The results in Figs. 4.10 were produced from a long simulation performed over a period of two days. This result closely resembles the analytical distribution displayed in Fig. 4.9, i.e. the analytical and numerical results should eventually converge although it takes a lot of data for this to happen. A lot of sampling has to be done or, alternatively, the program has to be run for a long time to get reasonable simulation results. Without analytical expressions to compare with, it is therefore easy for one to get a wrong conclusion from the simulation result only.

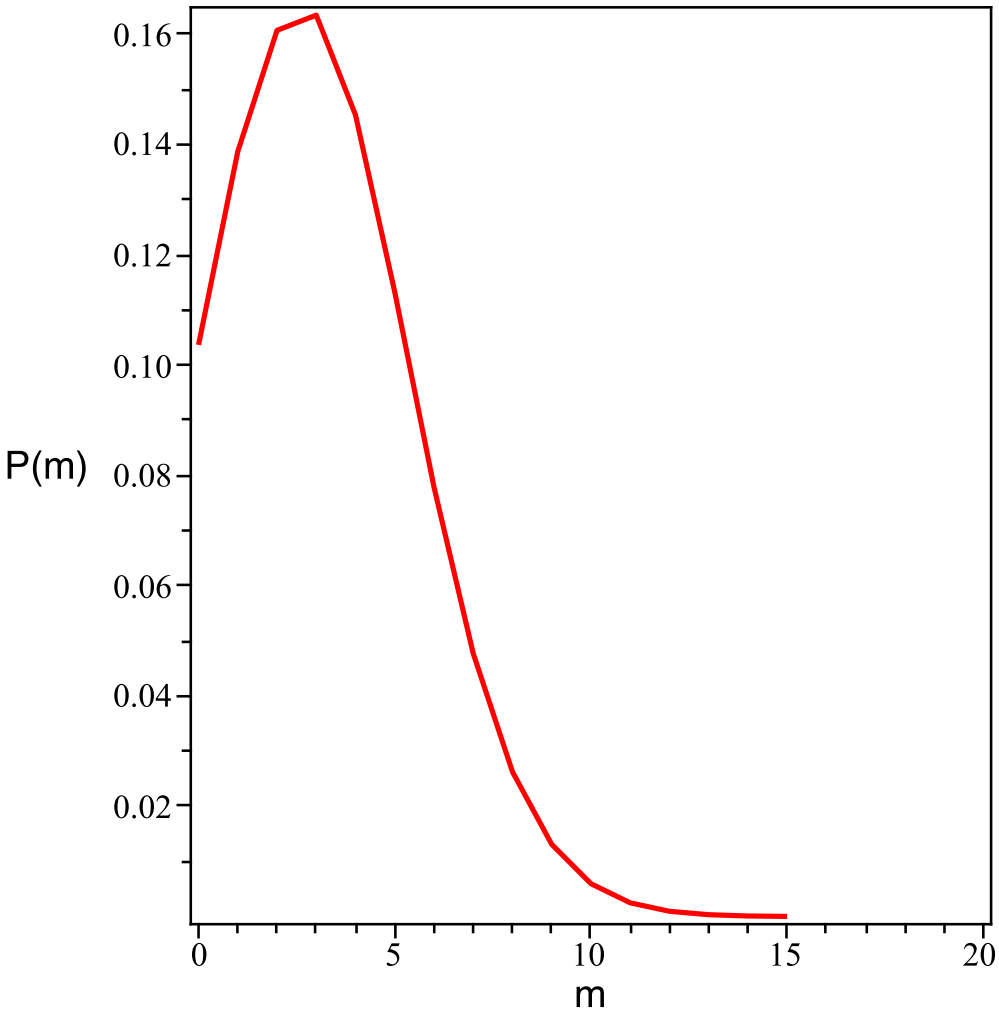

Figure 4.9: Analytical result showing the marginal probability distribution, $P(m)$, for $\alpha = 6.7$, $\varepsilon = 1.2$ and $K_{eq} = 0.33$ (as given in [68]).

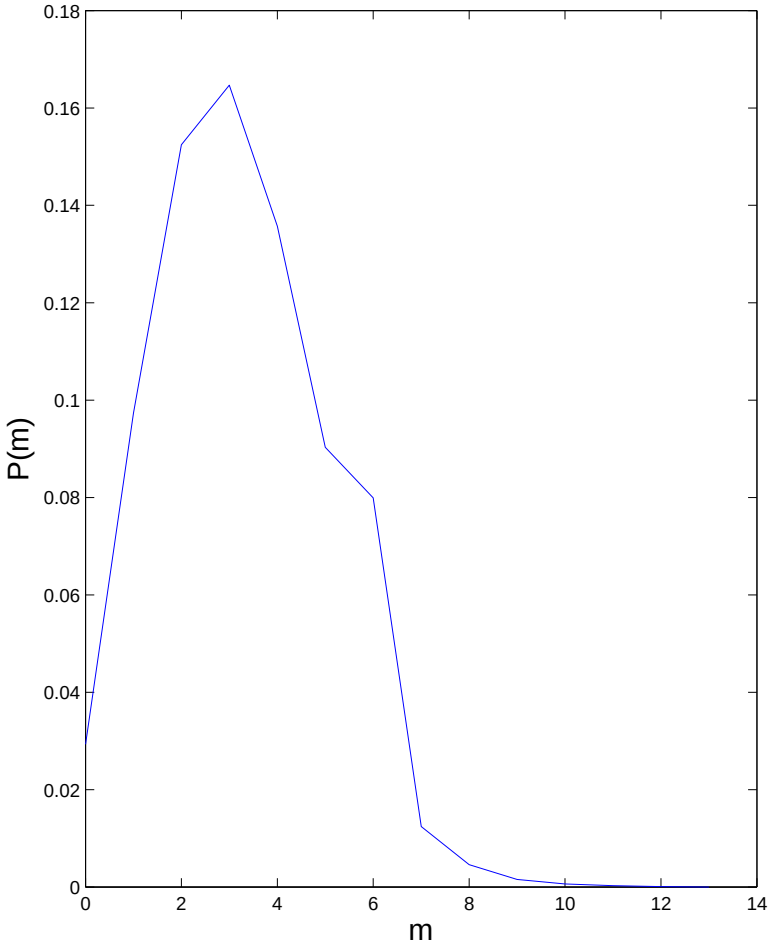

Figure 4.10: Simulation result showing the marginal probability distribution, $P(m)$, for $\alpha = 6.7$, $\varepsilon = 1.2$ and $K_{eq} = 0.33$ (as given in [68]). Two samples were collected from a simulation performed in two 24–hour intervals, each interval with 100000 mRNA molecules, i.e. a total of 200000 mRNA molecules.

Chapter 5

Discussions and Conclusions

Biological systems depend on the ability of the cells in them to function normally and to maintain behavior. Most of the important cellular processes, for example gene expression, operate with very low copy numbers. Chemical reactions which result in changes in numbers of molecules occur randomly, and the molecule numbers change by very small integer values. Involvement of each molecule is indispensable, a small change in the number of molecules of a particular species occasions pronounced relative changes in the reaction probabilities. The result is a stochastic environment where all events occur probabilistically.

The stochasticity complicates the mathematical modelling of biochemical systems which has to take into account the randomness of events, if the true nature of the system is to be revealed. Stochastic methods, such as the Chemical Master Equation (CME) [35], can be used to obtain the probability distributions describing the probabilities of the system being in any given state. Unless a model is simplified, the resultant CME is a huge set of differential equations and is difficult to solve due to huge dimensionality. Simple and solvable CMEs can be derived analytically for correspondingly simple models to obtain expressions for the probability distributions. However, such models do not represent real biological systems which are large and complicated, even though invaluable insight can be gained from them. Large and complicated models can be studied using simulations [8, 67]. One of the problems associated with simulations is that it is hard to obtain expressions for the probabilities and hence for other quantities, for example the moments. Expressions derived in terms of the model parameters can be easily used by other researchers. They also enable investigation of the properties of a system.

There are two common ways to study a stochastic biochemical system. One may decide to do a bifurcation analysis, which reveals qualitative changes in the behavior of the system as parameters change, or one can apply the theory of Metabolic Control Analysis. This latter theory uses control coefficients to measure responses of observables when parameters change. If one has the expressions for the probabilities, then one can assess the way each observable responds to a particular parameter change by calculating a control coefficient for it. The numerical value of a control coefficient tells one whether the observable increases or decreases when the parameter of concern changes, and also gives the magnitude of the response. A greater value for a control coefficient tells us that the parameter, or the step that it represents in a pathway, has greater control over the particular observable compared to others. On the other hand, by plotting graphs that represent the distribution functions and observing the way they change when values of parameters are changed, one can also tell if the system described undergoes stochastic bifurcations.

This study focussed on a simple solvable stochastic gene expression model without feedback. Genes typically appear in not more than two copies in a cell, and hence genetic systems are subject to stochasticity. In this model, the gene is considered to transition between two states, "on" and "off", and this occurs randomly due to stochasticity. This randomness also carries over to the production of the gene products, which are then produced in bursts [68]. The two-state scenario on the gene itself results in an infinite set of ordinary differential equations describing the probability rates of obtaining a certain number of mRNA molecules when the gene is in either state. I solved these equations to obtain the stationary joint probability distributions. Taking into account the fact that we may need to count the number of molecules regardless of the state of the gene, I derived the marginal probability distribution, as a sum of the two joint probability distributions. I also derived the conditional probability distribution functions from the two joint probability distributions,

which give the probability of obtaining a certain number of molecules given the gene is in a particular state. For each of these conditional distributions, as well as for the marginal probability distribution, expressions were obtained for the related moments, namely the mean, the variance, the skewness and the kurtosis. These four were basically used as the observables in this work and help to capture the numerical values of the important features of the distributions. I carried out sensitivity analysis using MCA and developed expressions for the control coefficients of the moments, which I evaluated using values of the parameters from the literature. The expressions for the probability distributions were also used to obtain graphical representations of the probabilities and bifurcation analysis was done.

The results show a successful extension of MCA to genetic systems. In fact, if one successfully derives expressions for the probability distributions, and hence the moments, one should be able to obtain expressions for the control coefficients. Moreover, the stochastic control coefficients obtained do satisfy one of the summation theorems governing classical MCA. The results also corroborate the general observation that noise decreases as the number of molecules increase and that the genetic system undergoes a qualitative behavioral change, stochastic bifurcation, as the parameters change. There are parameter regimes in which the bifurcation occurs and in these regimes, the parameter space is split into two regions where the system displays different types of behavior, basically shifting from having a maximum to being monotonic. The summation theorem enables us to link the system variables with the kinetics of the separate enzyme catalysed reactions in the system and its fulfilment helps us get the insight that reactions in the system occur antagonistically. While one reaction boosts the level of a certain metabolite, we expect another reaction to be utilizing the same metabolte and bringing down the level of this metabolite. This enables the cell to work normally and maintain behavior.

5.1 And where do we go from here?

I mentioned in section 1.3, Chapter 1, that my work was not covering translation. The next step in this work is to include this process. A three–stage model, involving gene activation/deactivation, mRNA and protein dynamics will be considered next, with the objective of obtaining the moments for the protein probability distributions and developing the control coefficients for them. This work has already begun. In section 4.3, Chapter 4, I indicated that the analysis of the equation

$$\left.\frac{dP(m|0)}{dm}\right|_{m=0} = 0$$

would be suspended for now. This was mainly due to time constraints in my program. This equation will be revisited with the aim to obtain the phase diagram, which, in parameter space, will show regions where bifurcations take place.

The model analysed here excluded feedback, as stated in the assumptions, section 2.2. The feedback refers to interference with the promoter activity by the proteins produced by the gene itself or other genes (for example see [28]). This should allow the model to reflect a real genetic system, characterized by a complex model and in which a gene does not operate in isolation.

An example of a more complex model that I intend to study is the lactose (*lac*) operon [1]. The operon encodes genes for lactose permease and for β-galactosidase. The former is required to allow lactose into the cell, while the latter catalyzes hydrolysis of lactose. In the absence of lactose, a repressor binds to the operator region of the DNA and inhibits transcription. The β-galactosidase reaction occasionally isomerizes lactose to allolactose instead of hydrolyzing the disaccharide. Allolactose in turn can bind to the *lac* repressor, inactivating it and allowing transcription of the genes in the operon to proceed. Since we

need the permease and hydrolytic enzyme to get lactose into the cell and to convert it into allolactose, how is the operon induced in the first place? Binding of the repressor to the operator is an equilibrium process. Occasionally, the repressor will dissociate even in the absence of allolactose and allow transcription to proceed. Induction of the operon is thus a paradigm for rare but biochemically significant events. Given that this system has been well studied, for example in [1], all the necessary kinetic parameters are known, which makes this system ideal for the theoretical study of rare events in molecular biology. My stochastic MCA theory will allow me to study the sensitivity of the induction process, including the frequency with which the uninduced operon "leaks" and allows its genes to be transcribed, to various parameters of the model.

Bibliography

[1] Daber R, Lewis M, *J. Mol. Biol.* (2009) **391**, 661–670.

[2] Antoine C, Gandrillon O, Belson G, *BMC Syst. Biol.* (2010) **4**:2.

[3] Minoru S H Ko, *J. Theor. Biol.* (1991) **153**, 181–194.

[4] Paolo V, Rosalind J A, Martins R E, *Phys. Rev. Lett.* (2008) **101**, 118104.

[5] Paszek P, Lipniacki T, Braiser A R, Nowak D E, Kimmel M, *J. Theor. Biol.* (2005) **233**, 423–433.

[6] Zhu R, Ribeiro A S, Salahub D, Kauffman S A, *J. Theor. Biol.* (2007) **246**, 725–745.

[7] Sanchez A, Kondev J, *PNAS* (2008) **105**, 5081–5086.

[8] Schultz D, Onuchic J N, Wolynes P G, *J. Chem. Phys.* (2007) **126**, 245102.

[9] Lipniacki T, Paszek P, Marciniak–Czochra A, Braiser A R, Kimmel M, *J. Theor. Biol.* (2006) **238**, 348–367.

[10] Tabor J J, Bayer T S, Simpson Z B, Levy M, Ellington A D, *Mol. BioSyst.* (2008) **4**, 754–761.

[11] Ozbudak E M, Thattai M, Kurtser I, Grossman A D, van Oudenaarden A, *Nature* (2002) **31**, 69–73.

[12] Thattai M, van Oudenaarden A, *PNAS* (2001) **98**, 8614–8619.

[13] Tomioka R, Kobayashi T J, Aihara K, *J. Theor. Biol.* (2004) **229**, 501–521.

[14] Fraser H B, Hirsh A E, Giaever G, Kumm J, Eisen M B, *PLoS Biology* (2004) **2**, e137.

[15] Raj A, van Oudenaarden A, *Annu. Rev. Biophys* (2009) **38**, 255–270.

[16] Colman–Lerner A, Gordon A, Serra E, Chin T, Resnekov O, Endy D, Pesce C G, Brent R, *Nature* (2005) **437**, 699–706.

[17] Zhou T, Chen L, Aihara K, *Phys. Rev. Lett.* (2005) **95**, 178103.

[18] Kaern M, Elston T C, Blake W J, Collins J J, *Nat. Rev. Genet.* (2005) **6**, 451–464.

[19] Kussel E, Laibler S, *Science* (2005) **309**, 2075–2078.

[20] Suel G M, Garcia–Ojalvo J, Liberman L M, Elowitz M B, *Nature* (2006) **440**, 545–550.

[21] Barkai N, Leibler S, *Nature* (2000) **403**, 267–268.

[22] Elowitz M B, Leibler S, *Nature* (2000) **403**, 335–338.

[23] Berg O G, Paulsson J, Ehrenberg M, *Biophys. J* (2000) **79**, 1228–1236.

[24] Goss P J E, Peccoud J, *PNAS* (1998) **95**, 6750–6755.

[25] Morton–Firth C J, Bray D, *J. Theor. Biol.* (1998) **192**, 117–128.

[26] Acar M, Mettetal J T, van Oudenaarden A, *Nat. Genet.* (2008) **40**, 471–475.

[27] Gillespie D T, *J. Phys. Chem.* (1977) **81**, 2340–2361.

[28] Hornos J E M, Schultz D, Innocentini G C P, Wang J, Walczak A M, Onuchic J N and Wolynes P G, *Phys. Rev. E* (2005) **72**, 051907.

[29] Innocentini G C P and Hornos J E M, *J. Math. Biol.* (2007) **55**, 413–431.

[30] Kepler T B and Elston T C, *Biophys. J.* (2001) **81**, 3116–3136.

[31] Tapaswi P K and Roychoudhury R K, *Stochastic Process. Appl.* (1985) **19**, 359–370.

[32] Shiraishi F, Tomita T, Iwata M, Berrada A A, Hirayama H, *Math. Biosci.* (2009) **222**, 73–85.

[33] Gillespie D T, *Physica A* (1992) **188**, 404-425.

[34] Gillespie D T, *J. Chem. Phys.* (2000) **113**, 297–306

[35] Gillespie D T, *J. Comput. Phys.* (1976) **22**, 403-434.

[36] Gillespie D T, *J. Phys. Chem.* (1977) **25**, 2340-2361.

[37] Fell D A, *Biochem. J.* (1992) **286**, 313–330.

[38] Westerhoff H V, Koster J G, van Workum M, Rudd K E, *Control of Metabolic Processes*, Plenum Press, New York (1990).

[39] Ahmed S, *J. Approx. Theory* (1982) **34**, 335–347.

[40] Shahrezaei V, Swain P S, *Proc. Natl. Acad. Sci. U.S.A.* (2008) **105**, 17257–17261.

[41] Ancarani L U, Gasaneo G, *J. Math. Phys.* (2008) **49**, 063508.

[42] Bruggeman F J, Bluthgen N, Westerhoff H V, *PLoS Comput. Biol.* (2009) **5**, e1000506.

[43] Gillespie D T, *J. Chem. Phys.* (2001) **115**, 1716–1733.

[44] Kacser H, Burns J A, *Biochem. Soc. Trans.* (1979) **7**, 1149–1160.

[45] Angeli D, Ferrell Jr. J, Sontag E D, *Proc. Natl. Acad. Sci. USA* (2004) **101**, 1822–1827.

[46] Krishnan J and Iglesias P A, *IEE Syst. Biol.* (2004) **1**, 54–61.

[47] Sontag E D, *IEE Syst. Biol.* (2004) **1**, 9–18.

[48] Wolkenhauer O, Kitano H, Cho K H, *IEEE, Contr. Syst. Mag.* (2003) **23**:4, 38–48.

[49] Wolkenhauer O, Gosh B K, Cho K H, *IEEE, Contr. Syst. Mag.* (2004) **23**:4, 30–34.

[50] Bayliss L E, *Living Control Systems*, The English Universities Press, London (1966).

[51] Grodins F S, *Control Theory and Biological Systems*, Columbia University Press, New York (1963).

[52] Milhorn Jr. H T, *The Application of Control Theory to Physiological Systems*, W. B. Saunders Co., Philadelphia, PA (1966).

[53] Kacser H, Burns J A, *Biochem. Soc. Trans.* (1995) **23**, 341–366.

[54] Heinrich R, Rapoport T A, *Eur. J. Biochem* (1974) **42**, 97–105.

[55] Heinrich R, Rapoport T A, *Eur. J. Biochem* (1974) **42**, 89–95.

[56] Kacser H, Burns J A, *Symp. Soc. Exp. Biol.* (1973) **27**, 65–104.

[57] Heinrich R, Schuster S, *The Regulation of Cellular Systems*, Chapman and Hall, 1966.

[58] Metzler D E, Metzler C L, Sauke D J, *Biochemistry: The Chemical Reactions of Living Cells*, Academic Press (2001).

[59] Acerenza L, Sauro H M, Kacser H, *J. Theor. Biol.* (1989) **137**, 423–444.

[60] Rocco A, *Phys. Biol.* (2009) **6**, 1–11.

[61] Kyung Hyuk K, Sauro H M, *Math. Biosci.* (2010), 109–119.

[62] Rao V C, Wolf D M, Arkin A, *Nature* (2002) **420**, 231–237.

[63] Paulsson J, *Phys. Life Rev.* (2005) **2**, 157–175.

[64] Hatzimanikatis V, *Metab. Eng.* (1999) **1**, 75–87.

[65] Small J R and Kacser H, *Eur. J. Biochem.* (1993) **213**, 613–624.

[66] Small J R and Kacser H, *Eur. J. Biochem.* (1993) **213**, 625–640.

[67] Gillespie D T, *J. Phys. Chem.* (1977) **81**, 2340–2361.

[68] Shahrezaei V and Swain P S, *Curr. Opin. Biotechnol.* (2008) **19**, 369–374.

[69] Rausenberger J, Fleck C, Timmer J, Kollmann M, *Prog. Biophys. Mol. Biol.* (2009) **100**, 57–66.