UNIVERSITY OF CALIFORNIA, SAN DIEGO

Many Body Theory of Stochastic Gene Expression

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in
Physics

by

Aleksandra M Walczak

Committee in charge:

Professor Peter G Wolynes, Chair
Professor Patrick H Diamond
Professor Katja Lindenberg
Professor José N Onuchic
Professor J Andrew McCammon

2007

The dissertation of Aleksandra M Walczak is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California, San Diego

2007

TABLE OF CONTENTS

LIST OF FIGURES

COMMONLY USED ABBREVIATIONS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
$mRNA$	Messenger Ribonucleic Acid
TF	Transcription Factor
$RNAP$	RNA Polymerase
GFP	Green Fluorescent Protein
$SCPF$	Self Consistent Proteomic Field

ACKNOWLEDGEMENTS

In all respects, educational, scientific and social, the time I spent at UC San Diego has been a great experience, which allowed me to grow and learn. I owe this to the people I met during my graduate studies, who have influenced my perspective on the world around me. I hope they will remain to do so for many years to come. I was incredibly lucky to have as an advisor, Peter G Wolynes, who was and is both a scientific guide and a mentor. He welcomed me into his group and encouraged my own scientific explorations, regardless of how far they took me. I cannot possibly list all the lessons Peter has taught me, but the most general one will guide me in my future work: nature usually makes sense, so we can understand it using logic with a dash of passion.

I have also received a lot of support from other UC San Diego faculty, a lot of them I am incredibly lucky to have on my committee. From my first year Patrick H Diamond opened up a new world of topics and continued throughout my graduate studies to show me new perspectives and directions. His mentoring in every respect of my scientific life helped me many times. I was also incredibly lucky to have met Katja Lindenberg, who would patiently listen to my stories and always ask the right questions. I hope I have picked up some of her wisdom. José N Onuchic taught me invaluable lessons by many times showing how the impossible can come about. If it were not for J Andrew McCammon's kindness and thoughtfulness I would never have come to UC San Diego. I benefited a lot from many insightful discussions with Terrence Hwa and Herbert Levine. I also thank my scientific collaborators and co-authors of my publications, especially Daniel Schulz.

I would also like to thank my family for their loving support. Throughout my education my parents have encouraged my curiosity and supported my choices. They supported by broad interests and taught me to ask myself questions.

I would not have such fond memories if it were not for the friends I have made in San Diego. My office mates, Jacob D Stevenson and Samuel Cho answered all my ridiculous questions and made coming to Urey Hall a great pleasure. Diego U Ferreiro reminded me how much fun science can be. I would like to thank all my friends, in particular: Defne Üçer, Virginia Vandelinder, Shane Keating, Dave Collins, J Kyle Campbell and Michael Buhl. Last, but definitely not least I would like to thank for his constant support, help, encouragement and patience Łukasz Cywiński, without whom I would not have the courage to undertake many things I have completed.

Throughout the course of my graduate studies I was supported by the Center for Theoretical Biological Physics through National Science Foundation Grants PHY0216576 and PHY0225630.

The text and data of Chapter 4, in full, has been published in "Self-regulating gene: an exact solution" by J. E. M. Hornos, D. Schultz, G. C. P. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic and P. G. Wolynes in Phys. Rev. E (**72**), 051907-1-5, (2005). The dissertation author was a contributing investigator and author of this article.

The text and data of Chapter 5, in full, has been published in "Self-Consistent Proteomic Field Theory of Stochastic Gene Switches" by A. M. Walczak, M. Sasai, P.G. Wolynes in Biophys. J. (**88**), 828-850 (2005). The dissertation author was the primary investigator and author of this article.

The text and data of Chapter 6, in full, has been published in "Absolute rate theories of epigenetic stability" by A. M. Walczak, J. N. Onuchic and P. G. Wolynes in Proc. Natl. Acad. Sci. USA (**102**), 18926, (2005). The dissertation author was the primary investigator and author of this article.

VITA

May 22, 1979	Born, Szczecin, Poland
2002	Master of Science in Physics Warsaw University, Warsaw, Poland
2002-2003	Teaching Assistant, Department of Physics University of California, San Diego
2003-2007	Research Assistant, Department of Physics University of California, San Diego
2007	Doctor of Philosophy University of California, San Diego

PUBLICATIONS

A. M. Walczak, J. N. Onuchic and P. G. Wolynes, "Absolute rate theories of epigenetic stability", Proc. Natl. Acad. Sci. USA (**102**), 18926, (2005)

J. E. M. Hornos, D. Schultz, G. C. P. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic and P. G. Wolynes, "Self-regulating gene: an exact solution", Phys. Rev. E (**72**), 051907-1-5, (2005)

A. M. Walczak, M. Sasai, P.G. Wolynes, "Self-Consistent Proteomic Field Theory of Stochastic Gene Switches", Biophys. J. (**88**), 828-850 (2005)

FIELDS OF STUDY

Major Field: Physics

Theoretical studies in biological physics.

Professor Peter G. Wolynes, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Many Body Theory of Stochastic Gene Expression

by

Aleksandra M Walczak

Doctor of Philosophy in Physics

University of California, San Diego, 2007

Professor Peter G Wolynes, Chair

The regulation of expression states of genes in cells is a stochastic process. The relatively small numbers of protein molecules of a given type present in the cell and the nonlinear nature of chemical reactions result in behaviours, which are hard to anticipate without an appropriate mathematical development. In this dissertation, I develop theoretical approaches based on methods of statistical physics and many-body theory, in which protein and operator state dynamics are treated stochastically and on an equal footing. This development allows me to study the general principles of how noise arising on different levels of the regulatory system affects the complex collective characteristics of systems observed experimentally.

I discuss simple models and approximations, which allow for, at least some, analytical progress in these problems. These have allowed us to understand how the operator state fluctuations may influence the steady state properties and lifetimes of attractors of simple gene systems. I show, that for fast binding and unbinding from the DNA, the operator state may be taken to be in equilibrium for highly cooperative binding, when predicting steady state properties as is traditionally done. Nevertheless, if proteins are produced in bursts, the DNA binding state fluctuations must be taken into account explicitly. Furthermore, even when the steady state probability distributions are weakly influenced by the operator state fluctuations, the escape rate in biologically relevant regimes strongly depends on transcription factor-DNA binding rates.

1

Introduction

The discovery of the double helix structure of the DNA [1] built a base for great progress in our understanding of genetics. Finally, genes were not just abstract units, but could be associated with a sequence of base pairs - a functional subunit of DNA, which coded for a specific protein. The mapping of a sequence of nucleic acids in the long lived DNA, via a sequence of nucleic acids in the short lived messanger RNA (mRNA), into a chain of aminoacids, which after modification and folding could become a functional protein, is summarized in the central dogma of modern molecular biology (Figure 1.1).

The DNA is the hereditary material of the cell, which is passed on from generation to generation. DNA is copied into mRNA in a process called transcription.

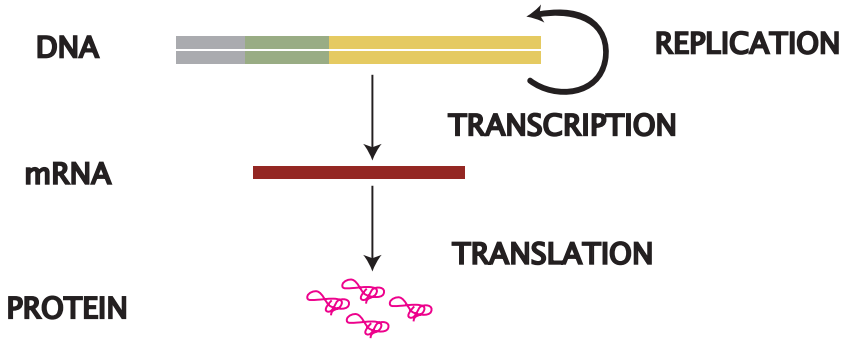

Figure 1.1: The central dogma of molecular biology. Genetic information stored in the DNA sequence is transcribed into mRNA, which is translated into proteins. The information in the DNA is copied onto a second strand in a process called replication.

The product of transcription, mRNA, undergoes translation, a set of reactions which convert the information stored in the nucleic acid into an aminoacid sequence. In addition, DNA undergoes replication, in which a copy of a chosen strand is made. As a result of these three reactions: transcription, translation and replication, the cell is able to make all the molecules necessary to reproduce itself [2, 3].

Yet knowing the sequence of the genome of a given organism does not give us all the information for understanding how, even a single cell organism will work. Not all genes are transcribed in each cell, and typically only a small number of genes is transcribed simultaneously [4]. The timing of when given genes are transcribed is crucial and determines processes such as development [5]. The production of proteins must not only be timed, but also must result in the production of the appropriate concentrations needed by the cell. In other words translation, transcription and replication of the genome must be regulated. The set of processes, which control the composing elements of the central dogma is termed gene regulation.

At any moment, a living organism is thus described not solely by its genome, but by the set of genes that it actually expresses. Proteins are not expressed independently, nor are they expressed at random. Which set of proteins is expressed depends on the point in the cell cycle, the environment of the cell and other needs of the organisms. A change of environment often results in the need for a different set of functional proteins, therefore gene expression patterns must be modified. So although each cell in a multicellular organism has the same DNA, each may express a different set of proteins, which results in the individual cells' specialization to perform different functions. In higher organisms this leads to the formation of tissues. In simpler organisms, such as the bacterium *Bacillus Subtilis*, it can result in different lifestyles: reproduction or spore formation. We see, therefore that genetic information is encoded not just in the DNA sequence, but also in the expression patterns of genes.

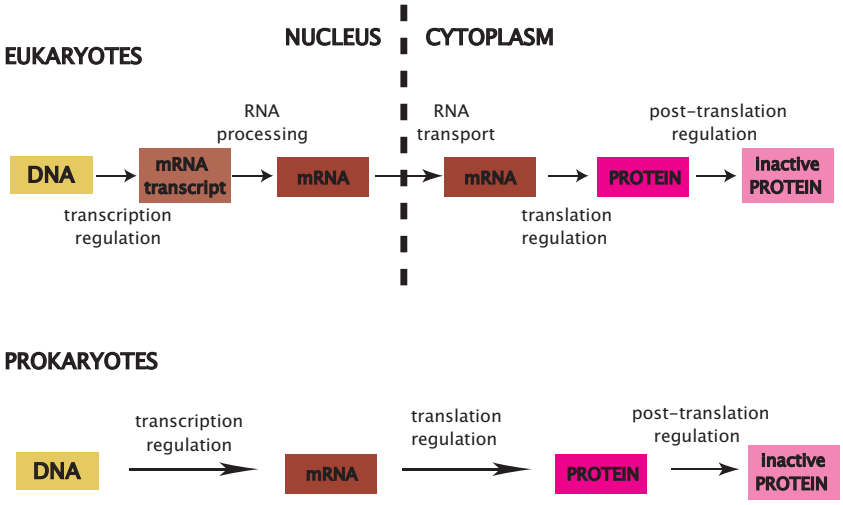

Figure 1.2: Gene expression may be regulated on all levels: transcription, translation and post-translational protein modification. Euakryotes (A) have additional forms of spatial regulation and mRNA modification. Prokaryotes (B) rely mainly on transcriptional regulation, although translational and post-translational control is also present.

Gene regulation occurs in many forms on all levels of gene expression (Figure 1.2). The most common form of gene regulation, present in all organisms from prokaryotes to many cellular eukaryotes, is transcription regulation [4]. Transcription regulation involves the interaction of proteins, themselves products of gene expression, with certain binding sites on the DNA. These special proteins, which are called transcription factors (TFs) bind to the DNA upstream of the initiation site of transcription and either enhance and repress the expression of a given gene. It is therefore easy to see that genes, their product proteins and in turn genes these product proteins regulate, form a complicated network of interactions, called a gene expression network (Figure 1.3) [6].

The great advancement in experimental techniques and their automatization over the last decades has enabled the identification of protein coding regions in the organisms DNA, referred to as the mapping of genomes. Currently the interest lies in mapping regulatory regions and the many interactions between the different genes [5]. Gene networks are often depicted as resembling other known large scale circuits, such as electric circuits or neural networks [7, 5]. Yet if we compare even

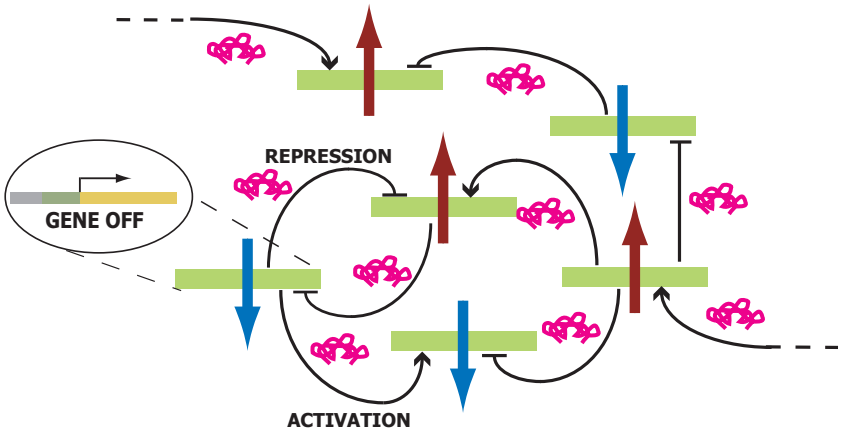

Figure 1.3: A schematic depiction of a genetic network. Genes interact by binding and unbinding of transcription factor proteins, which can act as activators (arrow) or repressors (bar). A gene may be regulated by many types of proteins and a given protein may regulate a couple of genes. As a result of regulation genes are either found in the on states (red up arrow)- protein production occurs at an enhanced rate, or off state- protein production occurs at a basal rate (blue down arrow).

two of these examples, we find many differences. Electric circuits are fabricated by forming hardwired connections between very well separated and characterized repetitive elements, such as diodes, capacitors and resistors. All interactions between these well defined elements occur on fast timescales. In gene networks on the other hand, there exists a diversity elements, each present in small numbers. The output of these "nodes" combine using combinational logic to control other elements, which may function on different timescales. These characteristics alone make a genetic circuit much more complicated. Furthermore, we as yet do not completely understand how a single switch functions in a network. So although great insight can be inferred from the large scale network properties of these genetic systems, in order to understand how these networks of interactions result in living, adapting and evolving cells, we must first understand gene expression regulation at the molecular level. It is worth noting at this point, that even the basic molecular interactions between protein and DNA, as biochemical macromolecules, remains an active area of research [8, 9]. However understanding the interactions on the atomic scale is not the purpose of this work. We will use the description

on the molecular, or often called kinetic rate equation level, and treat molecules as entities. Each macromolecular species will be described by the number of representatives of that species present in the system.

On the molecular level, the elements of a gene network are proteins and stretches of nucleic acids, which interact by means of chemical reactions. Chemical reactions are stochastic in nature - they are the emergent outcome of many processes and can therefore only be described to occur with a certain probability [10]. On top of this omnipresent source of stochasticity, the number of components of the chemical reactions involved in gene regulation is small [11]. There is typically one active copy of DNA per cell, a few copies of mRNA and tens to hundreds copies of a protein of a given species. As a result, gene expression is an inherently noisy process [12, 13, 14, 15, 16]. Yet cells function, moreover they often function in a predictable way. Therefore one of the challenges is to understand the role of noise in gene expression, which becomes a beautiful example of stochastic processes in a many-body system out of equilibrium.

Stochastic gene expression is a very broad subject, which may be studied both experimentally [17, 18, 19, 20, 14, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32] and theoretically [33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47]. Of course, I am not able to discuss all of the related topics in the scope of this thesis. I will review some of the work done by others in Chapter 3. Most of the attention on stochasticity in gene expression has focussed on noise originating from small numbers of mRNA and protein molecules of a given species [48, 25, 29, 28]. But the binding and unbinding of transcription factor proteins to the DNA may be out of equilibrium [49, 50, 51]. In most previous studies the binding and unbinding reactions are assumed to be fast enough to be well described by an equilibrium constant. However this is assumption is not always correct. In such cases, since there is only a single copy of the gene, one cannot average over the binding sites and the additional source of noise must explicitly be taken into account in the theoretical description. In the present study I theoretically investigate the role of

fluctuations in the binding of transcription factors to the DNA binding sites [40, 52, 53, 54]. I consider how this kind of noise interacts with noise arising from small molecule numbers and describe the emergent behaviour of the system. In Chapter 4 I propose a theoretical description which treats both kinds of noise on equal footing. I present an example of a genetic circuit, the steady state properties of which can be found exactly within this framework. In Chapter 5 I propose an approximation, which I call the Self Consistent Proteomic Field Theory (SCPFT). The SCPFT allows for a computationally efficient and universal treatment of networks. I use SCPFT to investigate the steady state properties of the simplest possible network- the toggle switch [55]. Comparison with numerical results is also made. In Chapter 6 I discuss the effects of the two kinds of noise and their interplay on the dynamical properties of simple gene circuits, namely on the lifetimes of the attractors in the simplest bistable network. A summary and conclusions are presented in Chapter 7. I start with a brief introduction to gene regulation.

2

A Review of the Biological Aspects of Gene Regulation

Gene expression regulation is a term which refers to all the processes that stand behind the fact that in each cell at a certain time a given group of genes is expressed, which results in a certain protein pool. All the processes that result in gene expression may be regulated: transcription, translation, modifications of mRNAs, post-translational modifications of amino-acid sequences. The number of levels of regulation typically increases with the complexity of the organism. However transcription regulation is present in all organisms, even the simplest - the bacterial infectant, the phage. For this reason I will be primarily concerned with the topic of transcription regulation. Gene expression does not occur in a vacuum, or even the experimental paradise of a well isolated system. All these reactions take place in a living cell, which must survive and reproduce. As a result, gene expression is also coupled to replication and cell division, but these topics are outside the scope of the present work.

Eukaryotic gene expression involves many more elements and nuances than does prokaryotic gene expression [2, 3, 4]. Eukaryotes are organisms that have a cell nucleus, containing the cell's DNA. Prokaryotes lack this compartmentalization- the DNA is in the cytoplasm. So one prominent example of the additional constraints, and therefore levels of regulation, which arise in eukaryotes, is the need to transport the mRNA out from the nucleus to the cytoplasm, where translation takes place. The models considered in this study will not deal with these addi-

tional constraints. For this reason, in order to introduce the generic elements of gene expression, I will focus on prokaryotic systems.

The aim of transcription is to produce a strand of mRNA corresponding to the information encoded in the given strand of DNA. DNA is transcribed in such a way that the RNA grows from the 5' to the 3' end. Transcription is carried out by a complicated enzymatic protein machinery, which requires the interaction of many subunits, which together are called RNA polymerase (RNAP). Of course, RNAP transcribes not only mRNA. There are also other forms of RNA, such as transfer RNAs (tRNA), ribosomal RNAs (rRNA), recently discovered small RNAs (sRNA), editing RNAs. Usually only one or a few genes are transcribed simultaneously [4]. RNAP binds to a region called a promoter upstream from the coding region (Figure 2.1 *A*). The first synthesized RNA base pair corresponds to position +1. The core RNAP consists of 4 subunits and a number of proteins called sigma factors, which together form the transcription unit. One of these sigma factors is responsible for the specific binding to the promoter as opposed to nonspecific binding to other parts of the DNA. After transcription is initiated the sigma factor dissociates. The promoter sequences are recognized by the sigma factors and have very few conservative base pairs (a group of 6 bp around -10 and -35). The strength of a promoter (that is how effective the transcription will be) can differ by 1000 fold depending on the complementarity of the promoter sequence and the RNAP-sigma factor complex (Figure 2.1 *B*).

Binding of the polymerase to the promoter results in the opening of the double stranded DNA. Negatively supercoiled structures unwind and open more quickly, which is referred to as topological activation. As in most cases in biology, there are exceptions to this rule. One exception, for example, is the gene which encodes the subunits of gyrase, which is the enzyme responsible for the negative supercoiling of the DNA. This group of interactions therefore form a topological negative feedback loop - yet another example of regulation in the process of expression. The average transcription rate is about 40 nucleotides per second. Transcription of the first 30

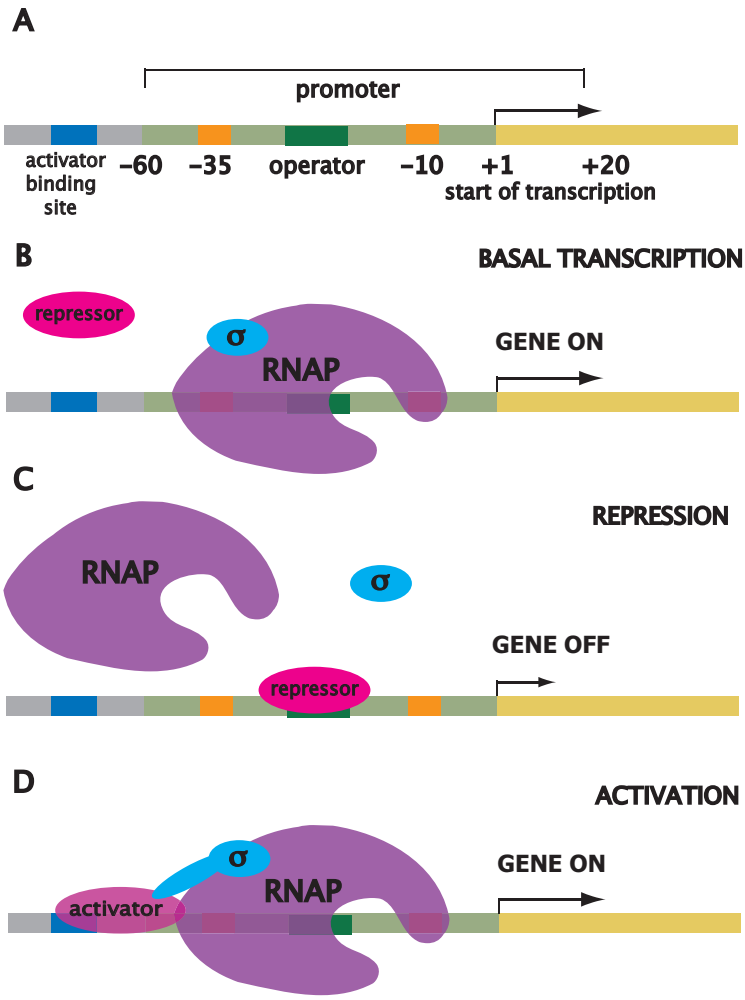

Figure 2.1: Transcription regulation on the molecular scale. (A) A detailed view of the promoter. The conserved base pairs recognized by sigma factors are at -10 and -35 bp. $+1$ is the first transcribed base pair. Transcription is carried out by an enzyme called RNA polymerase (RNAP) (B). Repressors bind to the operator and mostly act as roadblocks (C) enabling the RNAP to bind. Activators bind to the activator binding site (D) and often recruit sigma factors, which enhance the rate of transcription initiation.

base pairs, however, is very important, because the more quickly the RNAP leaves this region, the sooner another RNAP can bind and start transcribing. The first 9 or so base pairs are transcribed without the RNAP moving. The possibility of the RNAP dissociating and resulting in abortive initiation is very large in this region. If the initiation is successful the promoter is freed and another RNAP may bind. The described process of initiation takes about 1-2 seconds which is relatively long compared to other parts of transcription.

Another interesting point is, that there are very few specific promoters compared to other sites to which the DNA could bind nonspecifically. If the RNAP were to find these sites by 3D diffusion, which can be estimated to give association rate constants of the order of $k \sim 10^8 M^{-1} s^{-1}$, it would take much longer than the experimentally observed time, which corresponds to association rate constants of the order of $k \sim 10^{10} M^{-1} s^{-1}$. It is currently believed the RNA performs a random $3D$ walk to find any piece of DNA in the cell, or test tube and then it does a $1D$ random walk along the DNA [56]. Early experiments second this by showing that if one elongates the DNA the RNAP finds its target even faster [56]. In recent years there has been a significant experimental revival of interest in these problems [57, 58], with the advancement of techniques which allow for quantified measurement of motion in the cytoplasm.

The above elements are essential to initiate transcription. However, on top of these elements, gene expression is regulated. All of the processes mentioned above can be controlled: the binding of the RNAP, the opening of the DNA, the forming of a stable RNAP-DNA complex to name just a few. The most simple case is that of direct binding and unbinding of a transcription factor, which is a protein that controls the expression of the gene in question. We can distinguish activators and repressors. Repressors typically work as roadblocks (Figure 2.1 C). They bind to an operator sequence downstream or directly to some portion of the promoter sequence enabling the RNAP to bind in that given place. This blocking results in the repression of transcription. Activation may take on a few

forms (Figure 2.1 *D*). Many activators work by increasing the affinity of binding of the RNAP to the promoter. They recognize specific sites on the RNAP and form favourable interactions, as for example between acidic and base groups. Binding of the activator may increase the affinity of the polymerase to the promoter. Activator binding can also help recruit the polymerase by modifying its configuration with respect to the activator. An activator can act close to the promoter or further on downstream. It can also change the conformation of the DNA, by opening it and result in an enhanced transcription rate. The interactions between the activator and the polymerase or DNA discussed so far are electrostatic or hydrophobic in nature. Another possibility is allosteric control, where the binding of an activator to the polymerase changes its conformation and therefore its affinity to the DNA. The purpose of an activator is to enhance the rate of transcription and biological systems utilize whatever the detailed mechanism to obtain this enhancement. The important thing to remember is that the rate of transcription initiation results from chemical reactions between the activators or repressors and therefore depends on the concentration of these transcription factor proteins.

It is worth mentioning some more specialized forms of regulation. DNA looping results in repression, which seems not be connected to protein concentrations. But it is usually triggered by binding of a repressor, which enables the polymerase to bind. Another form of regulation results when an external signal changes the TF conformation in such a way that it changes from a repressor to an inducer, which recruits polymerase. An example is that of mercury where the promoter controls two genes, one of which is situated to the right of the promoter and the other one to the left. One of the genes is normally transcribed when the activator is bound. However when mercury is present in the cell, it is highly toxic and needs to be sequestered. Then the gene to the other side of the promoter needs to be transcribed to produce the proteins, which export mercury from the cell or convert the element into a less toxic form. Mercury molecules bind to the activator and induce a change in the twist of the DNA which disables the transcription of the

gene to the right, but starts the transcription of the gene to the left.

A common pattern in biological systems is the control of a gene by a few transcription factors. This allows for combinatoric control. A transcription factor, or a set of transcription factors can also regulate the expression of a group of genes. Such a group of genes, the expression of which is controlled together is called an operon. The most famous and oldest known example of such a form of regulation is the *lac* operon discovered by Jacob and Monod [59].

Transcription comes to a halt when the RNAP reaches the stop signal. Transcription termination may also be regulated, but I will not discuss this process in detail. The stop signal is usually a hairpin that the transcribed RNA forms. Hairpins are formed by rich GC (Guanine:Cytosine) regions that interact and bind. A hairpin is followed by a region of weak DNA-RNAP binding (many Us (uridines)) and such a sequence results in dissociation of the polymerase. Some stop signals require a protein called ρ, which binds to and releases RNA from a transcribing elongation complex allowing for termination.

Transcription termination is worth mentioning, because it allows for the existence of another possibility of control of gene expression called attenuation. The best known example of this is the TRP operon. Attenuation is a regulatory mechanism which couples transcription and translation. The TRP operon is repressed if the level of tryptophan is high. The operon starts of with a leader sequence, the mRNA structure of which may form hairpins. There are 4 possible arms:1, 2, 3 ,4. If the 3 : 4 arms form a hairpin transcription is halted in the attenuator (rich U region). Transcription is stopped until the leader region is translated. The leader sequence requires two tryptophans in a row. Tryptophan is a rare aminoacid, therefore if there is not enough tryptophan in the cell the ribosome will halt in this position, stopping translation and blocking the first arm. The 2 : 3 arms then form a hairpin enabling the formation of the 3 : 4 termination hairpin and the transcription and translation of the structure genes continues. If tryptophan is present the ribosome continues and stops at 2 (the end of the leader region) and

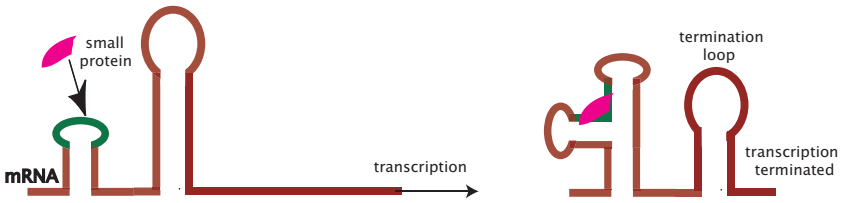

Figure 2.2: A schematic depiction of a riboswitch, which terminates transcription. A small protein binds to the mRNA, which is still being transcribed and changes its configuration, so it forms an hairpin loop which blocks the RNAP from transcribing further.

the $2:3$ hairpin cannot form, but the $3:4$ does. This terminates transcription.

Another form of transcription regulation uses sigma factors. Sigma factors regulate the specific binding to the promoter. If the conditions in the cell change, a different set of sigma factors may be used that recognize different promoter sequences and transcribe the newly necessary genes. This form of gene expression is used when a completely new set of genes has to be transcribed: heat shock genes, before sporulation or by phage to make the host cell express the genes needed by the intruder.

As was already noted, in this study we will be interested in transcription regulation, as it has been experimentally shown [18, 60] that translation is the less noisy step of gene expression. Noise arising in the transcription process simply gets amplified by close to deterministic translation. However I will briefly mention the main forms of translation regulation, albeit not as exhaustively as transcription regulation.

Once an mRNA transcript is produced, the translation process can be regulated in many ways. In eukaryotes this step of regulation is much more developed and important, since mRNA stability can vary over large timescales. In these organism mRNA needs to be exported from the nucleus in order for translation to proceed. Therefore transport of transcripts allows for a simple form of spatial regulation. Furthermore eukaryotic mRNAs must be capped and polyadenylated, and the introns must be accurately removed, which can lead to different proteins produced based on the same gene (alternative splicing). A very interesting feature,

which is very similar to transcription regulation by transcription factors and exists also in prokaryotes, is that of regulation by binding of non-coding RNAs (sometimes called small RNAs in bacteria) to the mRNA. This form of regulation can result in either repression or activation of translation, by enabling or preventing the binding of the ribosome respectively, and has been shown to have an effect on the stochastic nature of gene expression [61]. A more complicated form of RNA-RNA and RNA-protein interaction takes place in riboswitches (Figure 2.2), which are parts of mRNA that can bind a small target molecule (protein or RNA) and the binding event affects the translation process. An mRNA that contains a riboswitch is directly involved in regulating its own activity, depending on the presence or absence of its target molecule. The interaction of non-coding RNAs with mRNA is usually based on formation of secondary hair-pin like structures, which prevents translation. In riboswitches it sometimes involves the formation of more complicated tertiary structures. The main aim of these types of interactions, are to change the stability of the mRNA: either increase its lifetime, or tag it to be recognized by the cells degradation machinery. Once a protein is formed, post-translational modifications can take place, such as methylation, acetylation of the aminoacids or disulfide bond. However these events are much more common in eukaryotes than in prokaryotes.

In summary, transcription regulation is the most important and stochastic form of gene expression regulation in prokaryotes. Since the models presented in this work do not take into account many features of eukaryotic gene expression, such as the compartmentalization of expression processes and chromatin unravelling, I will associate gene expression regulation with transcription regulation.

3

The Physics of Gene Regulation - Background

3.A Experimental approaches to noise in gene expression

Work on the molecular details of gene regulation was pioneered by Jacob and Monod [59], who were able to discover how proteins can be selectively produced based to a large extent simply by analyzing the growth curves of *E.coli* populations in different media. Recently, great experimental advances in molecular biology and genetics have made it possible to study gene expression in great detail and to manipulate the regulatory pathways *in vivo*. The sequenced genomes allow for targeting and modification of specific genes, in specific genetic regulatory systems. Such a detailed map of genes on the chromosomes has been possible thanks to the invention of high throughput methods such as PCR (polymerase chain reaction) and CHIP-CHIP experiments. Plasmid techniques allow for insertion of desired genes into specific sites on the host chromosome. Although the procedures are far from automatic, they allow for a direct studies of the regulatory interactions in networks in great detail. With the aid of these techniques, even Jacob and Monod's *lac* operon has been found to have additional regulation pathways [62].

The study of noise expression is based to a large extent on incorporating genes which code for fluorescent proteins, such as GFP (green fluorescent protein) [63] and its many variations, downstream of promoter sequences which control the gene of interest. As a result, GFPs are produced at the same time as the genes' original

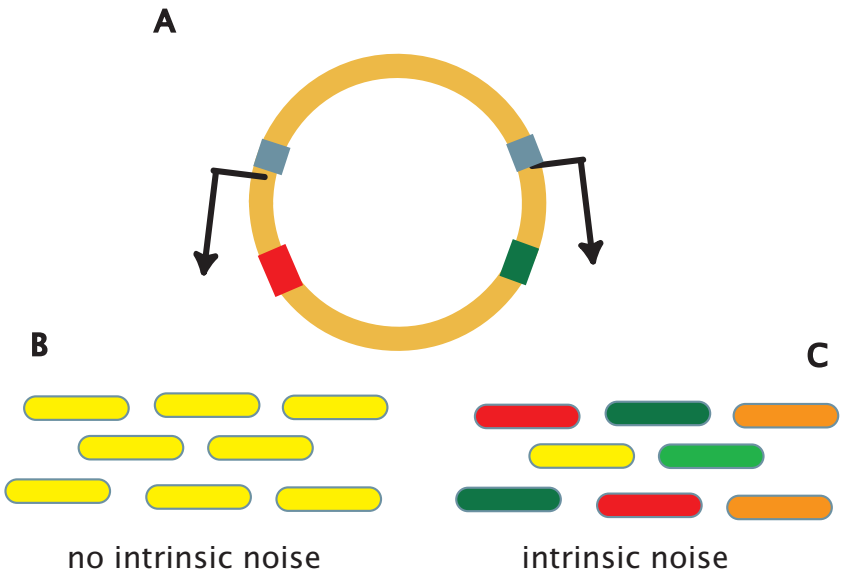

Figure 3.1: A schematic representation of the experimental system used to study intrinsic and extrinsic noise. Two colours of reporter proteins (red and green) are fused to two copies of the same gene A. If intrinsic noise is not present each cell expresses similar amounts of the two genes and the cells fluoresce in yellow B. If intrinsic noise is present the differences in the expression patterns of the two copies of the gene are visible as different ratios of red to green reporter proteins C.

product protein and their level, which can be measured using fluorescence counting under microscopes or in flow cytometers, is a signature of gene expression activity. These experiments make it possible to look at the expression patterns of genes in individual cells, as opposed to a population [17, 55]. As a result, cell to cell variability in the transcripts has been observed: some cells have high levels of gene expression, while others have low. Thus the stochastic nature of gene expression has been confirmed experimentally. The coefficient of variation, defined as the ratio of standard deviation to the mean is a commonly used measure of noise.

These methods, although easy to implement and visually pleasing, do have certain drawbacks. For example, the type of GFP protein which was used in early experiments, had a folding timescale on the order of a cell cycle. So the fluorescent yields did not give an accurate description of the current protein concentration of the gene transcript of interest. However, the fluorescent protein technique can,

and often is, supplemented by northern blot assays, which show the concentration of mRNAs in the cell. Furthermore the use of luciferase as a reporter of gene expression and more refined GFP proteins is becoming more popular.

One class of experiments considered two equivalent, independent gene reporters, with different coloured GFPs, placed in the same cell and controlled by identical promoters on the same prokaryotic chromosome, equidistantly from the origin of replication [18, 28, 60, 20, 64] (Figure 3.1). This has lead to the partitioning of noise into intrinsic and extrinsic components [37, 33, 20, 36]. The first are due to the stochastic nature of expression of a gene and results in differences between different expression profiles of the same kind of gene in the same cell. Extrinsic effects are due to differences in the environment in each cell, such as local protein concentrations, and affect the two reporters in the cell equally, but result in differences between the two cells. Extrinsic noise can be further subdivided into fluctuations of the rates of reactions that have an effect on the whole cell and fluctuations in the concentrations of proteins specific to the pathway of interest.

One of the main experimental challenges of the field is to isolate the system one wants to study and be certain there are no hidden interactions with other parts of the network. This task is indeed complicated, as even model networks often contain the possibility of hidden feedback loops in certain experimental conditions, as was the case with the *lac* operon [21, 65].

With the experimental techniques presently available we can not only ask questions, but we can propose new constructions, which will allow us to probe these issues experimentally [12]. One such example is the toggle switch [55], which is discussed in detail in Chapter 5.

Genetic regulatory circuits systems are an example of a stochastic interacting nonlinear elements which produce emergent behaviour, which can be studied experimentally. This is why it is an interesting problem for physicists.

3.B Modeling gene expression regulation

In order to describe gene switches using a model we need to describe a stochastic nonlinear process. The earliest models of gene networks were the Boolean models of Kauffman [66, 67], where each gene was assumed to be either on or off with a network of interactions between them. The experiments described above allowed for quick quantification of protein concentrations and this opened the doors for many modeling studies, which described the experimental observations [48, 68]. Most of these studies focussed on kinetic models. The simplest models which have a molecular description describe the time evolution of each species, protein numbers (n_i), mRNAs (m_i) using deterministic kinetic equations of the following form:

$$\begin{aligned} \frac{dn_i}{dt} &= g_i m_i - k_i n_i(t) \\ \frac{dm_i}{dt} &= \alpha_i f(\vec{n}) - \delta_i m_i(t) \end{aligned} \tag{3.1}$$

where $f(\vec{n})$ is typically a sigmoidal function of repression or activation, for example for repression (Figure 3.2) $f(\vec{n}) = \frac{1}{1+Kn_j^p}$, where p is the measure of nonlinearity called the Hill coefficient and K is the equilibrium binding constant [69, 70]. g_i, k_i and α_i, δ_i are respectively the synthesis and degradation rates of protein n_i and mRNA m_i. The mRNA equation is often eliminated and its effect is included through time delay in the protein dynamics:

$$\frac{dn_i(t)}{dt} = f(\vec{n}(t-\tau)) - k_i n_i(t) \tag{3.2}$$

These equations are relatively straightforward to analyze for fixed points of the dynamics and stability of the solutions around those points. They offer a lot of intuition and can be easily expanded to consider combinatoric control [35, 71, 72, 73, 74] and many species [75].

Recent experimental work has created a need for a stochastic description of genetic networks. The most general procedure when describing a stochastic system

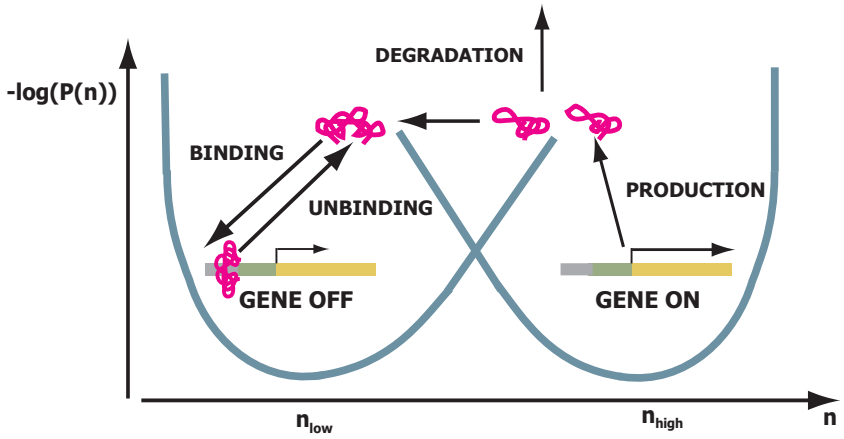

Figure 3.2: A schematic representation of a self-repressing switch. The gene may be found in two states: on or off. The state of the gene is regulated by binding and unbinding of the protein that gene produces.

is to start with an equation for the evolution of the probability distribution [76, 52, 53], which describes the probability of the system having a given numbers of molecules of a given species at a moment in time $\vec{p}(\{x_i\}, t)$, where $x_i = \{n_i, m_i\}$ and the vector notation allows for different possible states of the gene.

$$\frac{d\vec{p}(\{x_i\}, t)}{dt} = \sum_j w(x_j \to x_i)\vec{p}(\{x_i\}, t) - w(x_i \to x_j)\vec{p}(\{x_j\}, t) \qquad (3.3)$$

In order to work with these kinds of equations some approximations must be made. These have typically assumed large protein concentrations and equilibration of the gene expression state and have lead to Fokker-Planck [41, 77, 78, 79, 76] and Langevin descriptions [24, 25]. The Langevin approach in this context has only been pursued via simulations [24, 25], although both multiplicative and additive noise terms have been considered. The linear noise approximation [34], which expands the terms in the master equation in the inverse of the cell volume as opposed to large protein numbers in the Fokker-Planck approach, has also proved useful when describing experiments which focused on noise [20]. Both of these expansion methods are quite standard in nonequilibrium statistical mechanics and I will not describe them in detail. Derivations may be found in textbooks [10,

80, 81, 82]. All the approaches developed have assumed that the binding and unbinding of transcription factors to the binding sites may be assumed to be in equilibrium (Langevin), or close to it (linear noise approximation, Fokker-Planck). Kepler and Elston [76] included the operator state dynamics explicitly in their master equations. However they made the assumption that the DNA binding state is close to equilibrium in their analytical work, by looking at the Fokker-Planck and deterministic equations. They performed simulation studies in the regime on large DNA binding site fluctuations, however claimed this regime is of little biological interest. The generation function [10] has also been used [52, 83] and this technique is described in detail Chapter 4.

And of course there are simulation approaches [41, 77, 84, 85, 86, 75, 87, 76, 88, 89, 90], which are extremely useful but also pose problems. Most of these studies use a Monte-Carlo algorithm, adopted for simulations of chemical kinetics by Gillespie [91], in which one can also randomly choose the time window between reactions. With the traditional approach based on simulations it is impossible to exhaustively scan a wide range of parameters and calculate the phase diagram. Simulations are simply too computationally expensive - especially for slow binding kinetics: one run can take up to a day. Also a simulation study does directly not provide intuition about the relevant combination of parameters which act as control parameters of the problem.

Again we can see, that from a physicists perspective, methods of nonequilibrium statistical physics can be widely applied and developed using the example of gene regulatory systems. In the next chapter and in the whole of the thesis I hope to demonstrate that these systems are also ideal for studying many body effects in systems out of equilibrium.

3.C The formalism

In the present study we develop a formalism, which treats two kinds of noise on equal footing: noise arising from small protein numbers and noise arising from slow binding and unbinding of transcription factor proteins to the DNA binding site. For simplicity, we will refer to the strand of DNA which codes for a protein as a gene. To describe the state of a single gene, we have to specify the "state of the gene" - whether that gene is being transcribed at an enhanced or repressed level, and the number of proteins that gene produces present in the system at time t. We do this by introducing a joint probability distribution $P(\vec{n},t) = [P_{on}(n,t)P_{off}(n,t)]$ [40, 76]. The evolution of this probability distribution in time is governed by a master equation. For example for a self-repressor, such as that described in equation 3.1 with $n_i = n_j$ and neglecting the mRNA step within the described formalism:

$$\begin{aligned}
\frac{\partial P_{on}(n)}{\partial t} &= g_{on}[P_{on}(n-1) - P_{on}(n)] + k[(n+1)P_{on}(n+1) - nP_{on}(n)] + \\
&\quad + fP_{on}(n-p) - h\prod_{s=1}^{p}(n-s)P_{off}(n) \\
\frac{\partial P_{off}(n)}{\partial t} &= g_{off}[P_{off}(n-1) - P_{off}(n)] + k[(n+1)P_{off}(n+1) - nP_{off}(n)] + \\
&\quad - fP_{on}(n-p) + h\prod_{s=1}^{p}(n-s)P_{off}(n)
\end{aligned}$$

where p describes the order of the oligomer which acts as the repressor ($p = 2$ is a dimer, $p = 3$ a trimer), g and k are the synthesis and degradation rates and h and f describe the binding and unbinding terms. A generalization of this equation can be written down to account for mRNAs, with $\vec{P}(n,m,t)$ as the probability distribution.

To describe the state of an N gene system one would have to consider an 2^N state probability vector and the corresponding equations which would be coupled by the binding and unbinding terms. In section 5 I will describe an approximation which allows us to overcome this problem and make progress. I will also describe

a field-theoretical operator formalism proposed independently by Doi [92, 93] and Zeldovich [94, 95] and further developed by Peliti [96, 97] for diffusion reactions. The technique was reviewed by Mattis [98]. For each protein concentration a creation and an annihilation operator are introduced such that $a^\dagger|n\rangle = |n+1\rangle$ and $a|n\rangle = |n-1\rangle$. These satisfy $[a, a^\dagger]$=1. For a process only involving a single protein particle number, the state vector is defined as $\Psi = \sum_n P_(n,t)|n\rangle$, where $P(n,t)$ is the probability of having precisely n particles. The master equation 3.4 is written as $\partial_t \Psi = \Omega\Psi$ using a spinor hamiltonian for the dynamics of the DNA coupled to the proteins. Ω is a non Hermitian hamiltonian operator. Ω for the simple self-repressor gene switch is

$$\Omega = (\bar{g} + \delta g \sigma_z)(a^\dagger - 1) + k(a - a^\dagger a) + \mu^+(\sigma_x - 1) + \mu^-(-i\sigma_y - \sigma_z) \qquad (3.4)$$

where $\bar{g} = \frac{g_{on}+g_{off}}{2}$, $\delta g = \frac{g_{on}-g_{off}}{2}$, $\mu^+ = \frac{h(a^\dagger a)+f}{2}$, $\mu^- = \frac{h(a^\dagger a)-f}{2}$ and σ_i are just regular Pauli matrices. In this operator formalism averages are obtained by taking the scalar product with the bra $\langle 0|e^a$. The formulation allows one to easily guess trial function for the right hand states and perform a non-hermitian variational calculation [99, 40].

In chapter 6 I turn my attention to the question of the stability of the simplest gene switch by mapping it onto a bistable system. The theory of escape from a steady state of a two state system has a long tradition in physics and dates back to Kramers [100], who considered the formulation of the problem in configuration space as first proposed by Smoluchowski [101]. In the basic problem one has two steady states A and B separated by a metastable state at C, called the barrier (Figure 3.3). If the system finds itself in one of the states, say A, in the absence of noise it would stay there for ever. However as a result of stochastic motion, which in traditional problems is associated with temperature, after a certain time the system will escape and cross the barrier to the new minimum B. The rate of this process is called the escape rate and is the quantity of interest in the problem. The probability of finding the system in a given state, defined by the position of the particle may be described by a probability distribution $P(n,t)$. The evolution

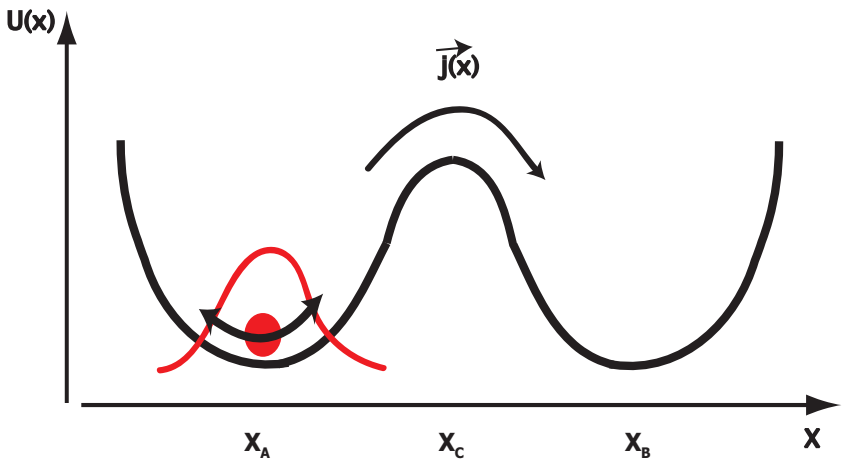

Figure 3.3: The traditional Kramers problem of escape over a potential barrier in a two state system. The two states are marked by different values of the reaction coordinate x, as x_A and x_B, with a barrier at x_C. There is a steady state probability distribution for the particle in a given well (red distribution in well A). The particle can escape due to noise.

of this probability is governed by a Fokker-Planck equation:

$$\frac{\partial P(x,t)}{\partial t} = \frac{\partial}{\partial x}[\frac{\partial D(x)P(x,t)}{\partial x} - v(x)P(x,t)] \tag{3.5}$$

and for potential problems $v(x) = -\frac{\partial U(x)}{\partial x}$, where $U(x)$ is the potential sketched in Figure 3.3. Also $v(x) = \Gamma f(x)$, such that in equilibrium $\Gamma^{-1}D = k_BT = \beta^{-1}$. This equation is simply a continuity equation for the current of particels $\vec{j}(x)$, such that equation 3.5 may be rewritten as $\frac{\partial P}{\partial t} + \frac{\partial \vec{j}}{\partial x} = 0$. The steady state current is simply found from $j(x) = -e^{\int_{x_0}^{x} dx' v(x')/D(x')} \frac{\partial}{\partial x}[e^{-\int_{x_0}^{x} dx' v(x')/D(x')} P_{st}(x)]$ with boundary conditions in the two wells and $P_{st}(x)$ is the steady state solution. Using the boundary conditions we can find the current between the two wells to be:

$$j(x) = \frac{1}{\int_{x_A}^{x_B} dx e^{-\int_{x_0}^{x} dx' v(x')/D(x')}} [e^{-\int_{x_0}^{x_A} dx' v(x')/D(x')} P_{st}(x_A) - e^{-\int_{x_0}^{x_B} dx' v(x')/D(x')} P_{st}(x_B)] \tag{3.6}$$

In general the integral in the denominator depends on the shape of the barrier. For simplicity we can consider a potential problem with constant diffusion $v(x')/D(x') = -\beta \frac{dU(x')}{dx'}$, since the purpose of this illustration is to develop intuition.

Equation 3.6 is the general form of the steady state current and to make progress we need to make approximations. We can also assume particles are removed upon reaching well B ($P_{st}(x_B) = 0$). For a sufficiently narrow and high barrier, we can approximate it using a harmonic potential $U(x) = U(x_C) - \frac{1}{2\sigma_c^2}(x - x_C)^2$ and perform the resulting Gaussian integral to give (after taking the limits of integration to $\pm\infty$):

$$j = P_{st}(x_A)\sqrt{\frac{\beta}{2\pi\sigma_C^2}}\exp(-\beta(U(x_C) - U(x_A))) \tag{3.7}$$

Using the steady state distribution $P_{st}(x_A) \approx \frac{n_A}{\sqrt{2\pi\sigma_A^2/\beta}}\exp(-\frac{\beta}{2\sigma_A^2}(x - x_A)^2)$ one obtains the rate of escape to be $k(A \to B) = \frac{\Gamma}{\beta} j n_A$:

$$k(A \to B) = \frac{\Gamma}{2\pi\sigma_A\sigma_C}e^{-\beta(U(x_C)-U(x_A))} \tag{3.8}$$

Often, instead of considering escape rates one considers the mean first passage time defined as $< T(x) >= \int_{x_0}^{x_C} P_{st}(x)T(x)$, where the passage times from the particular points are given by:

$$T(x) = \int_x^{x_C} dx' \exp(-\int_{x_0}^{x'} v(y)/D(y)dy)\int_{x_0}^{x'} \frac{\exp(\int_{x_0}^{y} v(y')/D(y')dy')}{D(y)}dy \tag{3.9}$$

We can see the general exponential dependence on the height of the barrier, which has been exploited when formulating the theory in chapter 6.

Many variations of the escape problem, both classical and quantum have been the focus of many scientific careers and resulted in a great number of limits and approximations, such as, for example, the Transition State Theory (TST) (for a review see [102]). The problem of escape from nonequilibrium attractors in many dimensions remains unsolved [103]. For the specific example of stability of a gene expression state, this problem has attracted a lot of attention [38, 104, 105]. We build on the basic intuition offered by Kramers' result to include nonadiabatic effects. Our approach is reminiscent of coupling the escape problem to a two-state system, often referred to as a spin-boson problem [106], where the two-state system is a spin, which interacts with other spins by emissions and absorption of bosons. The analogy between the two problems is clear in the operator formulation

described earlier (Equation 3.4). Once again, there are many possible approaches to this problem, for example that proposed in electron-transfer [80, 107, 108].

In summary of this chapter, I have hoped to give the reader some overview of the used and developed formalism, which is described in detail and with applications in the following chapters.

4

Self-regulating gene: an exact solution

Production of functional biomolecules in the cell is governed by a complex and diverse genetic network involving an intricate set of biochemical reactions. The mathematical description of this network is intrinsically nonlinear because the transcription of DNA is regulated by the binding reactions with the very protein products of the decoding process itself [109]. This description must also be stochastic because the genes are single molecules of DNA and their regulatory proteins are also present often in small numbers. The average behavior of a nonlinear, stochastic system cannot be inferred from macroscopic chemical rate laws alone [40, 39, 38, 37, 36, 35, 34, 33, 22, 110, 90, 83]. In this paper we examine the simplest model of an element of a gene regulatory network and show that its master equation admits an exact solution. In regimes where the binding/unbinding process is not significantly faster than the synthesis/degradation of the proteins, this solution is quantitatively different from the deterministic description [69, 111, 14].

In deterministic models of gene expression the concentration of various transcription factors controls the rate of protein production for a particular gene[75, 87, 111]. The stochastic analysis of gene switches treats the numbers of these various proteins, $n_1 ... n_N$, in a given cell as random variables [76, 88, 23, 18, 105] . If we ignore the mechanistic details of protein biosynthesis with their resulting time delays and mRNA fluctuations [34, 33], we can model each gene as a two state stochastic system. A single gene can then be described by a two component master equation

with one probability distribution $\alpha(n_1, ...)$ corresponding to situations where the DNA is free (*on* state) and a second component $\beta(n_1, ...)$ describing the distribution when the DNA has a repressing protein bound to it (*off* state). The dynamics of these genetic expression probabilities is described by coupled birth-death processes. Birth corresponds to protein synthesis while death occurs via degradation. The rates for protein production g_α and g_β are different for the free and bound states of the DNA. The rate for protein degradation is k, varying linearly with n. If the binding state of the DNA did not change, the stationary probabilities α and β would be described by Poisson distributions with mean values at g_α/k and g_β/k. We show that the time evolution from any initial state of this simple self-repressing switch to the stationary configuration can be written explicitly in terms of hypergeometric functions as in the theory of the threshold switch [39].

4.A The stochastic formulation

In the present model a single gene produces the same protein that represses its own activity. While not often found as an isolated entity, the self-regulating gene is a very common element of biological networks; for example 40% of *E. Coli* transcription factors negatively regulate their own transcription [21]. The master equations for this case are explicitly

$$\frac{d\alpha_n}{dt} = g_\alpha[\alpha_{n-1} - \alpha_n] + k[(n+1)\alpha_{n+1} - n\alpha_n] - hn\alpha_n + f\beta_n \text{ and} \tag{4.1}$$

$$\frac{d\beta_n}{dt} = g_\beta[\beta_{n-1} - \beta_n] + k[(n+1)\beta_{n+1} - n\beta_n] + hn\alpha_n - f\beta_n \text{ for } n \geq 2, \tag{4.2}$$

where α_n and β_n are the individual probabilities that the DNA is unbound and bound, respectively, while immersed in n proteins. h is the bimolecular rate describing the process of repressor binding to the DNA and f is the unimolecular rate describing release of the repressor protein from the repressor site. More generally, h can be a more complicated function of n if, for example, proteins bind as

oligomers [40]. In this case, we consider a mechanism of monomer binding. The binding/unbinding process does not alter the total number of proteins. Since a bound protein is still included in n, there is a need to modify the master equation for the states near $n = 0$. The gene cannot be in a bound state in which there are no proteins in the system ($\beta_0 = 0$). Thus we will use a set of equations in which a degradation reaction will transform the state where the only existing protein is bound (β_1) into the unbound state α_0.

$$\frac{d\alpha_0}{dt} = -g_\alpha \alpha_0 + k[\alpha_1 + \beta_1] \tag{4.3}$$

$$\frac{d\beta_1}{dt} = -g_\beta \beta_1 + k[2\beta_2 - \beta_1] + h\alpha_1 - f\beta_1 \tag{4.4}$$

$$\frac{d\alpha_1}{dt} = g_\alpha[\alpha_0 - \alpha_1] + k[2\alpha_2 - \alpha_1] - h\alpha_1 + f\beta_1. \tag{4.5}$$

4.B An exact solution

The master equations are differential-difference equations for t and n, respectively. The two sets of master equations need to be solved in the appropriate subspaces of n. The general solution may then be determined using the continuity condition at $n = 2$. The solution of equations (4.1) and (4.2) can be described in terms of the generating functions $\alpha(z) = \sum_{n=0}^{\infty} \alpha_n z^n$ and $\beta(z) = \sum_{n=0}^{\infty} \beta_n z^n$, where z lies in the complex unitary circle. The original probabilities for $n \geq 2$ can be recovered as derivatives of these generating functions at z=0: $\alpha(n) = \frac{1}{n!}\frac{d^n}{dz^n}\alpha(t,z)$ and $\beta(n) = \frac{1}{n!}\frac{d^n}{dz^n}\beta(t,z)$. The correct probabilities for the states in which $n < 2$ are calculated by using α_2 and β_2 derived from the generating functions in the modified master equations (4.3),(4.4)and(4.5). Various moments of the distribution, including the average number of proteins can still be expressed in terms of derivatives of these generating functions $\frac{\partial}{\partial z}\alpha(z,t)$ and $\frac{\partial}{\partial z}\beta(z,t)$ evaluated at $z = 1$. Before taking into account the boundary behavior the generating functions satisfy the first order partial differential equations:

$$\frac{\partial\alpha(z,t)}{\partial t} = (z-1)[g_\alpha\alpha(z,t) - k\frac{\partial\alpha(z,t)}{\partial z}] - hz\frac{\partial\alpha(z,t)}{\partial z} + f\beta(z,t) \tag{4.6}$$

$$\frac{\partial\beta(z,t)}{\partial t} = (z-1)[g_\beta\beta(z,t) - k\frac{\partial\beta(z,t)}{\partial z}] + hz\frac{\partial\alpha(z,t)}{\partial z} - f\beta(z,t) \quad . \tag{4.7}$$

The stationary solution of this system of equations is easily obtained. From equation (4.6), we can find β as a function of α and $d\alpha(z)/dz$. Substituting this expression for β in (4.7), a second order differential equation is obtained

$$\frac{d^2\alpha(z)}{dz^2} + p\frac{d\alpha(z)}{dz} + q\alpha(z) = 0 \quad . \tag{4.8}$$

Where the coefficients p and q are

$$p = \frac{g_\alpha + g_\beta + f + h + k - z(g_\beta(1+h/k) + g_\alpha)}{(k+h)z - k} \tag{4.9}$$

$$q = \frac{g_\alpha g_\beta z - g_\alpha(g_\beta + f + k)}{k(k+h)z - k^2} \quad . \tag{4.10}$$

Noting that p and q are rational functions of z with a simple pole at $z_0 = k/(k+h)$ and an irregular singularity at $z = \infty$, we see the structure of this equation corresponds to the confluent hypergeometric equation.

The dependence on z in the numerator of q can be eliminated by making the transformation

$$\alpha(z) = A \ \exp(zg_\beta/k) \ M(a,b,\eta) \quad , \tag{4.11}$$

which leads to the confluent hypergeometric equation in a canonical form. The normalization constant A guarantees that the sum of the probabilities is 1. The solutions are linear combinations of the Kummer functions M and U. The irregular function U does not satisfies the condition $\alpha_n \to 0$ when $n \to \infty$ and therefore is discarded. The resulting generating function α has the Kummer $M(a,b,\eta)$ parameters

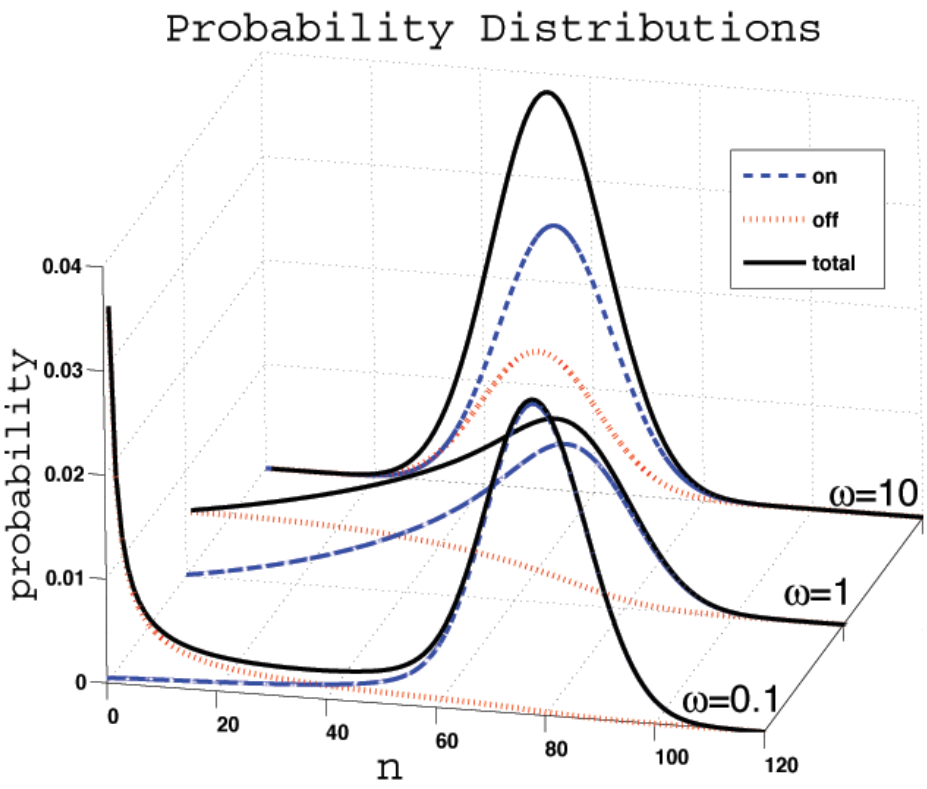

Figure 4.1: The probabilities of the gene expression as a function of the number of proteins n for the *on* state, the *off* state and the total. There are two peaks for small ω, but they converge to a single peak in the adiabatic regime of large ω. $X^{eq} = 100$ and $X^{ad} = 40$.

$$a = 1 + \frac{f}{k+h}(1 + \frac{h\ g_\alpha}{kg_\alpha - (k+h)g_\beta}) \tag{4.12}$$

$$b = 1 + \frac{f}{k+h} + \frac{h\ g_\alpha}{(k+h)^2} \quad , \tag{4.13}$$

and the argument of the function is

$$\eta = -\frac{(g_\beta(1+h/k) - g_\alpha)((k+h)z - k)}{(k+h)^2} \ . \tag{4.14}$$

As described above, α_n's for $n \geq 2$ can be calculated from the derivatives at $z = 0$. Explicitly these are [112]:

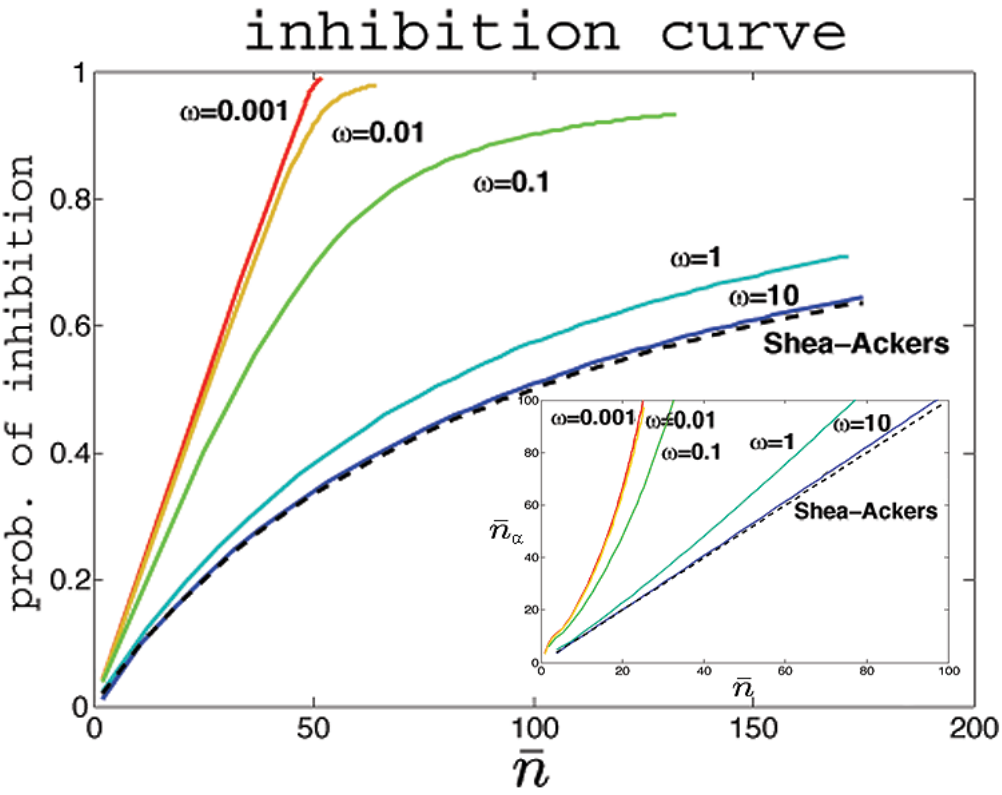

Figure 4.2: Total probability of the DNA being found in the *off* state as a function of the average number of proteins n̄. In the adiabatic limit (large ω) we approach the behavior given by the equilibrium mass action law as in the treatment of Shea-Ackers, where $P_\beta = \frac{\overline{n}}{\overline{n}+X^{eq}}$. $X^{eq} = 100$. In our model we find $P_\beta = \frac{\overline{n_\alpha}}{\overline{n_\alpha}+X^{eq}}$ exactly. The average number of proteins present when DNA is in the *on* state $\overline{n}_\alpha$ is different from $\overline{n}$, which includes the average number of proteins when the gene is *off* (inset).

$$\alpha_n = \frac{A}{n!}\sum_{s=0}^{n}\binom{n}{s}(g_\beta)^{n-s}\frac{d\eta}{dz}^s\frac{(a)_s}{(b)_s}M(a+s,b+s,\eta_0) \quad . \tag{4.15}$$

$\beta(z)$ can be calculated directly from (4.6) and the probabilities β_n for $n \geq 2$ are again derivatives at $z = 0$. It is worth noticing that in the limit where there is no protein synthesis at all in the *off* state ($g_\beta = 0$), there is only one non-zero term in the series for α_n ($s = n$). This leads to a simple expression for $\alpha_n = \frac{A}{n!}\frac{d\eta}{dz}^n\frac{(a)_n}{(b)_n}M(a+n,b+n,\eta_0)$ and $\beta_n = \frac{A(k+h)}{fn!}\frac{d\eta}{dz}^n[(\frac{hg_\alpha}{(k+h)^2}+b-1)\frac{(a)_n}{(b)_n}M(a+n,b+n,\eta_0) - (b-1)\frac{(a-1)_n}{(b-1)_n}M(a-1+n,b-1+n,\eta_0)]$.

The normalization constant A is determined by $\sum_{n=0}^{\infty}\alpha_n + \sum_{n=0}^{\infty}\beta_n = 1$. These

sums can be expressed in terms of $\alpha(1)$ and $\beta(1)$ and appropriate corrections to account for the states whith $n < 2$.

$$\alpha(1) + \beta(1) - \alpha(0) - \frac{d}{dz}\alpha(0) - \beta(0) - \frac{d}{dz}\beta(0) + \alpha_1 + \beta_1 + \alpha_0 = 1. \qquad (4.16)$$

4.C Comparison to the deterministic model

With these analytical solutions in hand, we are now in position to compare this exactly solved model with the commonly used deterministic mass-action approximation introduced by Shea and Ackers [69]. To simplify the discussion we introduce the following parameters: $\omega = f/k$, $X^{eq} = f/h$, and $X^{ad} = (g_\alpha + g_\beta)/(2k)$. The parameter ω measures how rapidly the DNA state can equilibrate in its proteomic cloud in comparison to the characteristic time for protein degradation, which measures how fast the cloud itself fluctuates. X^{eq} is the equilibrium constant of the binding/unbinding process. X^{ad} is a measure of the protein concentration, indicating the number of proteins when the system is half-inhibited.

The probability distributions for the protein number given the gene state (the total distribution $\alpha_n + \beta_n$, α_n for the *on* state, and β_n for the *off* state) are shown in figure 1. The values of the switch characteristics used for the figure are $X^{eq} = 100$ and $X^{ad} = 40$ and $g_\beta = 0$. These are typical values of the equilibrium switching threshold and mean protein copy number found in a small cell like *E. Coli*. For small values of ω the total probability distribution exhibits a two peak structure, at g_β/k and g_α/k, corresponding to repressed protein production, when the DNA has protein bound at small n, and to the higher production from the free DNA at large n. In this limit, the *on*-state behaves almost like an independent birth and death process since binding/unbinding become the slowest process in the system. Increasing the value of ω shifts both peaks to intermediate values, until there is only one peak at large ω. In the large ω limit, protein binding and unbinding becomes extremely fast. This "adiabatic" regime should be equivalent

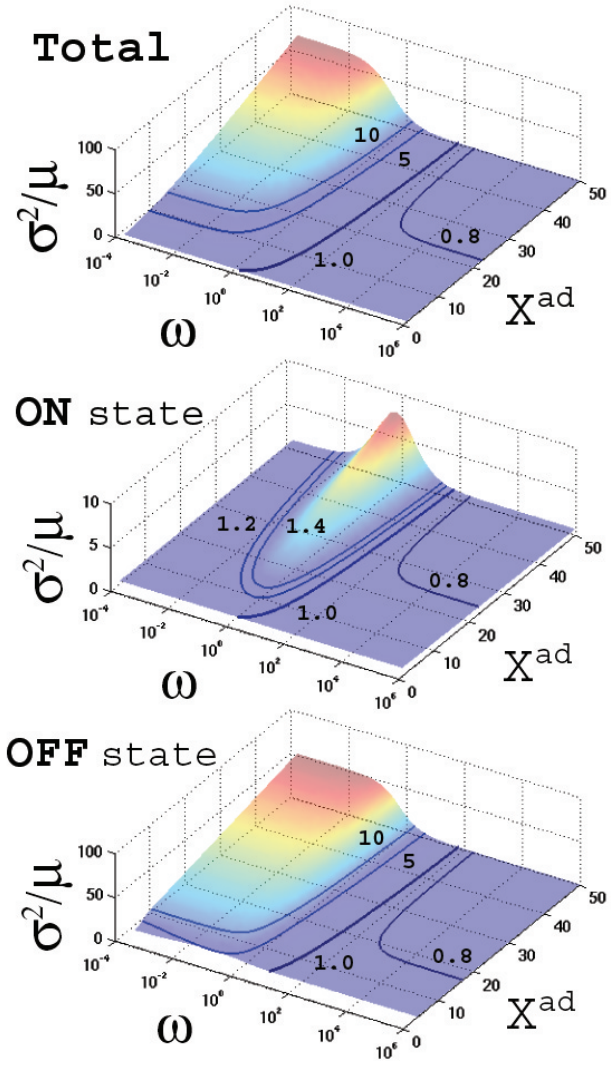

Figure 4.3: The Fano factor $F = \frac{\sigma^2}{\mu}$ for the self-repressing switch. Along the curve $F = 1$, all distributions are Poisson-like. The Total distribution is independent of DNA state while the *on* state has the DNA free and the *off* state has protein bound to DNA. In the limit of large ω the adiabatic regime is reached, with an almost Poisson behavior. This regime should be equivalent to the Shea-Ackers model. For intermediate ω's the overall fluctuations are large and therefore strongly deviate from Poisson. In the *on* state, the distribution tends to Poisson behavior for ω very small, since the system behaves almost like a birth-death process. $X^{eq} = 50$, $g_\beta = 0$.

to the Shea-Ackers model in which the gene itself is taken to have an equilibrated average probability of being *on* or *off*. Most of the characterized genes are known

to have high values of the adiabaticity parameter (*e.g.* when calculated from the transcription initiation rate obtained from [113]). Some systems exist, however, where ω is of order one (*e.g.* Cro protein in the λ-phage, parameters obtained from [104]). Also, the non-adiabatic regime may be important *in vivo*. For example, several *in vivo* mechanisms suggest that some proteins may be slow binders.

A more detailed understanding of the deviations from the Shea-Ackers approximation can be made by noting that in the Shea-Ackers model the probability of inhibition (P_β) is given by the equilibrium law of mass action as a function of the concentration of repressors. This concentration can be calculated using the first moments of the distribution $\frac{d}{dz}\alpha(z)$ and $\frac{d}{dz}\beta(z)$ at $z = 1$, again with the corrections from the terms with $n < 2$. Fig.2 shows how the exact solution for the master equation finally converges to the equilibrium approximation used by Shea and Ackers ($P_\beta = \frac{\overline{n}}{\overline{n}+X^{eq}}$) in the limit of large ω.

To directly probe the effect of fluctuations, figure 3 shows the probability distributions compared to those that would arise from Poisson statistics: a) independent of DNA state, b) when the DNA is free (α_n), and c) when the DNA has protein bound (β_n). The Fano factor $F = \frac{\sigma^2}{\mu}$ is plotted as a function of ω and X^{ad}, where μ and σ are the mean and standard deviation of the probability distributions. This factor would be one if the processes were purely Poisson. Notice that for very small ω, the Fano factor does limit to one when the DNA is in the *on* state. As discussed above, this is expected since, in this limit, the *on*-state behaves almost like an independent birth and death process. The overall fluctuations are however quite large for intermediate ω's and therefore their contributions cannot be ignored in the overall mechanism. Indeed the Fano factor remains large even at ω values large enough for the probability of inhibition to agree with the equilibrium behavior. This shows DNA binding noise cannot be neglected.

In the large ω regime (tending to the adiabatic limit), the Fano factor for the three distributions tends to values slightly smaller than one. This indicates an almost Poisson behavior as one would expect for near-macroscopic kinetics.

4.D Discussion

The exact solution presented here for the self-regulated gene in a stationary regime establish the basis for more complex problems yet to be solved. It provides an important analytical tool to understand the underlying mechanism governing these genetic networks. Already for this simple system, we notice that fluctuations become important for a large region of the parameter space. Figure 3 makes it clear that fluctuations cannot be ignored unless protein binding and unbinding are exceedingly faster than any other relevant time scale in the problem. Noise from binding/unbinding events dominates shot noise of protein synthesis and degradation up to quite high values of the adiabaticity parameter. Figures 1 and 2 also demonstrate the effects of fluctuations. For small ω, binding is slow and therefore the stationary solution for the gene probabilities shown in figure 1 has two well defined peaks. One peak corresponds to the repressed protein production (DNA with protein bound) and the other to the higher protein production (free DNA). As protein binding and unbinding become faster, these two peaks converge towards each other. Figure 2 shows how in this nonequilibrium system the probability of DNA being found in the protein-bound state deviates from the equilibrium mass action result. The self-repressing gene can become strongly anti-cooperative owing to non-adiabatic effects normally neglected in theories of gene regulation.

Some features of the genetic switch such as mRNA fluctuation and the time delays resulting from transcription and translation are not explicitly captured by this model. Although they might be essential in some cases, they may not always dominate the process of regulation. In prokaryotes, where there is no nucleus, transcription and translation occur within the same compartment, and mRNA is almost immediately translated [109]. Also many cases are being discovered where the regulation is performed by the RNA itself [114]. In cases like these, the approximation of having the synthesis of the transcription factors as one stochastic process seems plausible. This formulation of the problem of genetic regulation

and its analytical solution will help the study of the specific cases where mRNA fluctuations and time delays play a determinant role.

While an otherwise isolated non-interacting self-regulating gene is a biological rarity, it would be straightforward to construct in the laboratory. The exact solution presented here would then make such an experiment a beautiful simplified system for understanding the importance of fluctuations that govern gene networks.

4.E Acknowledgements

The text and data of Chapter 4, in full, has been published in "Self-regulating gene: an exact solution" by J. E. M. Hornos, D. Schultz, G. C. P. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic and P. G. Wolynes in Phys. Rev. E (**72**), 051907-1-5, (2005). The dissertation author was a contributing investigator and author of this article.

5

Self-Consistent Proteomic Field Theory of Stochastic Gene Switches

5.A Introduction

Genetic switch systems are an elementary means of regulatory control present in every living organism. Their complexity and details differ, but the general mechanism of the expression of a given gene being regulated by proteins, is believed to be universal [115]. They are building blocks of larger regulatory elements: genetic networks and signaling cascades. The pathways by which these systems operate is passed on from generation to generation. Understanding their stability and characteristics is therefore fundamental. A lot of previous work has considered a deterministic description of genetic switches [69, 14]. The need for a stochastic treatment of genetic switches due to the single copy of the DNA molecule and multiple protein molecules in the cell, has been largely recognized [104, 76].

The most general way of accounting for non deterministic processes is to write down the master equation for a given system. To define the state of the switch one must specify the DNA binding states of particular genes and the number of proteins of each type. The probability distribution even of a single switch consisting of two genes, the product proteins of which act as regulator proteins for the system, may not be determined exactly and approximations must be considered [38, 25, 104].

Several approaches to account for the probabilistic nature of chemical reac-

tions have been undertaken, ranging from the Langevin description of single genes [38], and two interacting gene switches [25], to the master equation reduced to a Fokker-Planck equation considerations [76, 24]. A dynamical action formulation has also been used [104] to determine the lifetimes of states of the switch. A popular alternative to purely analytical methods, which often need to make approximations or are limited to very simple model systems, has been to conduct stochastic simulations of genetic switches. Two types of simulations are mostly used. In the first the randomness of the system is introduced by means of a Monte Carlo algorithm with fixed time step [23]. The second is based on the Gillespie algorithm [91] to predict the probability of a given reaction occurring [75]. For single gene systems, stochastic simulations have shown that stochasticity in the system is responsible for the bimodal probability distributions [46], observed experimentally. These methods prove very useful, as they allow us to test the theoretical predictions on model systems, which might be hard to build experimentally. However this approach often does not enable us to gain intuition or insight into the mechanisms behind the functioning of the system. The aim of the present work is to gain a better and deeper understanding of the device physics of genetic switches. We therefore, contrary to many important previous discussions [88, 87, 68] do not present a specific concrete biological system, but discuss generic behavior and try to understand its sources. Our approximation also allows for an exact solution of a broad class of genetic switch systems without any further assumptions and with little computational effort. Hasty et al [14] present an overview of the existent theoretical approaches.

A popular approximation, assumes the DNA binding state reaches equilibrium much faster than the protein number state. Therefore the adiabatic approximation is often considered [69, 104, 111], allowing for a thermodynamic treatment [69] of the DNA binding state. The protein number fluctuations are then treated stochastically. Even before the statistical thermodynamics approach of Shea and Ackers [69] using partition functions, much previous work assumed the DNA binding and

unbinding can simply be accounted by an equilibrium constant, since the relaxation timescales for equilibration of the DNA state are much larger than those of the protein numbers, which require protein synthesis and degradation to change. The partition function approach has also been successful at looking at logic gates build from switches [35]. The adiabatic approximation is believed to hold true in many cases, judging by the experimental parameters of biological switches [111]. But as the experiments of, for example Becskei et al [22] show, not all switches need function in the adiabatic limit and the non-adiabatic limit may result in new phenomena. We therefore consider a wide range of parameter ratios in our discussion.

In this paper we explore more fully an approximation, previously used by Sasai and Wolynes [40] for the variational treatment of the problem, the self-consistent proteomic field (SCPF) approximation. Within this approximation one assumes the probability of finding the switch in a given state is a product of probabilities of states of individual genes. One can then solve the steady state master equation for the probability distribution of many regulatory systems exactly. We discuss the approximation and present a detailed study of different classes of genetic switches, some of which have never previously been considered theoretically. We consider several particular features of such systems, found in known switches, separately to be able to characterize their contributions to the behavior of the whole system. To be specific, starting from a symmetric toggle switch, we go on to compare the effects of multimer binding and of the production of proteins in bursts on the stability of the switch.

The stochastic effects prove to be modest for symmetric switches without bursts, especially if the genes have a basal production rate. We find the deterministic and stochastic SCPF solutions to have similar probabilities of particular genes to be on and mean numbers of proteins of a given species in the cell. However in the non-adiabatic limit, when the unbinding rate from the DNA is smaller than the death rate of proteins, the probability distributions have two well defined

peaks, unlike in the deterministic approximation or adiabatic limit of the stochastic SCPF solution.

We also show the effect of stochasticity on the observables becomes more apparent when proteins are produced in bursts. In these types of switches, the definition of the adiabatic limit, which was clear for the switches in which proteins are produced separately, is no longer simple. Our discussion shows that the properties of genes often analyzed in the deterministic limit, may be strongly influenced by stochasticity in this case. Randomness in a biological reaction system leads to quantitative and in many examples even qualitative changes from predictions of deterministic models.

We also discuss the differences in the behavior of an asymmetric and symmetric switch. We point to the mechanisms resulting in different types of bifurcations and show how they are influenced by noise. Within the SCPF approximation switches that are regulated by binding and unbinding of monomers, do not have regions of bistability. This holds true for both symmetric and asymmetric switches. When proteins are produced individually rather than in bursts, fast unbinding from the DNA can effectively minimize the destructive effect of protein number fluctuations on the stability of the DNA binding state. Furthermore a detailed analysis of the probability distributions show they have long tails and are far from Poissonian in both the adiabatic and non-adiabatic limit. We discuss the properties of the system in terms of clouds of proteins buffering the DNA. We show how fast or slow DNA binding characteristics and protein number fluctuations influence the stability of the buffering clouds leading to specific emergent behavior of observables. Throughout the paper a comparison is made between results of the exact stochastic solution with solutions of deterministic kinetic equations for the system, within the self-consistent proteomic field approximation.

We establish a base of potential building blocks of more complicated switches and systems, such as networks and signaling cascades, for which an exact solution within the present approximation can also be obtained. A detailed discussion of

these larger systems will be the topic of another paper. We also present limitations of the present style of analysis where exact solutions are not possible.

There are two aims of this paper. The first is to discuss the self consistent field approximation and show that it has an exact solution which may be extended to a large class of systems. This approximation lets one deal in a straightforward and computationally inexpensive manner with the effect of random processes on genetic networks. The second is to discuss the many components of biological switches present in nature and in engineered systems, in the necessary stochastic framework.

5.B The Self-Consistent Proteomic Field Approximation

The basic mechanism of gene transcription regulation in prokaryotes may be reduced to the binding and unbinding of regulatory proteins, repressors and activators, to the operator site of the DNA. If we use this simplified treatment, which neglects extra levels of regulation, such as the binding of RNA polymarase, effectively each gene can be described as being either in an active (on) state, when the repressor is unbound (activator bound), or in an inactive (off) state, with the repressor bound (activator unbound). The stochastic system of a single gene and its product proteins is described by the joint probability distribution $\vec{P}(n,t) = (P_1(n,t), P_2(n,t))$ of the number of product proteins in the cell n, and the DNA binding site state: on (protein not bound)- 1, or off (protein bound) - 2. To conserve probability $\sum_n (P_1(n,t) + P_2(n,t)) = 1$.

If one considers two interacting genes, the description in terms of a joint probability vector needs to be extended to four states: both genes may be on, or off, or one of the genes may be on, the other off. If the two genes do not interact, as would be the case for two self regulatory proteins, the probability of a finding the two gene system in a given state, defined by both the number of product proteins and the DNA binding site state, would be the the product of the states of particular genes

$P_{jj'}(n_1, n_2; t) = P_j(n_1; t)P_{j'}(n_2; t)$. This is generally not true for two interacting proteins, as is the case in a genetic switch. However, as a first approximation to the problem, one can ignore correlations between the spaces of the two genes and assume the space of the switch is a sum of spaces of the genes that compose it. Since we are looking for solutions in which the symmetry of the system is broken and different behaviors of the on and off state of a gene are possible, we must allow for different probability distribution functions for the on and off states. This is analogous to the unrestricted Hartree approximation in quantum mechanics, where allowing different spatial functions for spin up and spin down states results in breaking of the symmetry of the bound molecular orbital solution to the dissociated solution of two separate hydrogen atoms with opposite spin states for large internuclear distances. We therefore allow for multiple solutions for a given set of parameters. The total probability of having a given gene state i and n_i proteins of that type is simply given by $P_j(n_i, n_{i'}) = P_{j,j'=0}(n_i, n_{i'}) + P_{j,j'=1}(n_i, n_{i'})$.

The self-consistent approximation is a crude approximation since in the case of the genetic switch, the state of a given gene is determined by the number of protein products of the other gene. However, within this approximation, one can solve the master equation for the probability distribution exactly without any further approximations. This yields a powerful computational tool, which simultaneously gives useful insight.

5.C The Toggle Switch

For clarity of exposition, we show how the problem may be solved exactly within the self-consistent proteomic field approximation on a well defined system of the toggle switch. We then expand the method to apply to other systems. The elementary system we use as an example is composed of two genes, labeled 1 and 2, as presented in Fig. 5.1. Gene 1 produces proteins of type 1 which, act as regulatory proteins, say repressors, on gene 2. The product of gene 2, proteins of

type 2, in turn repress gene 1. In this simplified model, we assume that protein production occurs instantaneously upon unbinding of the repressor. For now, we assume that repressor proteins bind as dimers, since that is a common scenario in biological systems, but we do not treat dimerization kinetics explicitly. For simplicity the coupling form between the genes responsible for binding will be taken to be of the form $h_i n_{3-i}^p$, where p is the order of the multimerization of the repressor. This form is a small approximation to the more exact $h_i n_{3-i}(n_{3-i}-1)...(n_{3-i}-p+1)$. We have checked that using the simpler monomial does not influence the results in any regime discussed. We also do not account for the existence of mRNA molecules and the consequent time delays owing to their synthesis as intermediates. The extensions of the model are discussed later.

Within the self-consistent proteomic field approximation the set of master equations for the corresponding system is of the form:

$$\begin{aligned}
\frac{\partial P_1(n_i)}{\partial t} &= g_1(i)[P_1(n_i-1) - P_1(n_i)] + k_i[(n_i+1)P_1(n_i+1) - n_i P_1(n_i)] + \\
&\quad -h_i n_{3-i}^2 P_1(n_i) + f_i P_2(n_i) \\
\frac{\partial P_2(n_i)}{\partial t} &= g_2(i)[P_2(n_i-1) - P_2(n_i)] + k_i[(n_i+1)P_2(n_i+1) - n_i P_2(n_i)] + \\
&\quad +h_i n_{3-i}^2 P_1(n_i) - f_i P_2(n_i)
\end{aligned}$$

for $n \geq 1$ where the $i = 1, 2$ refers to the gene label. $P_1(n_1)$ describes the probability of gene 1 being in the on state and there being n_1 protein molecules of type 1 in the cell. The first term on the right hand side of the equations describes the production of proteins of type i with a production rate $g_j(i)$, where j=1,2, depending on whether the gene is in the on or off state. The second term accounts for the destruction of proteins with rate k_i. The binding of repressor proteins produced by the other gene is proportional to the number of dimer molecules present in the system n_{3-i} with rate h_i. We assume unbinding occurs with a constant rate f_i. Binding and unbinding contributes to the kinetics of the DNA binding states, as described by the last two terms. This set is supplemented by the $P_j(n_i = 0)$ equations to account for boundary conditions.

$$\begin{aligned}
\frac{\partial P_1(n_i=0)}{\partial t} &= -g_1(i)P_1(n_i=0)+k_iP_1(n_i=1) \\
&\quad -h_in_{3-i}^2P_1(n_i=0)+f_iP_2(n_i=0) \\
\frac{\partial P_2(n_i=0)}{\partial t} &= -g_2(i)P_2(n_i=0)+k_iP_2(n_i=1) \\
&\quad +h_in_{3-i}^2P_1(n_i=0)-f_iP_2(n_i=0)
\end{aligned}$$

For convenience, let us define $\sum_{n_i} P_j(n_i) = C_j$, the probability of finding the DNA binding site in a given state. One can now sum the $P_j(1)$ equations over the number states of the 2nd protein with $P_1(2) + P_2(2)$, and likewise the $P_j(2)$ equations. Due to the SCPF approximation, the only term affected is the repressor binding term $h_1(n_2^2)$, and since $\sum_{n_2} P_1(2) + P_2(2) = 1$, the summation results in $\sum_{n_2} h_1(n_2^2)(P_1(2) + P_2(2)) = h_1(C_1(2) < n_{12}^2 > +C_2(2) < n_{22}^2 >) = h_1F(2)$, where $< n_{j2}^2 >$ is the second moment of the number distributions of type 2 proteins produced when gene 2 is in the j-th state. The equations of motion of the moments of the probability distribution are of the form:

$$\begin{aligned}
\frac{\partial C_j(i) < n_{ji}^k >}{\partial t} &= g_j(i)[< (n_{ji}+1)^k > - < n_{ji}^k >]C_j(i) + \\
&\quad +k_i[< n_{ji}(n_{ji}-1)^k > - < n_{ji}^{k+1} >]C_j(i) + \\
&\quad +(-1)^j h_iF(3-i) < n_{1i}^k > C_1(i) + \\
&\quad +(-1)^{j+1} f_i < n_{2i}^k > C_2(i)
\end{aligned}$$

The steady state equations for the moments of the distributions that follow are closed form, the n_i^{th} order moment equation of motion depends only on the lower moments of the i^{th} gene and n_{3-i}^2.

To analyze the behavior of switches we introduce the following scaled parameters: the adiabaticity parameter $\omega_i = f_i/k_i$, which represents the characteristic rate of change of the DNA state compared to the characteristic rate of change in protein number, $X_i^{eq} = f_i/h_i$ measures the tendency for proteins to be unbound from the DNA, $X_i^{ad} = (g_1(i)+g_2(i))/(2k_i)$ the effective production rate and $\delta X_i^{sw} = (g_1(i) - g_2(i))/(2k_i)$ distinguishes between the two DNA states in terms

of protein dynamics. We present a detailed derivation of the moment equations in Appendix A.

The resulting equations for the zeroth moments couple to the higher moments by the interaction function $F(i)$. These lower moments can be solved self-consistently. The resulting solution predetermines all the other moments, which completely describe the probability distribution. Each gene therefore couples to the other gene by the influence of the self-consistently generated proteomic field. One could define the generating function and calculate the probabilities of having a given DNA binding state j for the i^{th} gene when there are n_i proteins of type i in the cell. In practice, it is easier to go back to the steady state master equation and solve directly for the probability distributions than sum an infinite number of moments. Rewriting the steady state master equation one gets:

$$\begin{aligned}
P_1(n_i) &= \frac{1}{X_i^{ad} + \delta X_i^{sw} + \omega_i \frac{F(3-i)}{X_i^{eq}} + n}[(X_i^{ad} + \delta X_i^{sw})P_1(n_i - 1) + \\
&\quad +(n_i + 1)P_1(n_i + 1) + \omega_i P_2(n_i)] \\
P_1(n_i = 0) &= \frac{1}{X_i^{ad} + \delta X_i^{sw} + \omega_i \frac{F(3-i)}{X_i^{eq}}}[P_1(n_i = 1) + \omega_i P_2(n_i = 0)] \\
P_2(n_i) &= \frac{1}{X_i^{ad} - \delta X_i^{sw} + \omega_i + n}[(X_i^{ad} - \delta X_i^{sw})P_2(n_i - 1) + \\
&\quad +(n_i + 1)P_2(n_i + 1) + \omega_i \frac{F(3-i)}{X_i^{eq}} P_1(n_i)] \\
P_2(n_i = 0) &= \frac{1}{X_i^{ad} - \delta X_i^{sw} + \omega_i}[P_2(n_i = 1) + \omega_i \frac{F(3-i)}{X_i^{eq}} P_1(n_i = 0)]
\end{aligned}$$

These sets of equations give recursion relations for $P_j(n_i)$ which one can use to express $P_j(n)$ as a function of $P_1(0)$ and $P_2(0)$. The normalization condition $\sum_{n_1}(P_1(n_1) + P_2(n_1)) = 1$ gives $P_j(0)$ in term of constants and the result is the probability function $P_j(n)$ as a series. The SCPF approximation reduces the two

gene problem to a one gene problem parametrized by the moments of the second gene, which can be worked out independently, as we have already shown and are represented by $F(2)$, which is a constant in terms of this calculation. To see the effect of the stochastic nature of the system we compare the exact solutions of the self consistent field approximation equations to the results that would follow from deterministic kinetic rate equations for the number of proteins of each type and the fraction of on/off DNA binding states for each gene:

$$C_1(i) = \frac{X_i^{eq}}{X_i^{eq} + n^2(3-i)}$$

$$n(i) = X_i^{ad} + \delta X_i^{sw}(C_1(i) - C_2(i))$$

where $n(i)$ is the number of proteins of type i present in the cell. The exact SCPF equations reduce to the deterministic kinetic equations in the limit of large ω and X^{ad} for the case discussed above. The $F(3-i)$ term in the stochastic SCPF equations is replaced by the $n^2(3-i)$ term in the deterministic kinetic rate equations. For the toggle switch, where repressors bind as dimers it is easily shown that the interaction functional may be rewritten in the form:

$$\begin{aligned} F(i) = {} & (X_i^{ad})^2 + X_i^{ad} + (\delta X_i^{sw})^2 + \delta X_i^{sw}(C_1(i) - C_2(i))(1 + 2X_i^{ad}) + \\ & -4\omega_i(\delta X_i^{sw})^2 \frac{C_1(i)C_2(i)}{\omega_i + C_1(i)} =< n(i) >^2 \frac{\omega_i + 1}{\omega_i + C_1(i)} + < n(i) > \end{aligned}$$

which in the large ω limit reduces to $F(i) =< n(i) >^2 + < n(i) >$. So for large mean numbers of proteins present in the cell, which corresponds to large effective production rates X^{ad}, $< n(i) >$ of the order of hundreds is a small correction to $< n(i) >^2$. We therefore reproduce the deterministic kinetics result. As shown by Sasai and Wolynes [40] the difference in the probability that gene 1 is active and that gene 2 is active, $\Delta C = C_1(1) - C_1(2)$, plays the role of an order parameter. We can now consider a family of switches and discuss their stability, sensitivity of regions of bistability to control parameters and types of bifurcations.

5.D The Symmetric Toggle Switch

For pedagogic purposes, we will start by analyzing the single symmetric toggle switch, such as discussed above in which repressors bind as dimers, with $\omega_1 = \omega_2 = \omega$, $X_1^{ad} = X_2^{ad} = X^{ad}$, $\delta X_1^{sw} = \delta X_2^{sw} = \delta X^{sw}$ and $X_1^{eq} = X_2^{eq} = X^{eq}$, as it is the most intuitive and shows the most generic behavior. It is an academic example, as even individual genes in switches engineered in the laboratory mostly have different chemical parameters. Yet a lot can be learned from this simple system.

5.D.1 The general mechanism of the phase transition

Figure 5.2 shows the phase diagrams for the system, $|\Delta C|$ as a function of reservoir protein number and the adiabaticity parameter for the exact SCPF equations for growing values of the parameter describing the tendency that proteins are unbound from the DNA, X^{eq}. The deterministic kinetics and exact SCPF approximations give qualitatively similar results. The analogous deterministic kinetic phase diagrams agree with the SCPF solutions in the large ω and X^{ad} limit, hence they become more similar with growing X^{eq}, as the bifurcation occurs at larger effective production rates for larger X^{eq}. For large fluctuations and a small unbinding rate, neither gene 1 nor gene 2 is favoured and the probability of a given gene to be on is determined solely by the effective production rate of the other gene and decreases in a quadratic manner as the number of repressor proteins grow (Fig. 5.3). Since the switch is symmetric, the system has one stable state, $\Delta C = 0$, where the probabilities of the genes to be on are equal. As the relative protein number fluctuations get smaller and the DNA unbinding rate grows, a proteomic cloud buffers the repressed gene, keeping it repressed. The symmetry of the system is broken and the solution bifurcates into two separate basins of attraction. For the stochastic SCPF equations the bifurcation takes place for larger effective production rates (larger X^{ad}), than for the deterministic equations, even in the large ω limit, which depicts their sensitivity to fluctuations.

The critical number of reservoir proteins necessary for the bifurcation of the solution to take place is the same in both approximations and is determined by $< n >_c = (X^{eq})^{\frac{1}{2}}$ (Fig. 5.3). In the discussed example $< n >_c = 32 = 1000^{\frac{1}{2}}$, for $X^{eq} = 1000$. For the deterministic kinetic switch the bifurcation takes place when $C_1(i) = (1 + < n(3-i) >^2 / X^{eq})^{-1} = 0.5$, due to the simple form of the interaction function equal to $< n(3-i) >^2 = (2X^{ad}C_1(3-i))^2$. So $C_1(i) = 0.5$ is equivalent to the $< n(3-i) >^2 / X^{eq} = 1$. In a noisy system larger effective production rates are needed to achieve the critical value of proteins. The interaction function in this case may be written as $F(i) =< n(i) >^2 \frac{\omega+1}{\omega+C_1(i)} + < n(i) >$, and $\frac{\omega+1}{\omega+C_1(i)} \geq 1$, always. So at $< n >_c$, $F(3-i)/X^{eq} > 1$ and the probability of the genes to be on is smaller than 0.5, therefore $C_1^{biff,SCPF}(i) < C_1^{biff,kin}(i)$. The mechanism of the bifurcation requires the two genes to be more likely to be unbound than bound for the phase transition to take place. The curvature of the nullclines presented in Fig. 5.2 can be simply worked out to be of the form $\omega = \frac{\zeta_1}{\xi_1 X^{ad2} + \xi_2 X^{ad} + \zeta_2} - \xi_2$, with ζ_i, ξ_i constants determined by the specific value of $C_1(1)$, $C_1(2)$.

5.D.2 Adiabaticity parameter dependence

As the adiabaticity parameter decreases the area of phase space which corresponds to multiple solutions decreases (Fig. 5.2). For very small values of the adiabaticity parameter, there exists only one solution which corresponds to a state in which the two genes are off. The value of ω below which only one solution exists decreases with the tendency for proteins to be bound, but exists for all values of X^{eq}. Therefore if the two genes have very high repressor binding affinities, the critical number of proteins necessary for the phase transition to take place cannot be formed, even for very high production rates. This region of parameter space where one solution is possible corresponds to a situation in which a buffering proteomic cloud may not form, due to a very fast destruction rate of proteins or a very small unbinding rate from the DNA. The critical number of proteins necessary for the bifurcation to occur grows with the tendency for proteins to be unbound from

the DNA (X^{eq}), as the cloud buffering the genes needs to be bigger and exhibit smaller relative protein number fluctuations, which effectively decrease with the growth of the adiabaticity parameter. This is further discussed in terms of the probability distributions. Therefore a monostable solution exists at all values of the effective growth rate, X^{ad}, for larger values of ω at large X^{eq} than at smaller X^{eq} values. The bifurcation point is a result of competition between the number of reservoir repressor proteins and the tendency for proteins to be unbound from the DNA. This is clear from the dependence of the number of proteins present in the cell at the bifurcation point on the relative values of X^{ad} and X^{eq}, but not the adiabaticity parameter ω.

5.D.3 Mean protein numbers

The total number of proteins present in the cell, produced both in the on and off state, asymptotically away from the bifurcation points is the same for the deterministic and stochastic approximations, and it is given by $< n(i) >= 2X^{ad}$, when $C_1(1) \approx 1$ the probability of the gene to be on is close to unity. The number of proteins of a given type present in the cell, when the gene that produces them is in the on state is always considerably smaller in the noisy system than the deterministic case (Fig. 5.3 *C*). Since the production rate in the off state was assumed zero, in the deterministic case no proteins of a given type are present in the cell if the gene is in the off state, unlike in the noisy system. Therefore the number of proteins in the deterministic system is nonzero only if the gene is on. But interaction of the DNA binding state with the proteins buffering it, results in a residual number of proteins present in the off state, for all values of ω. The region of bistability of the switch in parameter space grows as the binding rate increases with respect to the unbinding rate, stabilizing the DNA binding states. As the susceptibility of the system to fluctuations increases, the deterministic equations prove to be a poor approximation to describe the state of the system.

5.D.4 Gene-buffering proteomic cloud interactions

The stochastic nature of the system manifests itself also at the DNA level (Fig. 5.2). As the tendency for proteins to be unbound from the DNA grows, the area of parameter space, where multiple solutions are possible decreases, since a larger number of proteins is needed to reach a state in which two genes are more likely to be repressed (protein bound state), than at small X^{eq}. For small unbinding rates or large binding rates, regardless of the ratio of the rate of unbinding of repressors from the DNA to protein degradation, bistability requires smaller numbers of proteins, which correspond to larger relative fluctuations, than for large X^{eq}. Therefore a larger unbinding rate relative to the binding rate makes the system more susceptible to protein number noise. Competition between X^{eq} and $< n(i) >$ results in X^{eq}, for a given nullcline, being a parabolic function of X^{ad}, for the dimer binding case, with coefficients determined by ω and $C_1(i)$. This is easily generalized to higher order functions for higher order (p) oligomers, and results in p-order dependence. The switching region, by which we mean the region of parameter space between the bifurcation point and $\Delta C > 0.9$ decreases as the binding and unbinding rates become comparable (X^{eq} decreases). As discussed above, the probability of the genes to be on at the bifurcation point tends to 0.5 as the adiabaticicty parameter grows (Fig. 5.3), therefore the probability to be on has to increase by a smaller ΔC to reach $C_1(i) = 1$. Therefore the switching region decreases also as the unbinding rate from the DNA grows, since smaller effective production rates are needed to reach $\Delta C = 1$, than for small ω. Small values of ω correspond to large fluctuations in the DNA binding state, as well as the protein number state and result in destabilizing the gene-buffering protein cloud interactions. Hence very large effective production rates are needed for $\Delta C > 0.9$. Therefore the DNA unbinding rate must become considerably faster compared to the protein degradation rate for the switch to have two stable solutions in a large region of parameter space.

5.D.5 The probability distributions

A better understanding of the bifurcation can be gained from examining the probability distributions. Figures 5.4 A and B and 5.4 C and D show the evolution of the probability distributions of gene 1 and gene 2, respectively, to be on and off as functions of X^{ad}. The peak of the distribution decreases and the width spreads out as the control parameter grows, until it reaches the bifurcation point at $X^{ad} = 44$. Then the value of the probability function corresponding to the most probable number of proteins grows again. The spread of the functions grows as the effective production rate in the on state increases, however narrows with the increase of the adiabaticity parameter, as would be expected, since the DNA state fluctuations become smaller with ω. The average number of proteins in the cell in the on state ($\Delta C > 0.9$) does not show a dependence on ω. Yet as the unbinding rate from the DNA becomes very fast compared to the protein number fluctuations, the system switches often between the two states, hence a large number of proteins is present even in the off state. This results in a two peak - bimodal probability distribution (Fig. 5.4). If the DNA unbinding rate is small, the protein number characteristics follow the DNA state having time to reach a steady state within each well, before the DNA binding site switches into the other state, so the number of proteins in the off state falls to zero (Fig. 5.5 A and B). If ω is large, random fluctuations in the DNA state do not change the effective state of the system, since a residual high mean protein number is present even in the off state. In such a case lower effective production rates than for small ω result in higher protein yields and what follows smaller switching regions.

For small ω one might expect Poisson distributions of proteins in each of the DNA states, since the unbinding rate from the DNA is smaller than the protein degradation rate, so the proteins may reach a steady state without the DNA state changing. Hence, effectively proteins would feel only one well and be subject to a birth death process. However this is not true. The difference between the exact solution and a solution obtained within a Poissonian approximation to the state

of the system is surprisingly large, owing to the skewed tails of these distributions. Figure 5.5 C and D compares these probability distributions with distributions for the same system if one assumes a Poissonian probability function. The distributions obtained as an exact solution within the SCPF approximation are clearly not symmetric, but exhibit long tails towards zero. Therefore, although the most probable values of the two types of distributions are similar, noise has a destructive impact on the system, resulting in a larger probability of having a smaller number of proteins in the cell than expected based on a Poissonian distribution, whose higher moments are equal to the mean. Therefore a larger production rate is needed for one of the states to be favoured as a result of noise than predicted from a symmetric probability distribution. The most probable number of proteins in the on state, if the unbinding from the DNA is slow, is zero, unlike predicted by Poissonian distributions. The influence of noise on protein number fluctuations brings the protein number means down, as can also be seen from Fig. 5.3 C. Overall the spread of the probability distributions is large, and their characteristics for small values of the control parameters are different from those predicted by Poissonian distributions, let alone by deterministic kinetic equations, therefore the effects of stochasticity may not be neglected.

5.D.6 The nonzero basal effective production rate case.

The above analysis concerns a switch with a zero basal production rate, so proteins were not produced in the off state. In a number of biological systems (Ptashne and Gann, 2002) a non-zero basal production rate exists and we now turn to consider the effect of this on a symmetric switch. Figure 5.6 B shows the dependence of the bifurcation curves for different values of the effective basal production rate $g_2/(2k)$. Values smaller than one, when the death rate is larger than the production rate, show that for the symmetric switch assuming the effective production rate to be zero in the off state is a reasonable approximation. If the on state has a positive input to the number of reservoir proteins present due to

$g_2/k > 1$, the probability of the active gene to be on, even for very large on state effective production levels X^{ad} is smaller than one. Hence the off state contributes considerably to the steady state number of proteins. The solution which corresponds to the more active of the two states may effectively be an off state, since it has $C_1(i) < 0.5$, although the effective production rate in the on state in the bifurcated region of parameter space is much larger than in the off state (for example the $g_2/(2k) = 20$ line in Fig. 5.6 B). As the effective basal production rate increases, a larger production rate in the on state than for small $g_2/(2k) > 1$ is required to reach the critical number of proteins for the bifurcation to take place, which is given by $< n(i) >= 2X^{ad}C_1(i) - g_2/k(2C_1(i) - 1)$. For this reason even for the deterministic approximation at the bifurcation point, the two genes must be more probable to be off, as can also be seen for the exact SCPF solutions from the probability distributions (Fig. 5.7 B, C, E, F). Figure 5.6 A shows the dependence of the bifurcation curves on the adiabaticity parameter, which tend to the deterministic case for large ω. A closer analysis of the $g_2/k > 1$ case, since the $g_2/k < 1$ is analogous to the zero basal production rate case which was already discussed, show that mean properties of the system are in even better agreement with the deterministic solution than the $g_2 = 0$ case (Fig. 5.7 A and D). The genes have a non-zero probability of being in the off state, with the probability distribution of the off gene having a long tail towards higher protein numbers Fig. 5.7 E and F. In the off state the effective production rate $g_2/(2k)$ is small and the noise input is small, relative to the large protein numbers present in the system. The small effect of stochasticity results in the observed similar mean characteristics. Yet the form of the probability distributions for the genes to be on before the transition is especially broad, with a far smaller probability than those of the off state (Fig. 5.7 B, C, E, F). These clearly show that the two genes are more probable to be in the off state before the bifurcation point. Therefore although the average observables are similar for the deterministic and SCPF stochastic solutions, the predicted distributions are unusual.

5.D.7 Summary

The symmetric switch is based on a competition between the accessibility of the repressor site and the number of repressor proteins present in the cell. The bifurcation is solely a result of the nonlinearity of the system and introducing noise simply affects the region in parameter space where given states occur. The protein number fluctuations have a destructive role in determining the stability of the bifurcated solution, however fast DNA unbinding rates can compensate for the destabilizing effect of protein number fluctuations. In this region the stochastic solution predicts similar means to the deterministic case, but the form of the probability distributions which depends on a large number of higher moments is non-trivial. It is a result of the interplay of the DNA binding and protein degradation kinetics.

5.E The Asymmetric Toggle Switch

Most switches found in nature are not symmetric. For asymmetric switches, when proteins bind as dimers, the two genes interact, resulting in probabilities to be on, different from those imposed purely by the equilibrium between binding and unbinding. The steady state solution is a compromise between the tendency that repressors are unbound from the initially off gene (X_1^{eq} for the forward transistion, X_2^{eq} for the backward in the following discussion) and the effective production rate of the initially on gene (X_2^{ad} - forward, X_1^{ad} backward transition) (at least for the deterministic case). This results in the characteristic S-curve bifurcation diagram, as presented in, for example Fig. 5.12, with possible forward and backward transtions, and what follows hysteresis. We refer to the transition which occurs with increasing X_1^{ad} as the forward transition and that with decreasing X_1^{ad} as the backwards transition. Since X_i^{ad} is a well defined function of the probabilities that the genes are on, the simplicity of the deterministic equations allows for a completely analytic discussion of the asymmetric switch. The more complicated

form of the exact SCPF equations makes this approach impossible. However the deterministic rate solution offers valuable insight into the basic mechanism behind the transition.

5.E.1 The general mechanism

By combining the steady state equations of motion for the probabilities of the two genes to be on and noting that with a zero basal production rate $< n(i) >= 2X_i^{ad}C_1(i)$, one can derive the following form of the deterministic bifurcation curves:

$$X_1^{ad}(C_1(2)) = \frac{X_2^{eq\frac{1}{2}}}{2}(1 + \frac{(2X_2^{ad}C_1(2))^2}{X_1^{eq}})(\frac{1}{C_1(2)} - 1)^{\frac{1}{2}} \tag{5.1}$$

as a function of $C_1(2)$ and:

$$X_1^{ad}(C_1(1)) = \frac{X_2^{eq\frac{1}{2}}}{2C_1(1)}(\frac{2X_2^{ad}}{((\frac{1}{C_1(1)} - 1)X_1^{eq})^{\frac{1}{2}}} - 1)^{\frac{1}{2}} \tag{5.2}$$

as a function of $C_1(1)$. The transistion points are determined as the extrema of these functions, which are functions solely of the scaled parameter X_2^{ad2}/X_1^{eq} and are plotted on the bifurcation graphs. It is worth noticing that the bifurcation points $C_1(i)$ do not depend on the value of X_2^{eq}, the parameter describing the gene binding kinetics of the gene that is on initially. This is not true for the exact SCPF solution, which cannot be solved analitically, but the bifurcation curve has the more complex form:

$$X_1^{ad}(C_1(2)) = \frac{1}{2}((((\frac{1}{C_1(1)} - 1)X_2^{eq})^{\frac{1}{2}}\frac{\omega_1 + C_1(1)}{1 + \omega_1} + 1)^{\frac{1}{2}} - \frac{\omega_1 + C_1(1)}{1 + \omega_1})\frac{1}{2C_1(1)} \tag{5.3}$$

where $C_1(1)$ is a function of $\omega_2, X_1^{eq}, C_1(2)$ and X^{ad_2}. The bifurcation point is therefore determined by the protein (X_i^{ad}) and DNA (X_i^{eq}) characteristics and their mutual interactions (ω_i) of the two genes. The deterministic approximation therefore greatly simplifies the mathematical mechanism of the transition. This may lead to large errors when studying more complicated biologically relevant systems, where one considers asymmetric switches with non-zero basal production rates and

proteins are produced in bursts. The case of the non-zero basal production rate within the deterministic approximation also cannot be solved analytically. The general picture behind the transition is seen from the deterministic approach. The larger the tendency for proteins to be unbound from the DNA, the larger the effective production rate X_1^{ad} must be for the transition from one gene to be active to the other to be active to take place, since repressor proteins are less likely to bind to the on gene (i) at large X_i^{eq} than at small X_i^{eq}. However, if one considers a noisy system, it is effectively harder for proteins to stay bound to the initially off gene due to the destabilizing effect of DNA binding noise (Fig. 5.8). For the stochastic system, apart from very low values of the adiabaticity parameter ($\omega < 0.1$) (Fig. 5.11), there is a threshold number of reservoir proteins which will cause a rapid transition. If we start with a small effective production rate for one type of proteins and increase this rate, keeping the production rate of the other gene fixed at an initially higher value, the proteins produced by the gene with the initially smaller production rate, repress it gradually and ineffectively, until they reduce the probability of the gene to be on to one half, for the exact SCPF solution. The number of proteins present in the on state decreases much more rapidly with the change of X_1^{ad}, whether it be increase for the forward transition or decrease for the backwards in the examples presented, than the number of proteins in the off state grows (Fig. 5.10). Hence the probability to be on of the initially active gene shows a larger sensitivity to the change of X_1^{ad} than the off state probability. This leads to a rapid transition of the previously active gene to an inactive state (Fig. 5.9). Such behavior is described by Ptashne [109, 115]in the λ phage switch, who points out its role as a "buffer against ordinary fluctuations in repressor concentration". The observed system switches when the "repression probability" drops to 50%, as in the solutions of this model. Our analysis seconds Ptashne's hypothesis, since the deterministic system lacks this behavior, the transition is rapid and for certain values of parameters takes place when the probability of the initially on gene drops to 80% (Fig. 5.8). The buffering capabilities of the stochastic system are clearly

seen in the long tails towards $n = 0$ of the probability distibutions of the gene that is switching from the on to the off state (Fig. 5.9 A and B).

5.E.2 The effect of noise on the bifurcation mechanism

The mean number of proteins at the transition point differs for the deterministic and exact SCPF solution (Fig. 5.10). More repressors are needed to induce the transition in the deterministic approximation than in the stochastic system, since due to the form of the interaction function for the exact case, $F(i) =< n(i) >^2 (\omega + 1)/(\omega + C_1(i)) + < n(i) >< n(i) >^2$. A smaller number of proteins is therefore needed for the inactive gene to become competitive with the active gene. The mechanism of the transition is different from the symmetric gene case, where a critical number of proteins needs to be reached. The asymmetric switch is based on the competition between the probability that proteins of one kind will repress the opposing genes and the analogous probability for the other kind of proteins. The repression capability is governed by X_{3-i}^{ad2}/X_i^{eq}, which might be looked upon as the product of the probability of having a certain number of repressor proteins $(3 - i)$ in the cell and the tendency for them to be bound to the opposing gene (i). In fact, the transition point in the deterministic case is purely a function of such ratios, $X_{3-i}^{ad2}/X_i^{eq} = f(X_i^{ad2}/X_{3-i}^{eq})$. In both the stochastic and deterministic cases, the transition points are set by the interaction function which regulates the on and off state probabilities of a given gene $F(3 - i)/X_i^{eq} = C_2(i)/C_1(i)$. Inclusion of noise in the system effectively increases the nonlinearity of the system, which results in the already discussed buffering capabilities of the system. Stochasticity alters the very simple competitive mechanism seen in the deterministic kinetics to allow for more levels of control of the stability of the state of the system against random fluctuations.

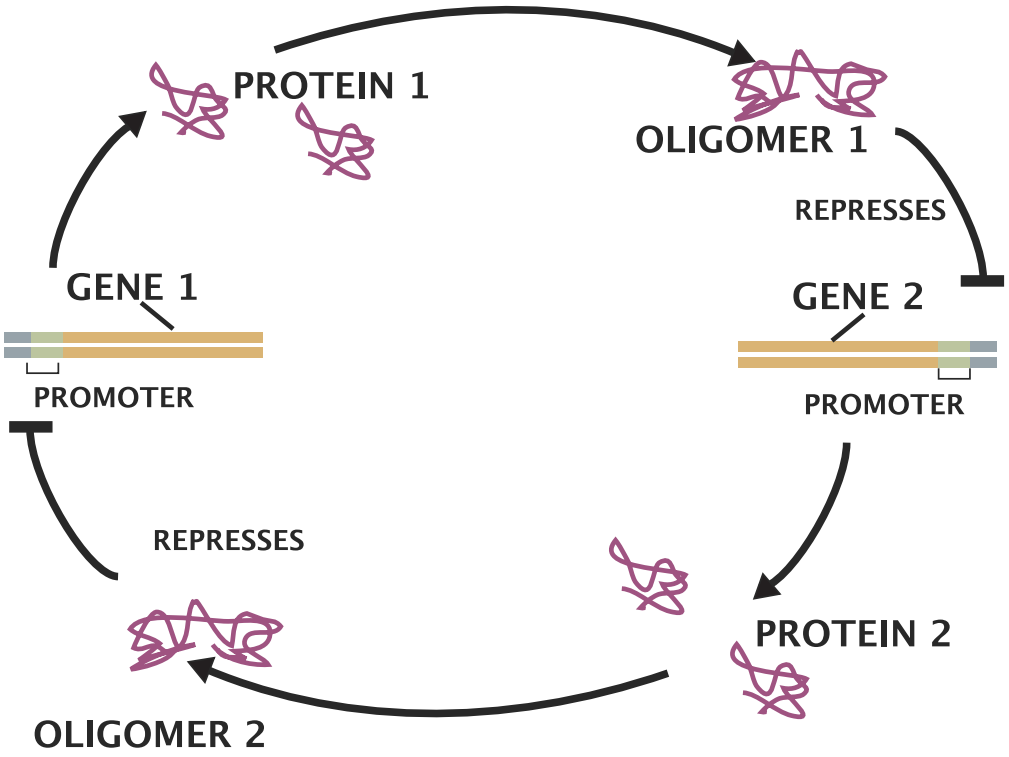

Figure 5.1: A schematic representation of the toggle switch. Gene 1 produces proteins of type 1 which repress gene 2 and gene 2 produces proteins of type 2 which repress gene 1.

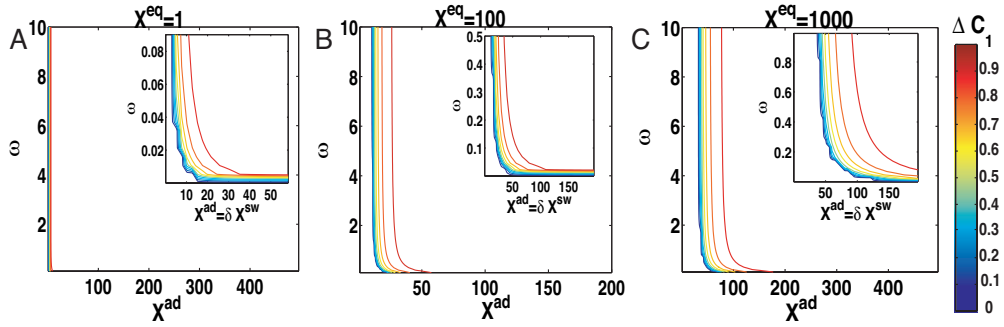

Figure 5.2: Phase diagram obtained as an exact solution within the SCPF approximation for the single symmetric switch when repressors bind as dimers with $X^{eq} = 1$ (A), 100 (B), 1000 (C). Contour lines mark values of ΔC.

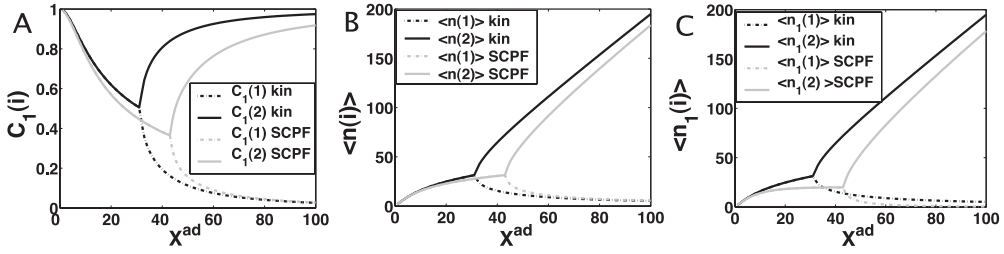

Figure 5.3: Probability that genes are in the active state (A), the mean number of proteins of each type present in the cell $< n(i) >$ (B) and the mean number of proteins of each type present in the cell if gene i is in the on state $< n_1(i) >$ (C) as a function of $X^{ad} = \delta X^{sw}$ for a symmetric switch. Exact solutions of the SCPF approximation equations compared with deterministic kinetic rate equations solutions, for a single symmetric switch, $X^{eq} = 1000$, $\omega = 0.5$.

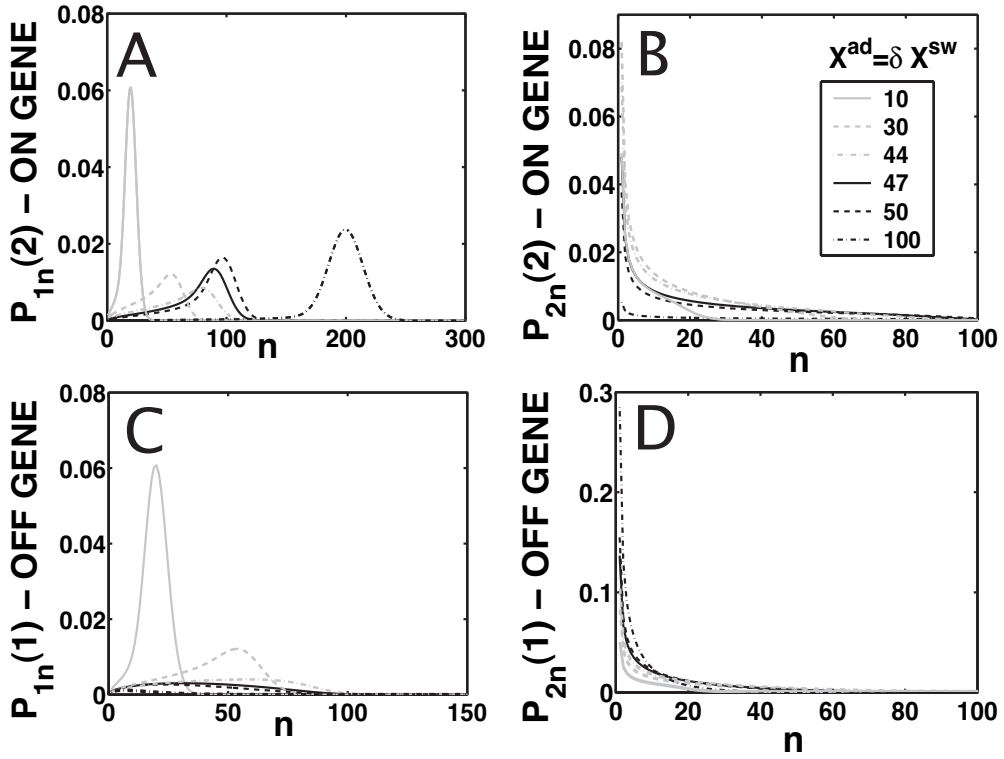

Figure 5.4: Evolution of probability distributions for the probability of the gene that will be active (on) after the bifurcation to be on (A) and off (B) and the gene that will be inactive (off) to be on (C) and off (D) as a function of the order parameter X^{ad} for a symmetric switch. The bifurcation occurs at $X^{ad} = 44$, $X^{eq} = 1000$, $\omega = 0.5$.

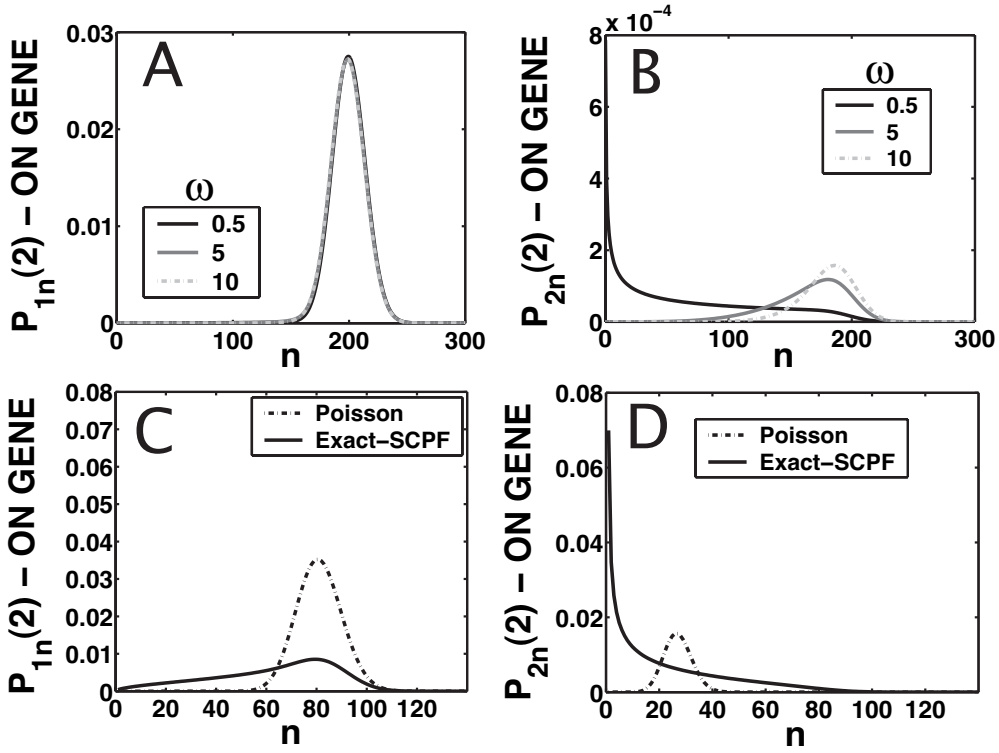

Figure 5.5: Probability distributions for the gene to be in the on state (A) and off state (B) for a gene in the active state for different values of the adiabaticity parameter $\omega = 0.5, 10, 100$. $X^{eq} = 100$, $X^{ad} = \delta X^{sw} = 100$. Comparison of probability distributions obtained by exactly solving the steady state equations in the SCPF approximations with analogous Poissonian distributions (C and D). Symmetric switch, $X^{ad} = 44$, $X^{eq} = 1000$, $\omega = 0.5$.

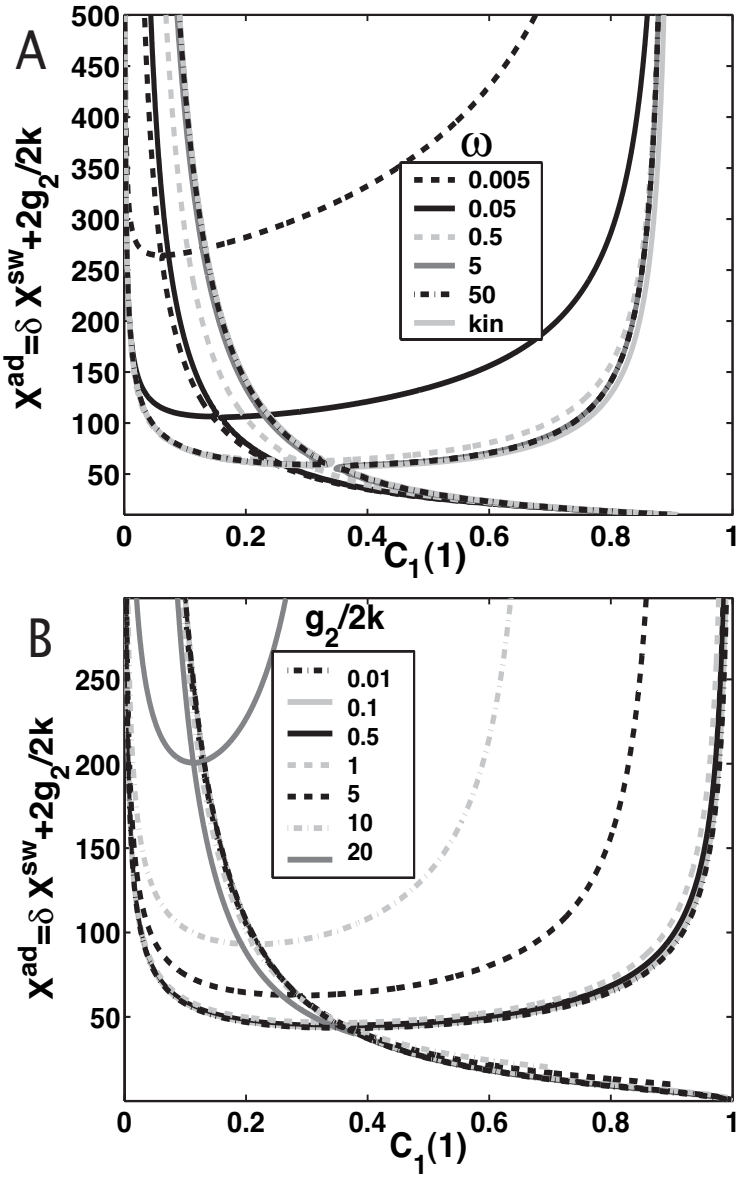

Figure 5.6: Nullclines for a symmetric switch when proteins bind as dimers when the effective base production rate $g_2/(2k) \neq 0$. Dependence on the adiabaticity parameter $\omega = 0.005, 0.05, 0.5, 5, 50$ compared to the deterministic equations solution (A), $g_2/(2k) = 5$. Dependence on $g_2/(2k) = 0.01, 0.1, 0.5, 1.0, 5, 10, 20$, $\omega = 0.5$. $X^{eq} = 1000$ (B).

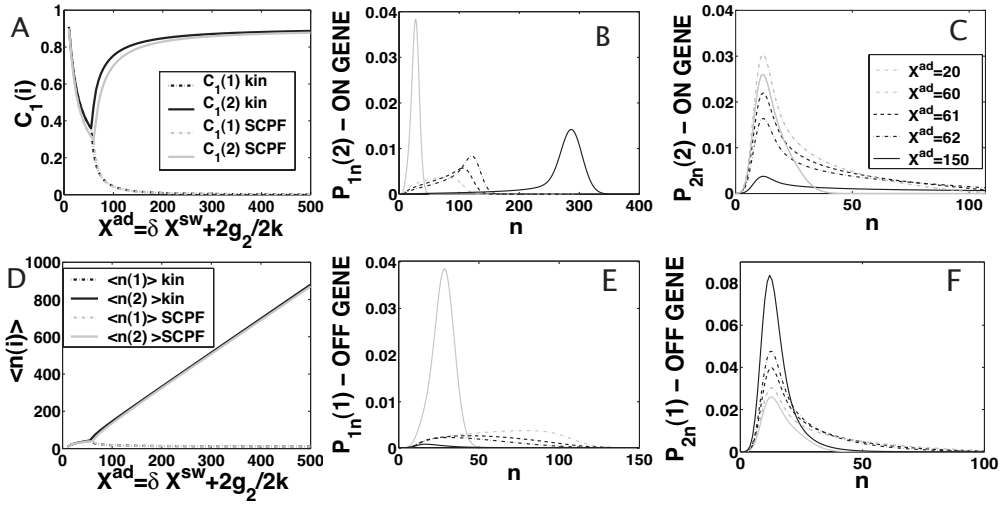

Figure 5.7: Probability of genes to be on (A) and mean number of proteins of a given type present in the cell (D) for a symmetric switch with an effective base production rate. Evolution of probability distributions for the probability of the gene that will be active after the bifurcation to be on (B) and off (C) and the gene that will be inactive to be on (E) and off (F) as a function of the order parameter X^{ad} for the same system. The bifurcation occurs at $X^{ad} = 61$, $g_2/(2k) = 5$, $\omega = 0.5$, $X^{eq} = 1000$.

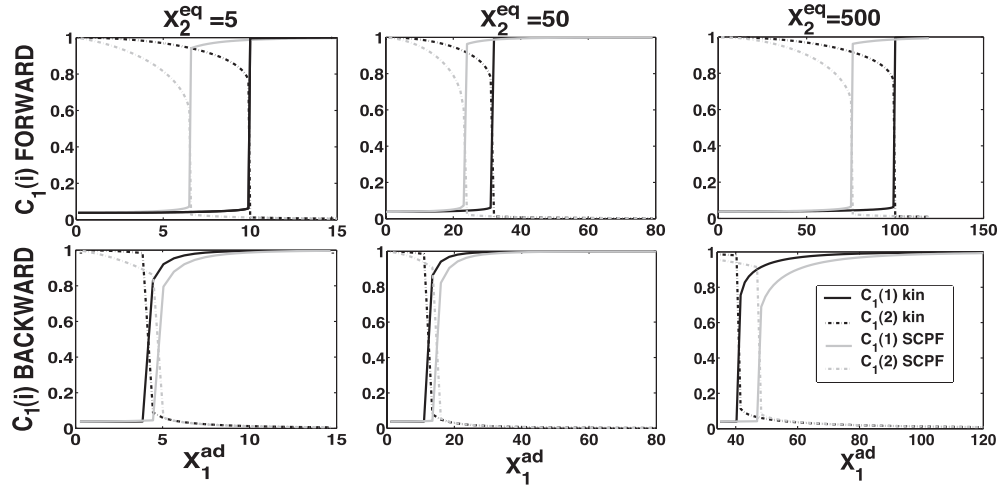

Figure 5.8: Dependendce of the probability of genes to be on in an asymmetric switch as a function of increasing parameters of one gene $X_1^{ad} = \delta X_1^{sw}$ in the forward (top) and backward (bottom) transition for different values of X_2^{eq}: 5, 50, 500. All other parameters fixed at $X_1^{eq} = 1000$, $\omega_1 = \omega_2 = 0.5$, $X_2^{ad} = \delta X_2^{sw} = 80$. Comparison of solutions of deterministic and exact SCPF equations.

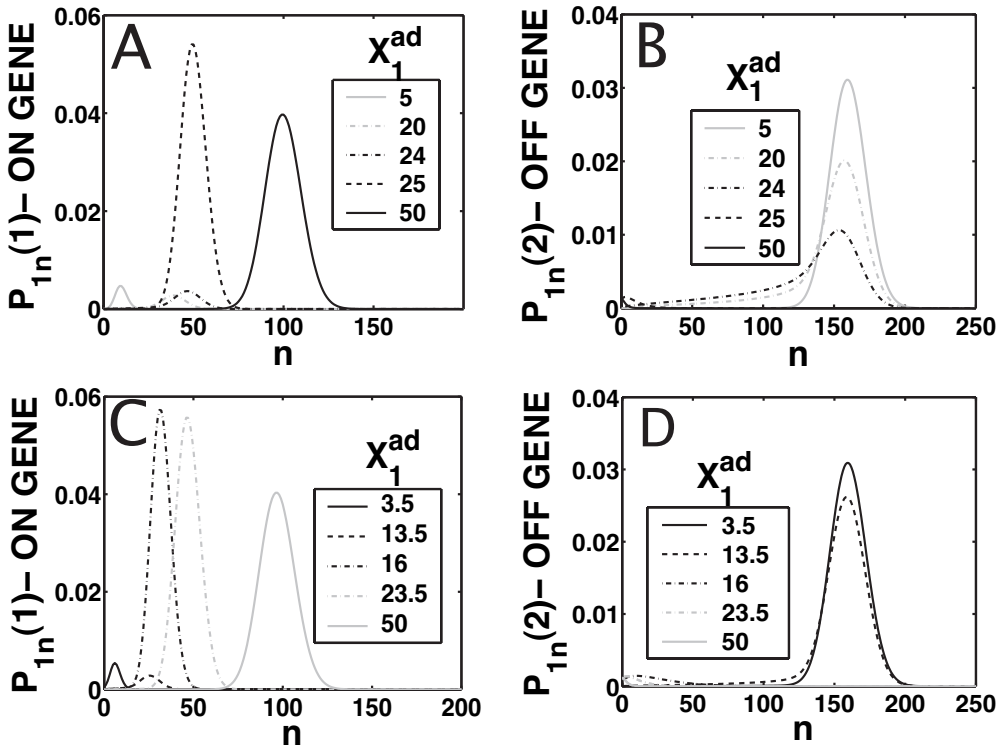

Figure 5.9: Evolution of the probability distributions for the two genes to be active for the forward transition (A) and (B) and the backward (C) and (D) as a function of $X_1^{ad} = \delta X_1^{sw}$ or an asymmetric switch. With $X_2^{eq} = 50$ with $X_1^{eq} = 1000$, $\omega_1 = \omega_2 = 0.5$; $X_2^{ad} = \delta X_2^{sw} = 80$.

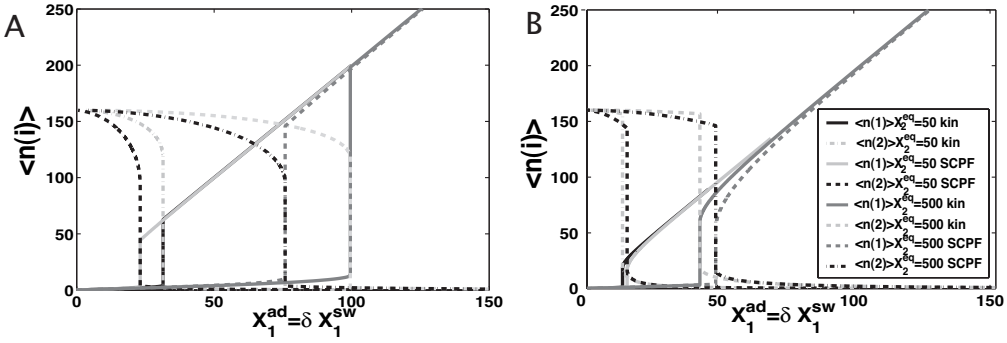

Figure 5.10: Mean number of proteins of each type present in the cell, according to exact solutions of the SCPF approximation and deterministic kinetic rate equations for an asymmetric switch. With $X_1^{eq} = 1000, \omega_1 = \omega_2 = 0.5, X_2^{ad} = \delta X_2^{sw} = 80$ and $X_2^{eq} = 50, 500$ during the forward (A) and backward (B) transition in an asymmetric switch.

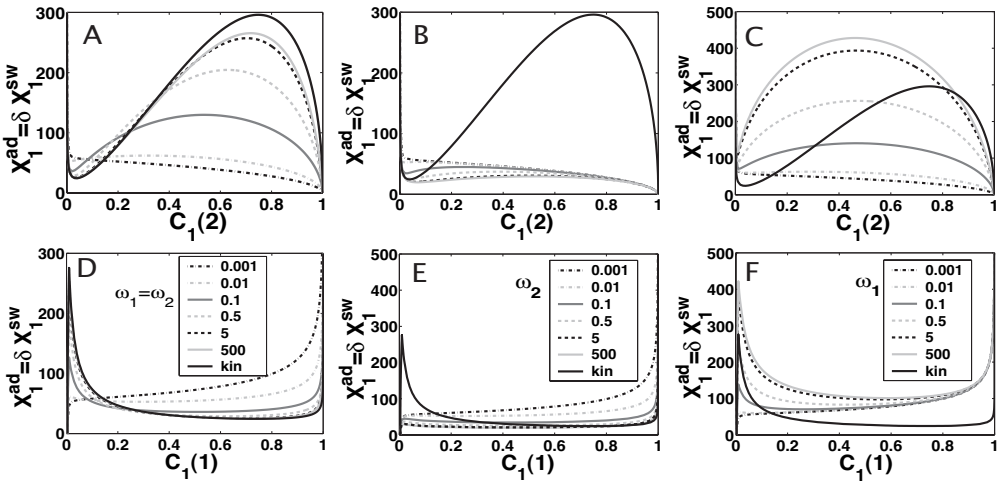

Figure 5.11: Bifurcation diagrams for an asymmetric switch, presenting $X_1^{ad} = \delta X_1^{sw}$ as a function of $C_1(2)$ $(A–C)$, and $C_1(1)$ $(D–F)$ for different values of the adiabaticity parameter: ω_1=ω_2 (A, D), ω_2, with $\omega_1 = 0.001 = const$ (B, D), ω_1, with $\omega_2 = 0.001 = const$ (C, F). $X_1^{eq} = 100$,$X_2^{eq} = 50$,$X_2^{ad} = \delta X_2^{sw} = 80$.

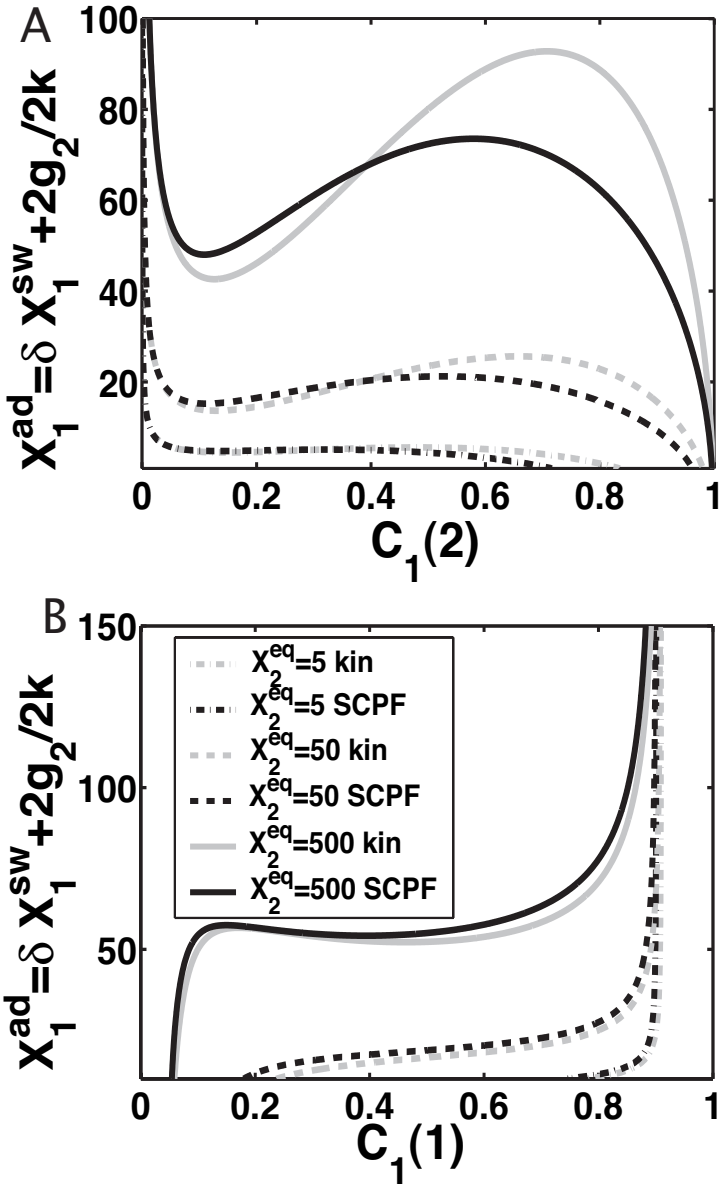

Figure 5.12: Bifurcation diagrams as a function of $X_1^{ad} = \delta X_1^{sw} + 2g_2/(2k)$ for $C_1(1)$, with $g_2(1)/(2k) = g_2(2)/(2k) = 5$ (A) and $C_1(2)$ $g_2(1)/(2k) = g_2(1)/(2k) = 0.5$ (B) for $X_2^{eq} = 5, 50, 500$. Comparison of exact solutions of the SCPF and deterministic kinetic equations for an asymmetric switch. $\omega_1 = \omega_2 = 0.5$, $X_1^{eq} = 1000$, $X_2^{ad} = \delta X_2^{sw} = 80$.

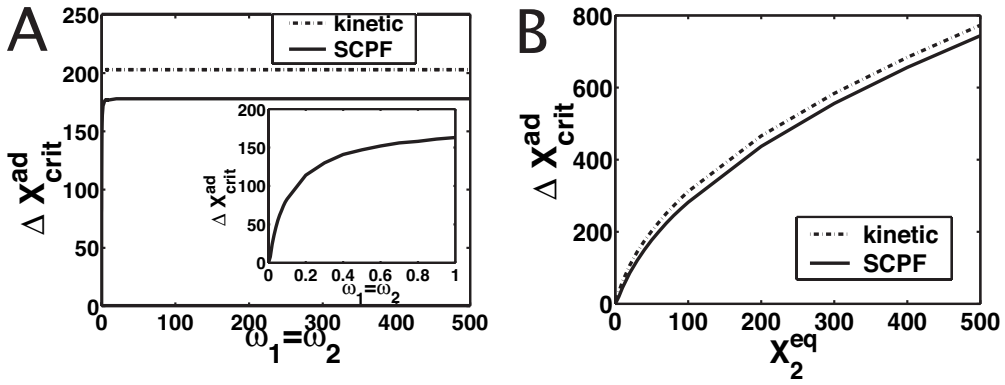

Figure 5.13: Region of $C_1(1)$ hysteresis for an asymmetric switch for the SCPF and deterministic approximations as a function of $\omega_1 = \omega_2$. With $X_2^{eq} = 50$ (A) and X_2^{eq} with $\omega_1 = \omega_2 = 100$ (B). $X_1^{eq} = 100$, $X_2^{ad} = \delta X_2^{sw} = 80$, $g_2/(2k) = 0.5$.

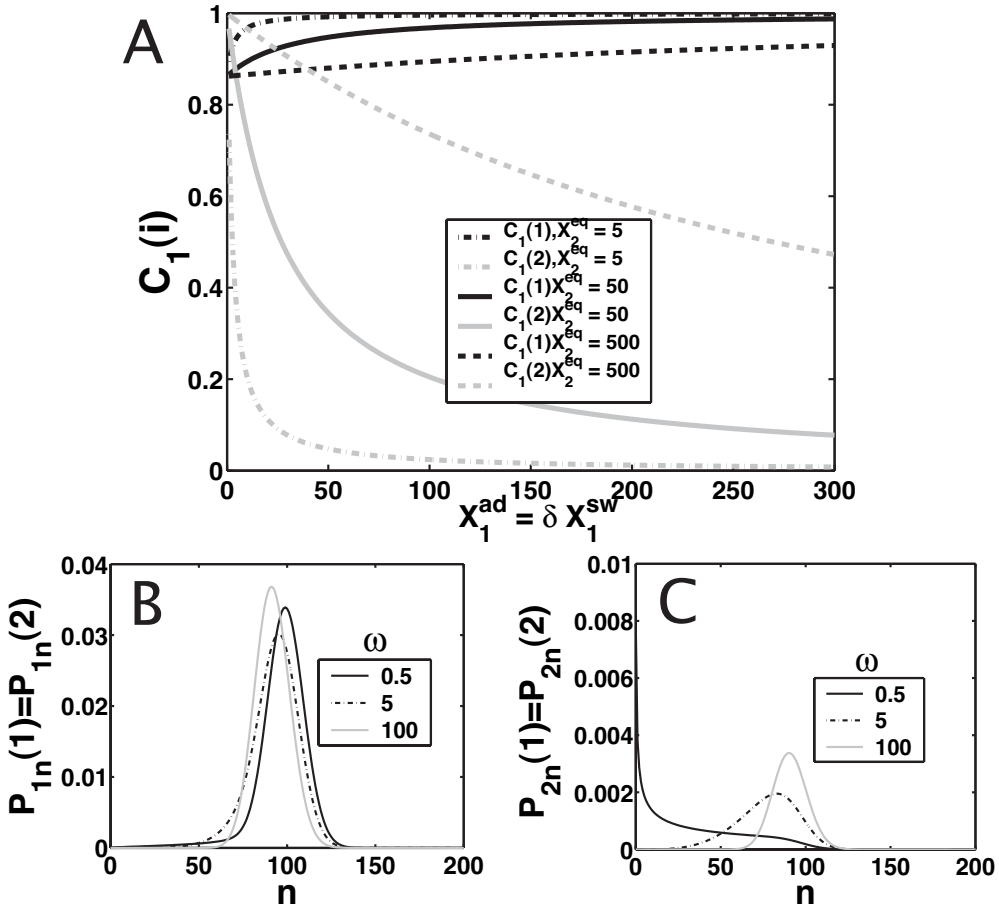

Figure 5.14: Probability distributions for an asymmetric switch. (A) Probability of genes in an asymmetric switch to be active when proteins bind as monomers, for different values of X_2^{eq}. $X_2^{ad} = \delta X_2^{sw} = 80$. Probability distributions for the gene to be in the on state (B) and off state (C) for a gene in the active state for different values of the adiabaticity parameter $\omega = 0.5, 5, 100$, when proteins bind as monomers to a symmetric switch. $X^{ad} = \delta X^{sw} = 50$. $X^{eq} = 1000$.

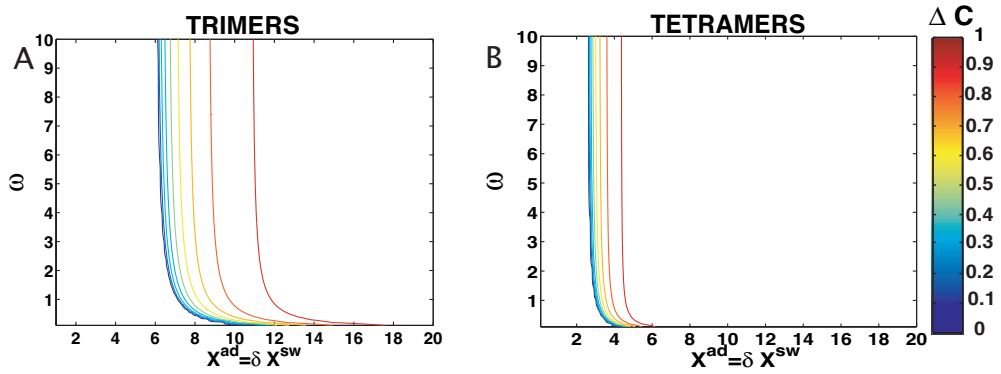

Figure 5.15: Phase diagram for the SCPF approximation for a single symmetric switch to which proteins bind as trimers (A) and tetramers (B) with $X^{eq} = 1000$. Contour lines mark values of ΔC.

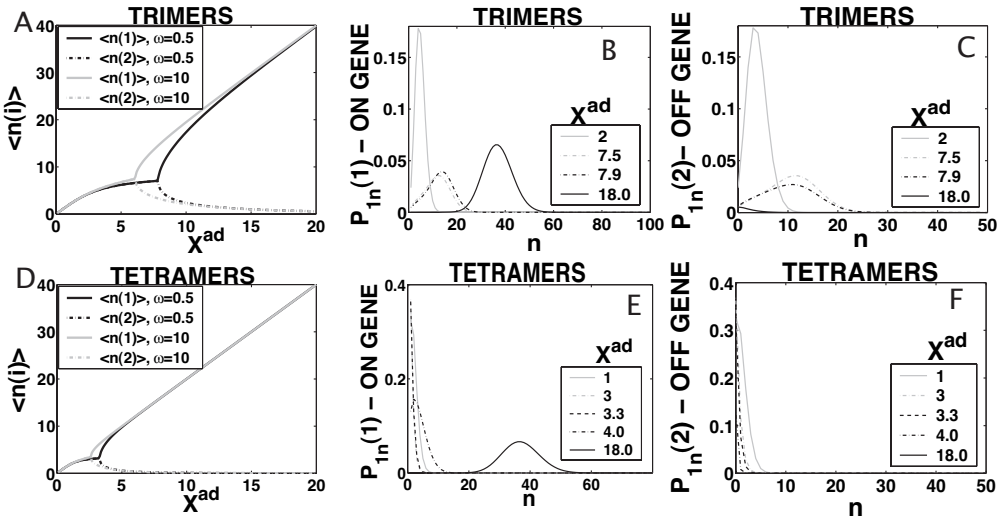

Figure 5.16: Mean number of proteins in the cell, for each type when proteins bind as trimers (A) and tetramers (D), $\omega = 0.5, 10$, symmetric switch. The evolution of the probability distribution for the probability of the gene that will be active and inactive after the bifurcation to be on as function of X^{ad} for a switch when proteins bind as trimers (B and C) and tetramers (E and F). $X^{eq} = 1000$,$\omega = 0.5$.

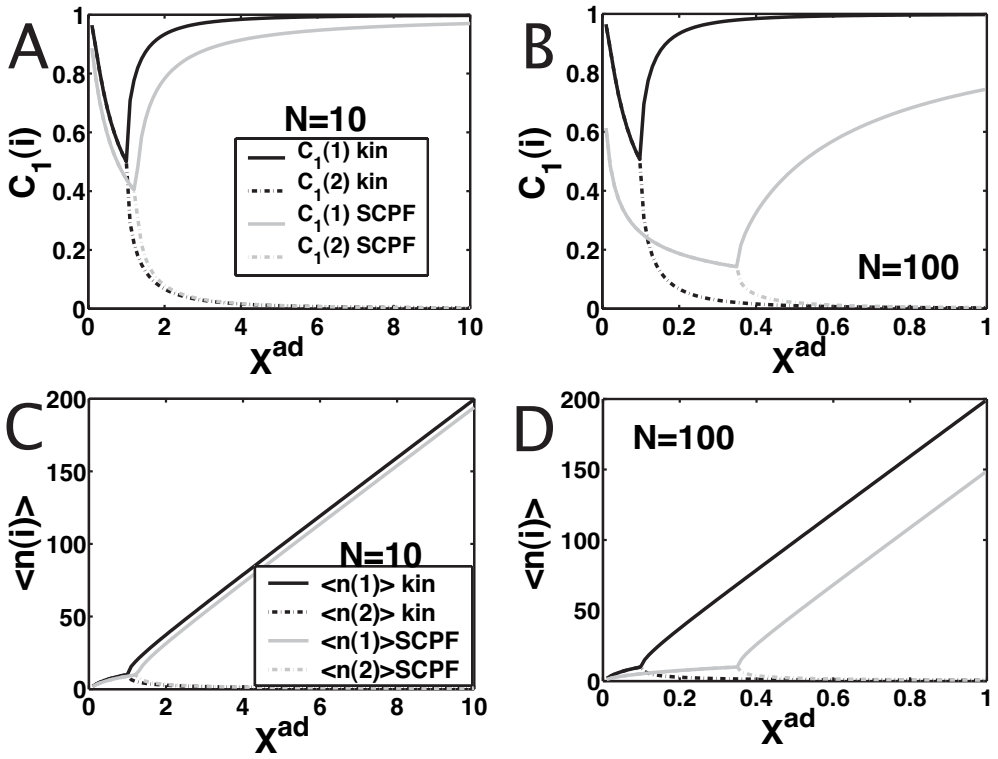

Figure 5.17: Probability that gene i is on when proteins are produced in bursts of $N = 10$ (A) and $N = 100$ (B). Mean number of proteins of each type present in the cell when proteins are produced in bursts of $N = 10$ (C) and $N = 100$ (D). Symmetric switch proteins bind as dimers, $X^{eq} = 100$, $\omega = 100$. Comparison of deterministic and stochastic solutions.

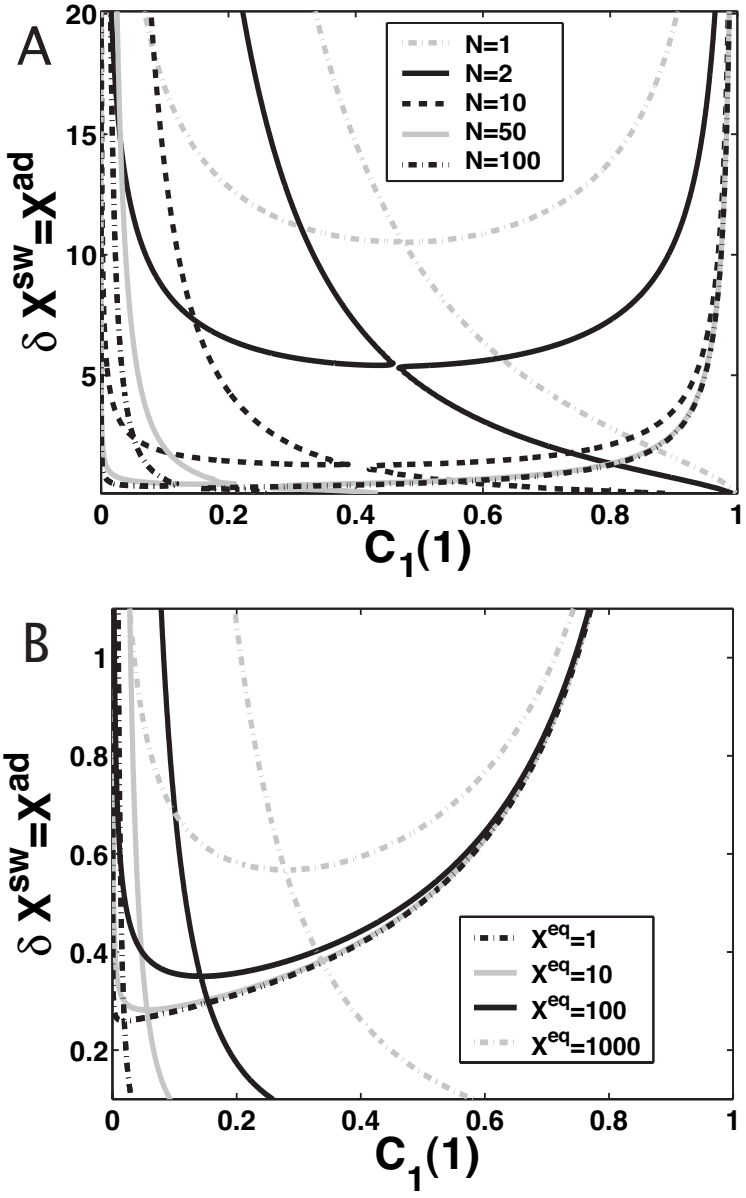

Figure 5.18: Bifurcation curves as a function of $X^{ad} = \delta X^{sw}$, $\omega = 100$ for different burst size values $N = 1, 2, 5, 10, 50, 100$, with $X^{eq} = 100$ (A) and for proteins produced in bursts of $N = 100$ (B) for different values of $X^{eq} = 1, 10, 100, 1000$.

Further comparison of solutions of the deterministic and stochastic equations leads to the same conclusions as for a symmetric switch. As the tendency for proteins to be unbound from the DNA grows, the difference in the critical number of reservoir proteins necessary for the transition to take place increases for both approximations. The critical number of proteins produced by a given gene necessary for the transition to take place for both genes is, in most cases (see ω dependence discussion), smaller for the exact solutions of the SCPF equations and the difference between the stochastic and deterministic result grows with both X_i^{eq} and decreases with ω_i (Fig. 5.10). It has a value of 15 for $X_2^{eq} = 500$, $\omega_1 = \omega_2 = 0.5$ and 2 for $X_2^{eq} = 500$, $\omega_1 = \omega_2 = 10$.

Consider the forward transition. The initially inactive gene is buffered by a cloud of repressor proteins. As one increases the effective production rate of the proteins produced by the inactive gene (X_1^{ad}), the number of proteins which are able to repress gene 2 grows slowly and linearly $< n(i) >= 2X_1^{ad}C_1(1)$, where $C_1(1) \sim const$ and forms a buffering proteomic cloud around it. In the results presented in the figures of this paper the tendency that proteins are unbound from gene 2, (X_2^{eq}), is smaller than X_1^{eq}, so gene 1 is able to produce enough repressors to form a stable buffering cloud around gene 2 and turn it into the inactive state at quite modest values of X_1^{ad}. If $X_1^{eq} < X_2^{eq}$, gene 1 produces proteins less effectively, as the probability of it being repressed is larger than in the previous case, and larger values of X_1^{ad} are needed to produce enough repressors to achieve a high effective probability of binding, X_1^{ad2}/X_2^{eq}. An example of how $X_1^{ad,crit}$ grows as $X_1^{eq} \to X_2^{eq}$, is seen by comparing the $X_1^{ad} \sim 33$ for $X_1^{eq} = 1000$, $X_2^{eq} = 50$ in Fig. 5.8 and $X_1^{ad} \sim 300$ for $X_1^{eq} = 100$, $X_2^{eq} = 50$ (Fig. 5.11).

5.E.3 Adiabaticity parameter dependence

The interaction of the buffering proteomic cloud with the DNA can be altered when the ratio of the DNA unbinding rate compared to the protein degradation rate is changed. For small ω_i values the unbinding rate of repressors to the DNA

is slower than the destruction of the produced proteins. Apart from very small ω values, as long as there is a critical number of repressor proteins in the buffering cloud, the off gene is repressed and it responds by turning on, only once the initially on gene is nearly totally repressed. Large adiabaticity parameters result in the efficient formation of the buffering proteomic cloud. For the initially off gene, a small DNA unbinding rate of the off gene, decreases the effectiveness of the buffering proteomic cloud around it, as the protein number state can reach a steady state before the DNA state does. The hindered DNA reaction to the protein number state effectively increases the tendency of repressor proteins to be unbound from the DNA, for a given X_1^{ad}. This in turn decreases the probability of the initially on gene to be on, leading to rapid, switching behavior as an be seen for gene 2 in the forward, or gene 1 in the backward transition for $\omega > 0.1$ in Fig. 5.11 *A*. The initially on gene reacts to the interaction function of the initially off gene, for which $F(i) \rightarrow < n(i) >^2 / C_1(i) + < n(i) >$ in the small ω limit. Therefore the interaction function is effectively increased for $C_1(i) \approx 0$, leading to the enhanced buffering. The reaction of the initially off gene is unaltered, as for $C_1(i) \approx 1$ $F(i) =< n(i) >^2 + < n(i) > \sim const$, if $C_1(i)$ remains close to 1. However if ω is very small (black dash-dot curve in Fig. 5.11 *A*), the buffering proteomic cloud is not given a chance to form due to a very high degradation rate of proteins and gene 2 is simply repressed in a gradual transition. If ω_1 is extremely small and ω_2 large, the buffering proteomic cloud around gene 1 cannot form and the probability of it to be off in the forward transition decreases gradually. A buffering proteomic cloud exists around gene 2, hence the backward transition is reminiscent of the deterministic result (Fig. 5.11 *B*). The most interesting case is shown in Fig. 5.11 *C*, where a large ω_1 acts as a buffer against fluctuations in the number of proteins, which repress gene 1. For large production rates of repressors the probability of gene 2 to be on for the forward transition decreases faster than in the deterministic solution, however the buffering cloud repressing gene 1 allows gene 2 to remain in the on state. A buffering proteomic cloud does not form around

gene 2 and it remains on until the number of proteins produced by gene 1 grows considerably, as the effective production rate, X_1^{ad}, is increased. The effective production rate of gene 1 must be very large to sustain a sufficient steady state number of proteins to repress gene 2 to the point that $C_1(1) < 0.5$, which leads to switching. For the backward transition the lack of a buffering proteomic cloud around gene 2 results in destabilizing gene 1 for larger X_1^{ad} effective production rates than for large ω_2 values. These examples show how certain combinations of values of adiabaticity parameters can lead to a system with a larger switching region than the deterministic model predicts. This property may be useful when engineering artificial switches. If one has a constraint on the production rates of the genes, one can use repressors with different binding affinities to achieve switching in the desired region of parameter space.

In this simple system slow unbinding from the DNA can compensate for the destabilizing of the DNA state by protein number fluctuations. As the probability of the initially active gene to be on gradually decreases, the initially repressed gene becomes active only once the probability of the other gene to be on has fallen bellow a certain values α. The susceptibility of the system to protein number fluctuations may be estimated by the value of α. For small ω, which is still able to sustain a buffering proteomic cloud, this values tends to 0.5. The incapability of the system to form a buffering proteomic cloud is much stronger if both adiabaticity parameters are small, since the reaction of both genes to the change in the number of proteins is hindered (Fig. 5.11 *A*). DNA state fluctuations contribute to effectively faster protein number fluctuations, therefore the exact solution exhibits the very small ω characteristics, where a buffering proteomic cloud cannot form, for a slightly wider range of the adiabaticity parameter than one would expect with a Poissonian distribution (results not shown). Combining these observations a switch works most effectively if the change of the DNA state compared to the protein number fluctuations of one gene is sufficiently smaller than that of the other gene, to allow for effective buffering.

5.E.4 The nonzero basal production rate

The asymmetric switch in which both genes have a nonzero basal effective production rate proves to be susceptible to noise. In Fig. 5.12, we show the dependence of $C_1(1)$, with $g_2(1)/(2k) = g_2(2)/(2k) = 5$ and $C_1(2)$, with $g_2(1)/(2k) = g_2(2)/(2k) = 0.5$ in the small ω_i limit. The stochastic solutions converge to the deterministic solutions for large ω. If gene 2 is initially in the on state, the majority of proteins are produced with the high fixed rate in the on state, as $g_1(2) >> g_2(2)$. The repression of gene 2 is in turn governed by the interaction function of gene 1. If X_1^{ad} is small the number of proteins produced in the on and off states by gene 1 are comparable. As the number of proteins produced by gene 1 grows faster the larger g_2 is, gene 2 gets repressed more effectively for smaller X_1^{ad} values. This results in a smaller number of repressors produced by gene 2 and the transition from gene 1 to be on to be off takes place for smaller X_1^{ad} - effective growth rate values, than for small g_2.

The deterministic solution is much more influenced by the production of proteins in the off state than the stochastic solution. In the exact SCPF solution slow DNA unbinding rates compared to protein degradation rates are another means of control of the stability of the DNA state against random protein number fluctuations. The state of the system is far less influenced by the exact protein numbers than in the deterministic solution. So until the probability of a gene to be on is larger than that to be off, the fraction of proteins produced with a smaller effective production rate in the off state is treated as a random fluctuation by the system. Once again the SCPF system demonstrates its susceptibility to protein number fluctuations.

The influence of the off state protein production on the total repressor yield may also be seen in the fast decrease of $C_1(2)$ and increase of $C_1(1)$ in the forward transition. If g_2 is considerably large its effect can also be seen in the stochastic solution, hence even when gene 1 is in the on state, it never reaches $C_1(1) = 1$, although gene 2 is totally repressed (Fig. 5.12 *A* and results not shown for gene

2). The magnitude of the probability of gene 1 to be on for very large effective production parameters strongly depends on the the tendencies of the proteins to be unbound from gene 1. As X_1^{eq} increases the asymptotic X_1^{ad} limit of $C_1(1)$ becomes smaller, as it is effectively harder for repressors to stay bound to the DNA. The gene is more likely to be in the off state, which however manages to sustain the necessary number of proteins produced by gene 1 to repress gene 2. As g_2 increases the region of bistability grows into areas of parameter space, in which the tendency of proteins to be unbound, X_2^{eq}, is larger than for small g_2. For small values of X_2^{eq} the number of repressors produced by gene 1 in the off state is sufficient to repress gene 2 and one observes a smooth and slow transition in terms of X_1^{ad}. If g_2 is considerably large the transition takes place for larger values of X_1^{ad} in the stochastic solution than in the deterministic solution, hence showing the large buffering region the interplay of DNA and protein number fluctuations provides. This also results in an effective similarity of the deterministic and stochastic solution. In regions of parameter space, in which the change of DNA state is rapid, the deterministic and stochastic solutions differ, apart from the large ω limit. Most experimentally observed proteins have very small basal production rates, which seconds our analysis, that it is functionally unfavourable for large basal production to occur. The dependence on other parameters is analogous to the case without a basal production rate.

5.E.5 The region of bistability

The backward transition, as already discussed, is analogous to the forward transition. In most cases, the regions of bistability (Fig. 5.11) in parameter space are reduced in size by noise. When engineering artificial switches, one may be interested in making sure the forward and backward transition takes place for considerably different production rates. We therefore consider how the region of bistability, defined as the difference in the critical effective production rate for the forward and backward transition, depends on the parameters of the model. For the

deterministic case the region of bistability depends on the tendencies that proteins are unbound from the DNA in a quadratic manner, as can easily be seen from the bifurcation equations 5.1, 5.2 and is demonstrated in Fig. 5.13. The SCPF solution shows the same behavior. For large values of the adiabaticity parameter the size of the region of bistability is independent of ω, as is the form of the bifurcation curve (Fig. 5.13). The approach to this plateau is very rapid and is given by the ratio of polynomials. However, the size of the region of bistability for the $\omega_1 = \omega_2$ never reaches that of the deterministic solution, as even in the large ω limit the greater nonlinearity of the interaction function $F(i)$ results in a more complex SCPF curve which does not reduce to deterministic solution, but $X_1^{ad}(C_1(2)) \rightarrow ((((C_1^{-1}(1)-1)X_2^{eq})^{\frac{1}{2}}+1)^{\frac{1}{2}}-1)/(4C_1(1)) \neq X_2^{eq\,\frac{1}{2}}(1+(2X_2^{ad}C_1(2))^2/X_1^{eq})(C_1^{-1}(2)-1)^{\frac{1}{2}}/2$. This effect is true for both curves, as the presented graphs show $C_1(1)$ hysteresis and the chosen equations $C_1(2)$. The same behavior is observed for the case with a zero and a nonzero basal production rate. The increase with X_2^{eq} is slightly slower in the $g_2 \neq 0$ case as the bifurcation curve is smaller by $|g_2/k(C_{1f}(i) - C_{1in}(i)) - ln(C_{2f}(i)/C_{1in}(i))/2|$.

5.E.6 Summary

After the transition, the number of proteins produced by the now on gene, follows a linear dependence on X^{ad}, similarly to the symmetric switch. The number of proteins in the cell is independent of the DNA dynamical characteristics, as those remain constant in that region of parameter space. The number of proteins of the off gene, rapidly falls before the transition takes place. Based on the bifurcation diagram of Fig. 5.12 the phase transition is discontinuous, for a certain region of the parameter space, where switching may occur. That region may be roughly estimated by the parameters of the genes which must be competitive, $(X_1^{ad}/X_2^{ad})^2 \approx X_2^{eq}/X_1^{eq}$. This has a major implication for biological systems, such as the λ phage, where many mechanisms are used to achieve balance between two genes. The first order phase transition, as opposed to the second order present in

the symmetric system, is a results of the breaking of symmetry and is clearly seen in the evolution of probability distributions in phase space (Fig. 5.9). The gene that is on after the transition rapidly increases its probability of being on, whereas the off gene decreases with a rapid drop in the number of proteins it produces.

5.F The Case when Proteins bind as Monomers

The equations presented above can easily be augmented to describe the binding of monomers or higher order oligomers by changing the form of the binding term to $h_i n_{3-i}^p$, where $p = 1$ for monomers. The equations remain solvable for any value of p.

5.F.1 Monomers do not make good repressors/activators

The behavior of the system is quite different if we consider the case when proteins bind as monomers. For a symmetric switch there is no region of the parameter space, in which one observes switching. The SCPF equations may be reduced to a single quadratic equation:

$$2\delta X^{sw} C_1(i)^2 + (X^{eq} + X^{ad} - \delta X^{sw}) C_1(i) - X^{eq} = 0 \tag{5.4}$$

which has at most only one positive solution. Therefore the probability of one gene to be in the active state is always equal to that of the other to be in the active state and no switching is observed. The equation (5.4) is independent of ω, the adiabaticity parameter, therefore it is solely a consequence of the lack of nonlinearity in the binding of proteins and cannot be influenced by very slow DNA unbinding rates. By writing down deterministic equations we can also show that when proteins bind as monomers switching does not occur. A similar equation to (5.4), also independent of ω, holds for asymmetric switches. It also has one positive solution, therefore the parameters of the model predetermine the solution and each gene has a probability to be on determined by its kinetic rates. Since the rates are different for the two genes, the gene with the larger production rate

will be in the active state, repressing the weaker gene (Fig. 5.14 A). In naturally occurring biological switches and those developed experimentally proteins bind as dimers, or higher order multimers (Ptashne, 1992). We see cooperativity contributes to improving the efficiency of a switch. A switch controlled by monomers is shown to react ineffectively to changes in the repressor concentration, just as in the case of the asymmetric switch in our model discussed above. Monomers do not have the ability to stabilize a broken symmetry state, therefore the solution is fragile to kinetic rates and inefficient. Effectively monomers do not make good repressors/activators. Ptashne and Gann [115] explain the cooperativity process between two monomers by claiming that one monomer bound to the DNA increases the "local concentration" of proteins around the binding site through weak protein-protein interaction, thus causing the second to bind cooperatively. Our model lacks spatial dependence, therefore shows this effect need not be thought of as due to changes in local concentration, but actually is required by the insufficient nonlinearity for monomers, which cannot produce bistability.

5.F.2 Bimodal probability distribution

Although the probabilities of the two genes to be on are equal for the whole region of parameter space and the mean number of both types of proteins in the cell is the same as in the deterministic case, the probability distributions are bimodal when the DNA unbinding rates are slower than the protein number fluctuations (Fig. 5.14 B and C). The mechanism of this small ω behavior has already been discussed on the example of the symmetric switch when proteins bind as dimers. This is analogous to the case when DNA fluctuations induce a probability distribution with two peaks for the single gene with an external inducer [46]. In fact the SCPF approximation has reduced this two gene system to an effective one gene system with an external inducer. A bimodal distribution in the small ω case is also observed for the asymmetric switch, when proteins bind as monomers.

5.G The Case when Proteins bind as Higher Order Oligomers

Switches in which effector proteins bind as higher order oligomers are omnipresent in nature and have been realized experimentally in artificial switches [26]. We considered the binding of trimers ($h_i(n_{3-i}) = h_i n_{3-i}^3$) and tetramers ($h_i(n_{3-i}) = h_i n_{3-i}^4$) in symmetric switches. The equations of motion have the same form as before, but the interaction function $F(i)$ accounts for the higher moments. For proteins binding as k^{th} order oligomers it has the form $F(i) = C_1(i) < n_1^k(i) > + C_2(i) < n_2^k(i) >$. As shown when discussing the dimer binding switch, the k^{th} order moments have a simple form in the creation operator representation.

5.G.1 The general mechanism

From Fig. 5.15 one notes that in order for the system to act as a bistable switch a considerably smaller number of reservoir proteins is needed than in the case of the dimer binding switch. As the multimericity number grows the area of bistability of the switch in parameter space grows. Since we assumed only one type of protein repressed a given gene, binding of higher order multimers is an effective model of cooperativity. Therefore we expect the system to have a larger region of bistability the higher the order of the binding multimer. The evolution of the system in parameter space when trimers bind is qualitatively similar to the dimer binding scenario (Fig. 5.16 *B* and *C*). Fast DNA unbinding rates stabilize the system and the bifurcation takes place for smaller effective production rates, for large ω than for small ω (Fig. 5.16 *A* and *D*). The critical number of proteins necessary for the bifurcation to take place is independent of the adiabaticity parameter and decreases with multimericity: $< n >_c = 32$ for dimers binding, $< n >_c = 8$ for trimers binding and $< n >_c = 4$ for tetramers binding. This along with the narrow probability distributions (Fig. 5.16 *E* and *F*), small ω dependence when tetramers bind (Fig. 5.15), shows that one binding event determines the result, hence DNA binding rates do not play a role. Once there are $< n >_c$ proteins of a given type

in the cell, a tetramer repressor will bind and stay bound. In the deterministic case the probability of the genes needs to fall to $(p-1)/p$, where p is the order of multimerization of the repressor, for the bifurcation to take place. That along with the need for the number of repressors to be comparable with the tendency for proteins to be unbound from the DNA sets the critical number of proteins necessary for the bifurcation. Hence the bifurcation occurs when both genes are more probable to be on than off, for both tetramers and trimers. Therefore for the tetramer system a large buffering proteomic cloud is not needed to stabilize the DNA binding state of the switch and the characteristics of the system are practically independent of the adiabaticity parameter.

5.G.2 Tetramer binding results in nearly deterministic characteristics

In naturally occurring systems the production of the critical number of proteins is slowed down by relatively high multimerization rates and spatial dependence arising from the need of a large number of particles to diffuse together. These elements, which we neglect in our simple model constitute what might be called the cost of multimerization. This analysis also explains why most repressors and activators bind as dimers and tetramers, not trimers or pentamers. The effect of trimers binding is not different from that of dimers: a buffering proteomic cloud needs to be formed, the state of the system is quite influenced by noise, the switching region (region in X^{ad} parameter space from the bifurcation point to $\Delta C > 0.9$) is quite large. Yet in a real system there is an effective cost of trimerization: the energy of trimer formation and a need for the diffusion of particles. For tetramers the effect of stochasticity becomes negligible. Effectively one tetramer is sufficient for the bifurcation to take place. The binding of tetramer repressors may be thought of as a mechanism for increasing the deterministic nature of the switch.

5.G.3 Binding of higher order oligomers as a competitve mechanism

This analysis, although it neglects some important features, allows for a more quantitative formulation of cooperativity. Since most biological switches are asymmetric, cooperativity is also used as a means of making genes with smaller chemical rates more competitive. Tetramer binding seems to have a different role than that of lower order multimers. It may be used by genes which need to react to very small concentrations of proteins, for example they turn on degradation mechanisms when even a small number of toxic molecules is present. Or they may act as an extra mechanism stabilizing the existent state of a gene, as seems to be the case for the cI gene of the λ phage. It seems tetramers are used as having either a stabilizing role or that of a drastic, all or none response to the protein distributions in the system. This formulation of the problem is naturally oversimplified, but it allows for general observations.

5.H The Case when Proteins are Produced in Bursts

Many proteins in biological systems, for example the Cro protein in λ phage are produced in bursts of N of the order of tens. We consider a symmetric switch, where proteins bind as dimers and are produced in bursts of N. The derivation of the moment equations for this case is presented in Appendix B.

5.H.1 The general mechanism

We discuss the effect of bursting phenomena on the example of a symmetric toggle switch when proteins bind as dimers, as that can offer the most insight, when compared to previous results. In this case switching takes place for much smaller values of the effective production rate parameter X^{ad} compared to when proteins are produced separately. Therefore even in the large ω limit, noise resulting from large protein number fluctuations plays a role in defining the region of stability of the switch, as the criterion of large X^{ad} is not reached. The number of proteins in

the cell when the bifurcation occurs is determined by the tendency that proteins are unbound from the DNA and does not change when proteins are produced in bursts. For the rates discussed in Fig. 5.17 the critical mean number of proteins present in the cell at which the bifurcation occurs is $n_c = 10 = X^{eq} = 100^{\frac{1}{2}}$. If proteins are produced in bursts of $N = 10$, as in the left hand figures, this value of n_c is achieved when $X^{ad} > 1$, that is proteins must get produced at a higher rate than they are destroyed to be able to sustain the steady state number of 10 proteins in the cell. In the figures on the right hand side of Fig. 5.17 proteins are produced in bursts of $N = 100$. In this case even when the degradation rate is larger than the production rate, the critical steady state number of proteins necessary for the bifurcation to take place, can be reached and a bistable switch is possible. A bistable switch can exist if the degradation rate exceeds the production rate even for burst sizes present in biology. For $X^{eq} = 100$, the order of the tendencies for proteins to be unbound from the DNA in the λ phage, the value of N for which $X_c^{ad} < 1$ is smaller than $N = 20$, the burst size for Cro proteins in the λ phage. X^{ad} at the critical point decreases as function of N (Fig. 5.18 *A*) and depends on the tendency that proteins are unbound from the DNA X^{eq} (Fig. 5.18 *B*) and the adiabaticity parameter, ω (Fig. 5.19).

If proteins are produced individually the span of the non-adiabatic regime is clear from Fig. 5.19. It corresponds to $\omega < 1$. The bifurcation curves show small discrepancies for larger values of the adiabaticity parameter. However for larger burst sizes there is a continuous change in the form of the bifurcation curves with ω. All of the solutions differ substantially from the deterministic treatment, as shown in Fig. 5.17 *A*.

5.H.2 The influence of the adiabaticity parameter on the bifurcation mechanism

Contrary to the $N = 1$ case, the effective production rate at the bifurcation point X_c^{ad}, grows with the increase of the adiabaticity parameter, for considerably

large burst sizes, as in the $N = 100$ example in Fig. 5.19. In this case each gene produces a large number of repressors at a time. The bifurcation takes place in a region with $X^{ad} < 1$, which corresponds to very small effective production rates, which denote very large death rates. Therefore in the region of parameter space before the bifurcation takes place both genes remain repressed ($C_1(i) < 0.5$) in the steady state, as opposed to the previously discussed situations, in which both genes had equal probabilities to be active ($C_1(i) > 0.5$). For large N bursts, the bifurcation takes place when one of the genes becomes unrepressed in the steady state. That is when the repressor cloud buffering the DNA becomes destabilized, not when the cloud forms as in the smaller N examples. For large N bursts, if the rate of unbinding from the DNA is fast compared to the protein degradation rate, larger effective production rates are needed for the buffering proteomic cloud to stabilize the DNA state, than for small ω (Fig. 5.19 C). The larger X^{ad} is, the more repressor molecules are present in the system, which corresponds to larger protein number fluctuations, which are necessary for one of the genes to become unrepressed. For slower DNA unbinding rates, the buffering proteomic cloud is smaller, since the protein number reaches a steady state before the DNA state does. Therefore the buffering proteomic cloud is destabilized at smaller values of X^{ad}. Hence, in the case of small ω the unrepressing bifurcation takes place for smaller effective production rates than for large ω. However if the unbinding rate from the DNA is very small, $\omega < 0.01$, X_c^{ad} as a function of the adiabaticity parameter grows again, as this corresponds to effectively large death rates, which need very high production rates to sustain a proteomic cloud buffering the DNA. If the effective production rate is too small in this case, the steady state number of proteins is too small to form the buffering proteomic cloud, although the burst size is enormous. In the very small ω limit the bifurcation cloud needs to be formed for the bifurcation to be possible, as in the mechanism present in the small N case. The value of X^{ad} at the bifurcation point in both the large and small ω limit is strongly governed by protein and DNA binding state fluctuations in the

system. For this reason the deterministic solution fails. It assumes the incorrect mechanism, in which the bifurcation is a result of repressing one of the genes. Such a scenario is possible if the death rate of proteins is slow enough to allow for the existence of $< n(i)_c >$ repressor molecules in the system at very small production rates ($C_1(1)^{biff,kin} = 0.5$) (Fig. 5.17 A and B). One can see that the order of taking the adiabatic limits in the steady state for proteins produced in large bursts is subtle and depends strongly on the parameters of the system, as the bifurcation is governed mainly by relative protein and DNA fluctuations, both of which are very large. Furthermore, the deterministic solution is closer to the small ω limit, which corresponds to slow DNA unbinding rates compared to protein number fluctuations. Deterministic results may therefore be misleading in the bursting situation, even for large ω.

The steady state comes about as a result of different mechanisms depending on the burst number N and the order of reaching the steady state by the protein and DNA binding site dynamics changes depending on ω. For small burst sizes, slower DNA unbinding rates require larger effective production rates to reach the steady state number of proteins necessary to form the buffering proteomic cloud than for large N. For larger burst sizes, faster DNA unbinding rates destabilize the buffering cloud of proteins for smaller effective production rates than in the small N case.

5.H.3 Consequences of bifurcation at smaller X^{ad} values

The divergence from the deterministic solution at the bifurcation point increases with the burst size, as is expected due to the enormous noise effect due to large N, on a system with a constant and independent of the burst size number of proteins at the bifurcation point. As already noted the number of proteins in a cell, is in the range of tens to hundreds, even if they are produced in bursts. This number is reached for smaller effective production rates for larger burst sizes than for small N values. Therefore systems where proteins are produced in bursts display

smaller values of X^{ad} and are more susceptible to noise if the number of proteins in the cell is to be of the order which is observed experimentally. Furthermore the noisy burst systems even for very large values of X^{ad} do not converge as closely to the deterministic solution as they do for the single protein production example. This can be seen from the form of the steady state moment equations. The interaction function $F(i)$ for the $N = 1$ case in the limit of large ω and X^{ad} converges to $F(i) \rightarrow < n(i) > + < n(i) >^2$ whereas the deterministic solution corresponds to $F(i) =< n(i) >^2$. Therefore for large mean values of proteins the two are equal. However in the case when $N > 1$, $F(i) \rightarrow < n(i) > (1 + (N-1)/2) + < n(i) >^2$, which requires $N << 2 < n(i) >$ for the effect of bursting to be negligible at very large N. The values of the effective production rate that correspond to values of the proteins seen experimentally seem to be small. Therefore we can say that effectively the role of bursting is to enable for the existence of a bistable solution at lower effective production rates, which determines a region of parameter space which has been previously unstudied. In this region one cannot make the adiabatic assumption that the change in the DNA state can be integrated out due to a separation of timescales. That assumption leads to erroneous results, predicting a region of bistability where explicit treatment of both timescales suggests monostability. Furthermore, for very large N, the region of bistability decreases with the adiabaticity parameter, making the disagreement of the stochastic solutions with those of the deterministic rate equations larger. The adiabatic approximation and the full solutions converge only in the regime of large ω and X^{ad}, the second of which is never fulfilled at the bifurcation point or for biological concentration for systems in which proteins are produced in large bursts.

5.H.4 Dependence on the DNA Binding Coefficient

Just as increasing the burst size, decreasing the tendency for proteins to not be bound to the DNA results in a different switching mechanism. The probability of the genes to be on falls to far smaller values than the 0.5 of the $N = 1$ case. If

the burst size is large both genes have a very low probability of being on before the critical number of proteins necessary for bifurcation is achieved. The same effect is observed if proteins are more likely to bind to the DNA (small X^{eq}) (Fig. 5.18 B). When the genes are more probable to bind a repressor and successful unbinding events are rare, earlier bifurcations in terms of X^{ad} result. As X^{eq} increases, the probability of the genes to be on at the bifurcation point decreases as repressors have a higher tendency of unbinding.

For very high values of the adiabaticity parameter, corresponding to high unbinding rates form the DNA binding site, the stable solution which corresponds to the off state and the unstable state merge and the system is monostable again, with only the on state present. This limit is also reached by keeping X^{ad} fixed but taking the burst size $N \to \infty$.

5.H.5 Probability distributions

In the case of the rates used in Fig. 5.20, $n_c = 32$ is the same as for $N = 1$, but we note a tenfold decrease in X_c^{ad} compared to when proteins are produced separately. When proteins are produced in bursts, the probability distributions have tails towards larger n, as opposed to the distributions for individual protein production. The mean number of proteins in the system for given states of the switch is similar to that of the $N = 1$ case, however the distributions with bursts are much broader, as could be expected. In this case even very fast unbinding rates from the DNA cannot correct for the enormous protein number fluctuations and one must explicitly keep track of the change of the DNA binding state. A system in which proteins are produced in bursts is very noisy, especially compared to the nearly deterministic case of proteins binding as tetramers.

5.H.6 Nonzero basal effective production rate

If there is a nonzero basal production rate the difference between the deterministic and stochastic solutions is also qualitative even for relatively small burst

sizes. In this case proteins are also produced in the off state, so there the number of repressors produced by the off gene after the bifurcation is nonzero, but equal to the burst size N, since $< n(i) >= N(X^{ad} + \delta X^{sw}(2C_1(i) - 1)) \rightarrow^{C_1(1)\rightarrow 0} Ng_2/k$. This number is equal for both the stochastic and deterministic solutions and is equal to 10 in the examples presented in Fig. 5.21 C and D. So production in bursts maintains a high level of repressor proteins, even for very small g_2/k values if the burst size is large. When using experimental data one must be very careful to consider the burst size when assuming the basal production level is zero. Furthermore, the value of the interaction function of the gene in the off state ($C_1(i) \sim 0$) for the stochastic case is much larger than for the deterministic case, due to the multiplication of $< n(i) >^2$ which gives $F(i) \rightarrow < n(i) >^2 (1 + k/(2g_2)) + Ng_2/(2k)$, for large ω, the effect of which is shown in Fig. 5.21 A and B. The number of repressor proteins produced by the off gene decreases as $g_2 \rightarrow 0$, as expected and the probability of the on gene to be active tends to one. The dependence of the effective production rate at which the bifurcation occurs on the adiabaticity parameter is analogous to that of $g_2 = 0$ case. The probability distributions for the gene which is active after the bifurcation in the on and off state are presented in Fig. 5.22 A and B, for large unbinding rates from the DNA, and Fig. 5.22 C and D, for small unbinding rates from the DNA. They exhibit maxima around $2X^{ad}$ for the on state and $2g_2/(2k)$ for the off state and display behavior analogous to that of proteins produced separataly, apart from the different curvature of the slopes for $n < N$ and $n > N$. For small ω values the protein numbers reach a steady state before the DNA states, hence we observe bimodal probability distributions. The mechanism of competition in this noisy burst system is different than in the single protein production case. If the gene is in the on state, probability states with higher n values are strongly occupied and there is hardly any probability flux into the lower n states. In the off state however, a flux pushes the system into the lower n states, essentially trapping it there, hence the difference in the slopes, as can be seen in Fig. 5.22 C and D. This is also true for the $g_2 = 0$ system when

proteins are produced in bursts.

5.I Limitations of the SCPF Treatment

The examples presented above cover a large class of two gene switches, all of which are exactly solvable within the SCPF approximation. An exact solution may be obtained within this approximation for systems of genetic networks and switching cascades. However the SCPF approximation does not allow for an exact analytical solution of all systems. If we try to model one of the simplest natural systems where regulation is achieved by means of a switch, that is the λ switch, we encounter a problem. The genes in the λ switch, apart from having a toggle like regulation, also exhibit auto-regulation, that is cI proteins can bind to OR3, repressing the cI gene, and the Cro proteins can bind to OR1 or OR2, enabling the RNA polymarase from transcribing the Cro gene [109, 115]. If we expand the master equation to account for self-regulation we add a $h_i n_i^p$ binding term to the $P_j(n_i)$ equations. Therefore the k^{th} moment equation will display a dependence on the $k+p^{th}$ moment and the set of equation will not exhibit closure. One can find the probability distribution for a single self-regulating single gene. However if we consider as system like the λ phage, where self regulation is also combined with regulation by another gene, the problem is no longer solvable exactly and demands a cutoff of the hierarchy or other approximations. We can nevertheless treat these systems using the variational method, as proposed by Sasai and Wolynes (Sasai and Wolynes, 2003). The fact that self-regulation renders the system incompletely solvable within the SCPF approximation, is not surprising, since it corresponds to the exact solution for such a system. Gene i is influenced only by the number of proteins it produces. It is independent of the state of the other gene. Therefore, as one would expect the full solution should depend on all moments of the distribution of gene i. However for systems such as the λ phage, we can treat all inter gene regulation effects exactly and truncate the self-regulation equation at

the highest order of the inter gene interaction, which would be six, corresponding to, for example, 3 cI proteins binding to the 3 operator sites.

5.J Conclusions

The self-consistent proteomic field approximation for stochastic switches reproduces many intuitive notions about their behavior. It proves to be a a very powerful tool that allows for the consideration, of all but one, of the basic building blocks of more general switches and networks. A switch with a self-repressing/activating gene cannot be solved exactly within the SCPF approximation, as in this case the approximation is equivalent to the full solution. Therefore the probability distribution is determined by an infinite number of moments. The probability distributions obtained for the systems considered in this paper are not symmetric and exhibit long tails. This anticipates problems for using the variational principle for finding probability distributions when one accounts for correlations between the two states. The possibility to expand this method to consider networks and cascades will allow for are more realistic treatment of complex systems with emergent behavior at low computational costs.

One can account for the mRNA step in the system by a adding a deterministic step which using a deterministic kinetic rate equation translates the number of mRNA molecules into proteins produced in bursts. This is a valid procedure, as as separately shown by [36] and [37], transcription noise is just amplified in the translation process. Therefore treating the mRNA step deterministically simply introduces another constant into the discussed case of proteins produced in bursts. Therefore the presented treatment of proteins produced in bursts with a modified effective production rate is a simple model of including mRNA in the system. Of course, the effect of mRNA is much more complicated, as it also introduces, for example time delay, between binding and production. This model in the present state neglects these effects.

Our analysis of a large class of switches, shows how particular elements contribute to the emergent behavior of functioning switches. Comparison of the stochastic and deterministic treatments of a single gene switch shows convergence in the region of fast rates of unbinding from the DNA compared to protein number fluctuations and large effective production rates. For symmetric switches when proteins are produced separately the two solutions converge after the bifurcation, but often differ when defining the region of parameter space, where the bifurcation occurs. The agreement between the deterministic and stochastic solutions, is especially good for symmetric switches, with $N = 1$ and a non-zero basal production rate. However even though the mean repressor protein levels in the cell are similar in both approximations, the probability distributions are broad and far from Poissonian, i.e. they are not completely characterized by these means. If the adiabaticity parameter is small ($\omega < 1$) the protein number state reach a steady state before the DNA binding state and we observe a bimodal probability distribution. For the symmetric switch noise has a destructive effect on the region of bistability. Increasing the adiabaticity parameter facilitates the formation of a buffering proteomic cloud around a gene, which leads to repression at lower effective production rates than for small ω.

As was already mentioned, the symmetric switch is hard to design and build experimentally. The asymmetric switch, which is the experimental toy system, is much more susceptible to noise than the symmetric switch and stochasticity has not only the destructive effect on the region of stability one might expect, but also introduces new phenomena and can be utilized to increase the bistable region. This is of fundamental importance, since experimentally one deals with asymmetric switches and these offer greater possibilities in artificially engineering new systems. As can also be learned from the asymmetric switch as well as from the analysis of binding of different oligomers, the region of bistability of a switch grows with increasing the interaction function. When creating artificial switches, one may argue a large region of bistability may be desired, so the switch reacts by the

forward or backward transition to very specific concentrations or production levels of a protein. If the experimental setup constrains the protein production rates, this can also be achieved by modifying the adiabaticity parameters of the system, which ensures the transition remains rapid and effective. Asymmetric switches, exhibit first order phase transitions. This size of the region of phase space, in which the forward and backward transitions occur grows with the tendency that proteins are unbound from the DNA of both genes. Large adiabaticity parameters stabilize the buffering proteomic cloud around the repressed gene and lead to the formation of an effectively repressing cloud for smaller numbers of repressors, in the forward transition, than for small ω, for the active gene.

Experimental data available at this point [111], suggest biological switches function in regions of high adiabaticity parameters from the deterministic point of view. Nevertheless, even for large values of adiabaticity parameters one must account for the DNA binding site fluctuations explicitly when proteins are produced in bursts. The deterministic solutions give qualitatively wrong results in biologically relevant areas of parameter space. The stochastic solutions for large burst sizes suggest that the bifurcation of the solution is a result of destabilizing of the repressor cloud buffering the DNA, not formation of the cloud as for smaller burst systems. The probability distribution therefore exhibit tails towards large n values, not as in the small N case towards small n values. The deterministic kinetics remains unchanged for large burst sized, unlike the stochastic kinetics, hence presenting results derived from a wrong mechanism. The definition of the adiabatic limit, when proteins are produced in bursts is not clear as in the $N = 1$ case, when it corresponds simply to $\omega < 1$. This ambiguity does not allow one to integrate out the degrees of freedom corresponding to the change in DNA binding site occupation. Such an approximation leads one to erroneously identify the regions of bistability. The switch with a nonzero basal production rate when proteins are produced in bursts results in probabilities to be on and mean numbers of proteins in the cell very different from those of the deterministic solution, even for

small effective basal production rates. If proteins are produced in bursts assuming that a small effective basal production rate may be approximated by a zero rate may be misleading. Binding of proteins produced in bursts results in a bifurcation transition for smaller values of the effective production rate. It is also a mechanism for making two genes in an asymmetric switch more competitive.

Binding of higher order oligomers leads to results closer to those of deterministic treatments, with narrower probability distributions. This can be experimentally used to stabilize DNA binding states. In this simple model tetramers seem to be the most optimum binders. The close to deterministic all or nothing switching they offer may be worth the effective cost of the energy of multimerization and diffusion of particles. Binding of higher order oligomers may be viewed as a simple model of cooperativity, which increases the competitiveness of genes in an asymmetric switch. Within the SCPF approximation monomers do not make good switches due to lack of nonlinearity in protein concentration. They do not exhibit a region of bistability. This model neglects any structural DNA-protein interactions and spatial dependence. Hence this conslusion is simply a result of the lack of cooperativity in the system. For small adiabaticity parameters, they do however exhibit bimodal probability distributions, unlike in the large ω limit.

The thorough investigation of different components of gene regulatory networks using the self-consistent proteomic field approximation provides a tool kit for engineering new switches and networks. Based on our analysis, if one would want to build a strong component of a switch out of a gene with relatively small chemical parameters, one could use components that utilize binding of tetramers and that produce proteins in bursts. This is what the Cro gene in the λ switch uses.

5.K Acknowledgements

The text and data of Chapter 5, in full, has been published in "Self-Consistent Proteomic Field Theory of Stochastic Gene Switches" by A. M. Walczak, M. Sasai, P.G. Wolynes in Biophys. J. (**88**), 828-850 (2005). The dissertation author was the primary investigator and author of this article.

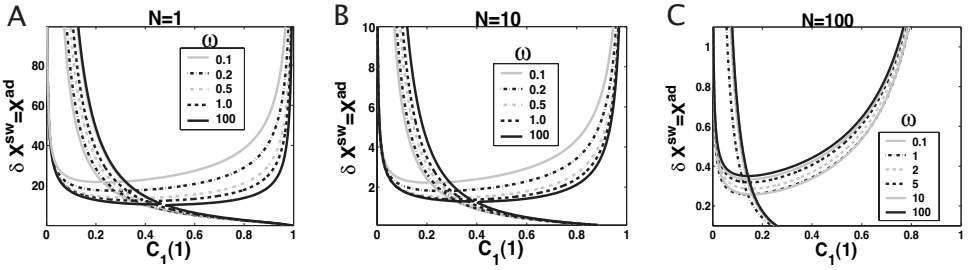

Figure 5.19: Bifurcation curves for proteins produced separately $N = 1$ (A), in bursts of $N = 10$ (B) and $N = 100$ (C) as a function of $X^{ad} = \delta X^{sw}$ for different values of the adiabaticity parameter.

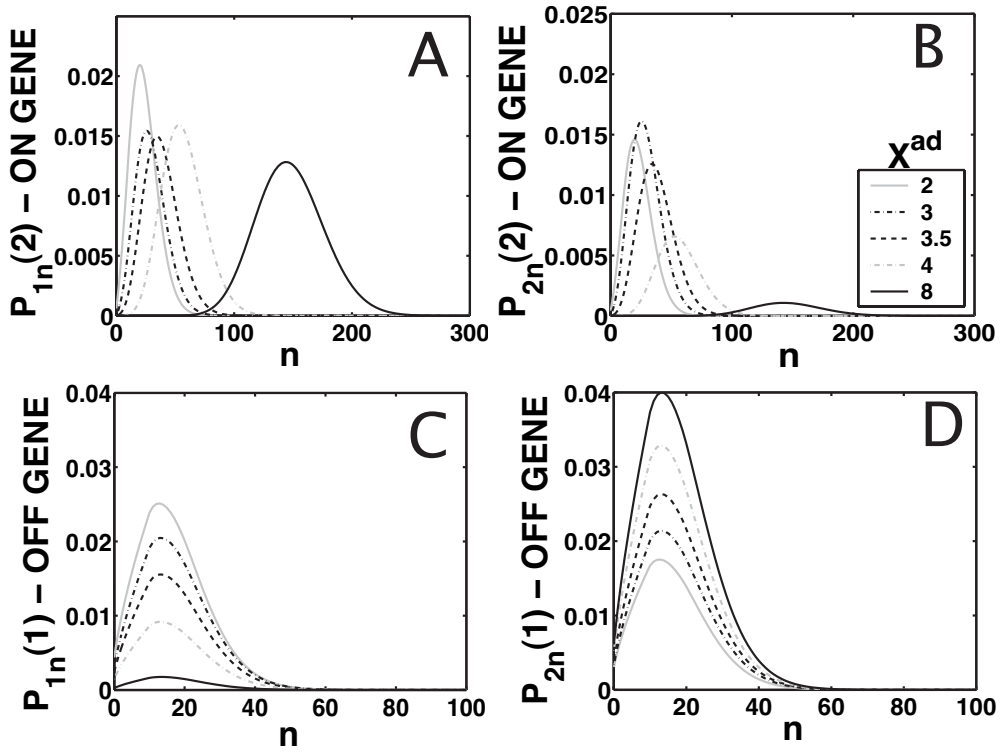

Figure 5.20: The evolution of the probability distribution of the gene that is active after the bifurcation, to be on (A) and off (B) and the gene that is inactive to be on (C) and off (D) as a function of X^{ad} for a switch when proteins are produced in bursts of $N = 10$, $X^{eq} = 1000$, $\omega = 100$. Bifurcation point at $X^{ad} = \delta X^{sw} = 35$.

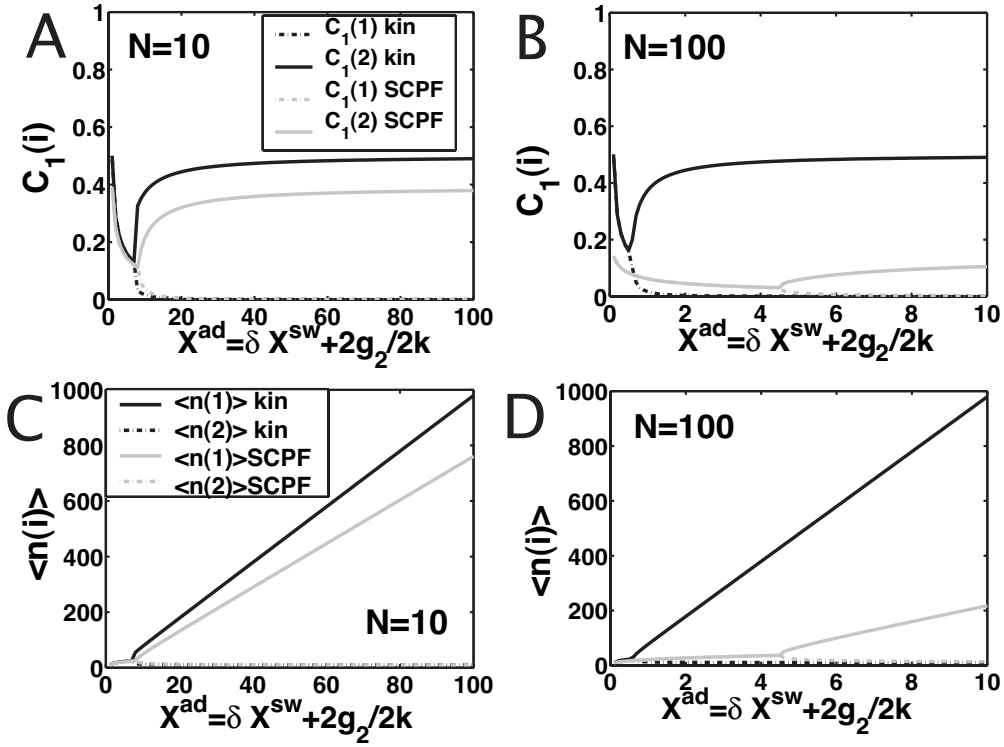

Figure 5.21: Probability that gene i is on when proteins are produced in bursts of $N = 10$ with a basal effective production rate $g_2/(2k) = 0.5$ (A) and $N = 100$, with a basal effective production rate $g_2/(2k) = 0.05$ (B). Mean number of proteins produced by each gene in the two cases (C and D). Symmetric switch, proteins bind as dimers, $X^{eq} = 100$, $\omega = 100$. Comparison of deterministic and stochastic solutions.

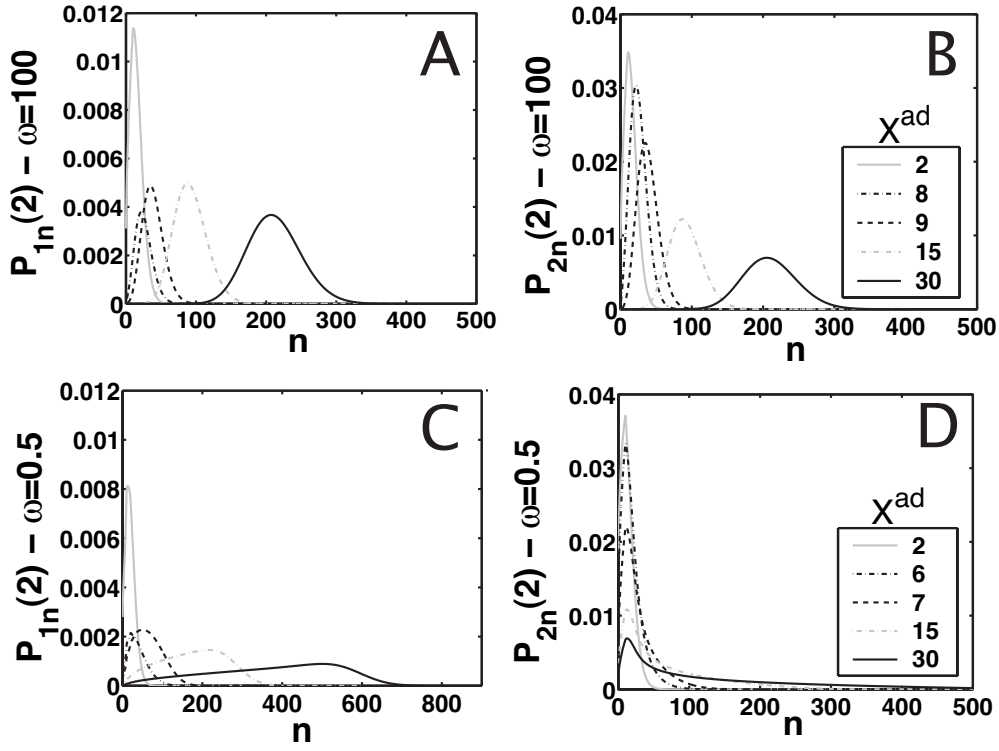

Figure 5.22: The evolution of the probability distribution when protein are produced in bursts. The evolution of the probability distribution of the gene that is on after the bifurcation, to be on for $\omega = 100$ (A and B) and $\omega = 0.5$ (C) and off (D) as a function of X^{ad} for a switch when proteins are produced in bursts of $N = 10$ with a basal effective production rate $g_2/(2k) = 0.5$, $X^{eq} = 100$. Bifurcation points at $X^{ad} = 8$ ($\omega = 100$) and $X^{ad} = 6$ ($\omega = 0.5$).

6

Absolute rate theories of epigenetic stability

Information may be passed from one cellular generation to another not just in the form of the DNA sequence, but also in the long lived expression patterns of genes. The epigenetic state of the cell i.e. which genes are expressed at a given time, is determined in part by binding and unbinding of transcription factor proteins to the DNA. The genes with their partner proteins form complex dynamical systems known as genetic networks, which can have many steady states i.e. an attractor landscape [6, 40]. The attractors are more stable than the individual molecular protein-DNA adducts, because the proteomic atmosphere of gene products renews the DNAs binding state, ultimately creating auto-catalytically its own proteomic atmosphere [59, 109, 40]. The attractors of such a genetic network may be associated with distinct cell types [6, 5]. Experimental evidence for this view has recently been presented [116, 19]. The growing experimental interest in this problem [19], as well as a number of theoretical puzzles involving the stability of the attractors [104, 117], call for a flexible and intuitive theory of the lifetime of such genetic network attractors. Some progress has already been made towards the goal [41, 38, 104, 39, 118], but existing formalisms are cumbersome, certainly when compared with the theory of activated events in molecular systems based ultimately on transition state ideas [108, 119]. Our goal here is to present a simple treatment of the noise induced transitions between two attractors on a landscape that is parallel to the treatment of simple molecular rate processes, which starts

with Wigner's absolute rate theory [120]. In chemical kinetics, the ratio of escape is proportional to the probability of rare configurations equally likely to become reactant or product. These rare configurations represent a stochastic separatrix of the motion.

While thermal atomic motions cause the escape from energy minima in molecular physics, the noise in genetic networks comes from the probabilistic nature of the chemical reactions, since only a small number of proteins and individual copies of the target DNA are involved. Unlike molecules, genetic systems being far from equilibrium, cannot strictly be described by thermodynamic free energy functions. The stochastic separatrix for molecular activated events is a dividing surface passing through saddle regions of the free energy. We argue that, even in the absence of a free energy function, the notion of a stochastic separatrix between basins of attraction remains a good approximation [41, 39, 118] and allows a treatment of stochastic switching along the lines of a transition state theory with dynamic corrections involving the rates of elementary processes [121, 108].

The dynamics of gene networks involves two very different processes whose rates must be compared- protein synthesis and DNA binding. The complexity and energy consuming nature of protein synthesis, in prokaryotic cells, generally causes changes in protein number to take longer than the diffusion controlled binding times of transcription factors, even at their low concentrations. For this reason, it has been argued that one can describe the binding of the transcription factors to each DNA binding site as an instantaneously equilibrated process, when considering protein production. For steady states this approximation appears to be reasonably accurate. It has, however, also been noted [20, 34, 52, **?**], that the DNA state fluctuations may influence the protein number state fluctuations. Here we show that the impact of DNA state fluctuations on the escape process, is considerable in a rather wide parameter regime for which the steady states are not much influenced by the DNA state fluctuations (Figure 6.1). For biologically relevant parameters, the DNA occupancy fluctuations may significantly increase the spon-

taneous switching rate from a given attractor. In what we call the nonadiabatic limit, where DNA state fluctuations dominate the protein number fluctuations, individual binding and unbinding events of the transcription factors are directly responsible for the transition. For much of the adiabatic regime, although the influence of DNA fluctuations on the steady state protein levels is negligible, these fluctuations still modify the lifetime of a state - we will call these transitions "weakly adiabatic". DNA occupancy fluctuations can only be neglected at very high values of the rate ratios, in what we call the strongly adiabatic limit.

As Sasai and Wolynes [40] have pointed out, the stochastic theory of a simple genetic switch, can be considered analogous to the physicists' Kondo problem or the chemists' electron transfer process, where DNA occupancy plays the role of a spin or electronic state variable [108]. In a simple and intuitive way, here we exploit these analogies to compute the lifetime of a genetic switch, using the idea of a landscape with a stochastic separatrix [41], much like in the earlier proposed threshold model [39]. Our treatment is quite analogous to that used for characterizing adiabatic vs nonadiabatic regimes of quantum rates [108, 119]. The approach, we present is easily generalizable to a many gene system. In the general case the present approximation yields the lifetime of a given state of the switch, which is governed purely by a few local properties of the landscape and does not require computing complicated trajectories. Global properties, sensed by the most probable escape paths, generally enter rates for far-from-equilibrium systems [103]. The present approach must, therefore, be admitted to be approximate. The simplicity hopefully will make up for some inaccuracy.

Several treatments of the mean first passage time between epigenetic states have already appeared. Most of these studies assume the DNA state equilibrates on a much faster time scale than the protein number [76, 122]. We refer to this as the adiabatic limit. In this limit the protein number states may then be treated as a continous variable giving an expression for the mean first passage time à la the Smoluchowski theory of diffusive rates as sketched by Bialek [38]. A more rigorous

approach finds the rate by constructing the most probable escape path [104] or by calculating the distribution of paths [41, 77]. These methods are powerful, but they are hard to visualize, especially for more complex switching systems. While the usually invoked adiabatic limit seems to be appropriate for simple switches in prokaryotes, it is not an obviously correct approximation for switches that have more complex operators, in which multiple protein elements must combinatorically assemble at a given site, slowing the binding [35]. Nonadiabatic effects should also play a significant role in eukaryotic systems where chromosome restructuring, which may be quite slow, dominates the epigenetic transition. Artificially engineered switches [65] may be constructed with parameters spanning the entire phase diagram.

6.A The Simplest Switch

For illustration we will present our ideas using the simplest example of a system in which we can consider the escape from one minimum to another - a bistable self-activating switch [**?**]. We emphasize the approach is more generally applicable. The self-activating switch consists of a single gene, which may be found in one of two states: on or off. In the off state proteins are produced at a basal level, but in the on state proteins are produced at an enhanced level, leading to a number, n, of proteins in the cell at any moment. The proteins act as activators by binding to the same operator site as the gene governing their production. We assume they bind as dimers with a rate $h(n) = hn(n-1)/2$. The unbinding of the transcription factors is described by a rate f. We neglect time delays due to mRNA synthesis etc. (which admittedly may play a key role), so that protein population dynamics is governed by a birth death process. Protein degradation occurs with a rate k, production with the activated rate $g_\uparrow$ in the on state and the basal rate $g_\downarrow$ in the off state. The system is characterized by a two state joint probability distribution $\vec{P}(n)$, describing the probability of having n proteins in the system

and the DNA binding site being in the bound (on-↑) or unbound (off-↓) state. A recent combined experimental and theoretical study [65] has brought attention to the bistability of a switch in previously unexplored limits, when the degree of operon repression is small. Our discussion will turn also to the nonadiabatic limit. Here the equilibration of the DNA and changes in the protein number occur on comparable time scales.

To compute escape rates from the steady state attractors one must determine the stochastic separatrix [41]. In the adiabatic limit, the position $n_A^\dagger$ of the minimum of the total probability distribution $P(n) = P_\uparrow(n) + P_\downarrow(n)$ is given by the condition of zero mean protein flow $dn/dt|_{|_{n=n_A^\dagger}} = (fg_\downarrow + h(n)g_\uparrow)/(f + h(n)) - kn|_{n=n_A^\dagger} = 0$. For a bistable switch, this equation is satisfied by three values of n; - one solution gives the separatrix, the other two the positions of the high and low protein number stable steady state attractors, $n_A^\downarrow$ and $n_A^\uparrow$. In the nonadiabatic limit the stochastic separatrix refers both to the DNA and protein number state. This results in a different value of the critical separatrix numbers $n_N^\dagger$ in the nonadiabatic, and $n_A^\dagger$ in the adiabatic limits. The direction of flow changes when the DNA state changes. Therefore the position of $n_N^\dagger$ corresponds to that number of proteins needed for the system to have comparable probability to be in the on or the off state. For simplicity we can approximate in the large n limit $h(n) = h/2n(n-1) \approx hn^2/2$ and determine the position of the nonadiabatic separatrix by means of mass action, using the chemical equilibrium constant K^{eq}: $n_N^\dagger = V\sqrt{K^{eq}}$, where $K^{eq} = 2f/(hV^2)$, where V is the cell volume. The steady state attractors in the nonadiabatic limit are determined by the birth-death processes in the particular DNA states: $n^\downarrow = g_\downarrow/k$ in the off state and $n^\uparrow = g_\uparrow/k$ in the on state. To function as a switch $n^\downarrow$ must be less than $n_N^\dagger$ and $n^\uparrow$ must be greater than $n_N^\dagger$. We can rewrite the adiabatic separatrix positions in terms of the volume scaled equilibrium constant $K^{eq}V^2$, which scales with $n_N^{\dagger 2}$, as $n_A^\dagger = K^{eq}/n^\uparrow$ and $n^\downarrow < n_A^\downarrow < n_A^\dagger < n_N^\dagger < n_A^\uparrow < n^\uparrow$.

6.B Nonadiabatic Rate Theory

Here we compute the rate of escape of the system from the low protein number attractor to the high protein number attractor (k_{on}) and vice versa (k_{off}). Since in the nonadiabatic limit the low protein number attractor corresponds to the off DNA occupancy state and the high protein number state corresponds to the on DNA occupancy state, the transition from the low protein number state to the high protein number state is requires the binding of an activator. Without the possibility of binding and unbinding, the dynamics in each attractor would be described by stochastic destruction and production of proteins alone, resulting in fluctuations of the mean protein number around each steady state. Consider a system maintained in the off DNA binding state and that now has $n^{\downarrow}$ proteins. The initial probability of being in the off DNA state, with precisely $n^{\downarrow}$ proteins present is $p_{off}(n^{\downarrow}) = P_{\downarrow}(n^{\downarrow})/(P_{\uparrow}(n^{\downarrow}) + P_{\downarrow}(n^{\downarrow}))$. $n^{\downarrow}$ may be generally assumed to be close to the mean number of proteins in the off state ($n^{\downarrow} = g_{\downarrow}/k$). If a binding event now occurs at time $t = 0$, the gene spontaneously flips into the on state and proteins are now produced at an enhanced rate. The protein number increases towards the mean number in the high protein state ($n^{\uparrow} = g_{\uparrow}/k$). If the activator does not unbind before the number of proteins becomes characteristic of the on state attractor a successful switching event will have taken place and the protein number will now fluctuate around the on steady state value. However, since, we are in the nonadiabatic limit, the timescales to reach the steady state for both the DNA binding state and protein synthesis and degradation are assumed comparable, so an activator may in fact unbind before reaching the separatrix at $n_N^{\dagger}$. If an activator does unbind during that time, the gene returns to an off state, albeit with a slightly higher number of proteins than initially. Another binding event will repeat the above scenario, until the protein number safely crosses the separatrix at $n_N^{\dagger}$ and the steady state corresponding to an activated gene is reached (Figure 6.2 *a*). The average time needed to cross the barrier from an initial point

$n^{\downarrow}$, which is also the time allowed for a unbinding event to occur, is the mean time to reach $n_N^{\dagger}$ for the enhanced production rate. The initial rate of binding an activator $h(n^{\downarrow}) = h/2n^{\downarrow}(n^{\downarrow}-1)$ must be modified to account for the possibility of unbinding again before the system crosses the separatrix. Summing of these attempted crossings, results in an expression for the rate of escape from the off state minimum ($n^{\downarrow}$) to the on state ($n > n_N^{\dagger}$) in the nonadiabatic regime given by:

$$k_{on}(n^{\downarrow}) = p_{off}(n^{\downarrow})h(n^{\downarrow})e^{-\int_{t(n^{\downarrow})}^{t(n_N^{\dagger})} f dt} \tag{6.1}$$

The exponential term gives the successful fraction of attempts to reach the protein number based separatrix, launched from the steady state $n^{\downarrow}$. The total time to reach the separatrix is given by $t(n_N^{\dagger}) - t(n^{\downarrow})$, as determined by the average flows in the initial DNA state and the mean time for an unbinding event to occur is f^{-1}. Explicitly, the escape rate from the off state, becomes $k_{on}(n^{\downarrow}) = p_{off}(n^{\downarrow})h(n^{\downarrow})((g_{\uparrow} - kn^{\downarrow})/(g_{\uparrow} - kn_N^{\dagger}))^{-f/k}$. The power-law term describes the motion on the surface with enhanced production after binding of the activator. In the nonadiabatic limit, the probability distributions for the on and off states are unimodal. Therefore it is unlikely for the gene to be in the on state is the number of proteins is small, thus $p_{off}(n^{\downarrow}) \approx 1$. If the protein number is large and the unbinding rate is comparable to the death rate this expression yields:

$$k_{on}(n^{\downarrow}) \sim h(n^{\downarrow})e^{-\frac{f}{k}(n_N^{\dagger} - n^{\downarrow})} \sim h(n^{\downarrow})e^{-\kappa\frac{\sqrt{(K^{eq}V^2)^3}}{n^{\uparrow 2}}} \tag{6.2}$$

where $\kappa = hg_{\uparrow}^2/(2k^3)$. In the extreme nonadiabatic limit $\kappa \to 0$, the first attempt may be successful hence the result simplifies to $k_{on}(n^{\downarrow}) \sim h(n^{\downarrow})$.

A similar calculation can be carried out starting from the other steady state. The escape rate from the on state, with $n^{\uparrow} > n_N^{\dagger}$ proteins, is given by the rate of binding of an activator at time $t = 0$, providing the system is in the on state $p_{on}(n^{\uparrow}) = P_{\uparrow}(n^{\uparrow})/(P_{\uparrow}(n^{\uparrow}) + P_{\downarrow}(n^{\uparrow}))$, reduced by the probability that an activator rebinds before the protein number decreases to numbers characteristic of in the off state ($n < n_N^{\dagger}$). The time available to rebind is calculated using protein production

at a basal level. The k_{off} rate is therefore:

$$k_{off}(n^{\uparrow}) = p_{on}(n^{\uparrow}) f e^{-\int_{t(n^{\uparrow})}^{t(n_N^{\dagger})} h[n(t)]dt} \tag{6.3}$$

For the off rate the mean free path before a rebinding event depends on the mean number of proteins in the system n. The argument of the exponential still describes the number of rebinding events. In the strongly nonadiabatic case, $p_{on}(n^{\uparrow}) \approx 1$, and for very large mean protein numbers the escape rate tends to:

$$k_{off}(n^{\uparrow}) \sim f e^{-\frac{h}{4k}({n^{\uparrow}}^2 - {n_N^{\dagger}}^2)} \sim f e^{-\frac{\kappa}{2}\frac{{n^{\uparrow}}^2 - K^{eq}V^2}{{n^{\uparrow}}^2}} \tag{6.4}$$

Due to the timescale separation in the nonadiabatic limit the system may be approximated as a two state system. The ratio of the escape rates, therefore yields the ratio of the probabilities to be in the individual steady states. The equilibrium constant for the "dressed" genetic states in the nonadiabatic limit $K^{GS} = k_{off}/k_{on}$ therefore becomes $K^{GS} \approx (n_N^{\dagger}/n^{\downarrow})^2 exp(-\kappa/2) = K^{eq}V^2/{n^{\downarrow}}^2 exp(-\kappa/2)$. When $\kappa = 0$ the proteomic atmosphere has no effect on the relative stability of the DNA occupancy, which follows the ordinary mass action law.

The formulae described above provide quite intuitive representations of specific escape mechanisms. These results may also be formally obtained via the path integral solution of the master equation by using the method described by Wang, Onuchic and Wolynes [123] for kinetic protein folding. This result also coincides with the heuristic approach of Ninio [124].

6.C Adiabatic Rate Theories: Weak and Strong Regimes

In the nonadiabatic limit the switch reaches the separatrix within the time for a few binding events, as schematically portrayed in panel *a* of Figure 6.2. In what we call the weakly adiabatic regime, the escape process proceeds differently. The DNA occupancy responds quickly to the changing proteomic atmosphere reaching a local steady state before the protein number changes by a large amount. The

average occupancy then determines the average local rate of protein synthesis and degradation. A few binding and unbinding events are required in the nonadiabatic limit, but in the adiabatic limit those events are much too common to allow the direct mechanism. One is tempted to equate the local diffusion rate to that coming from synthesis and degradation. But this temptation can only be rigorously indulged at an extraordinary high binding rate. Instead a random, but cyclic process of binding, growth and unbinding churns the protein number like a turbulent surf. The cyclic motions of eddies in an ocean wave, if interrupted contribute to a diffusive transport of flotsam to the shore. In the same way, in most of the weak adiabatic regime, protein numbers fluctuate from the mean flow through this "churning mechanism". The protein number, changes slightly with each cycle of binding/growth/unbinding and eventually reaches the separatrix point due to the resulting diffusive motion. One can show the system acts as if it were diffusing along an effective potential, whose gradient gives the mean flow expected from the average occupancy $V(n) = g_{eff}(n) - kn$ (panel b in Figure 6.2). The diffusion rate in this outwardly adiabatic regime though depends on the nonadiabatic events. Only at very high adiabaticity is diffusion ascribable to birth-death alone.

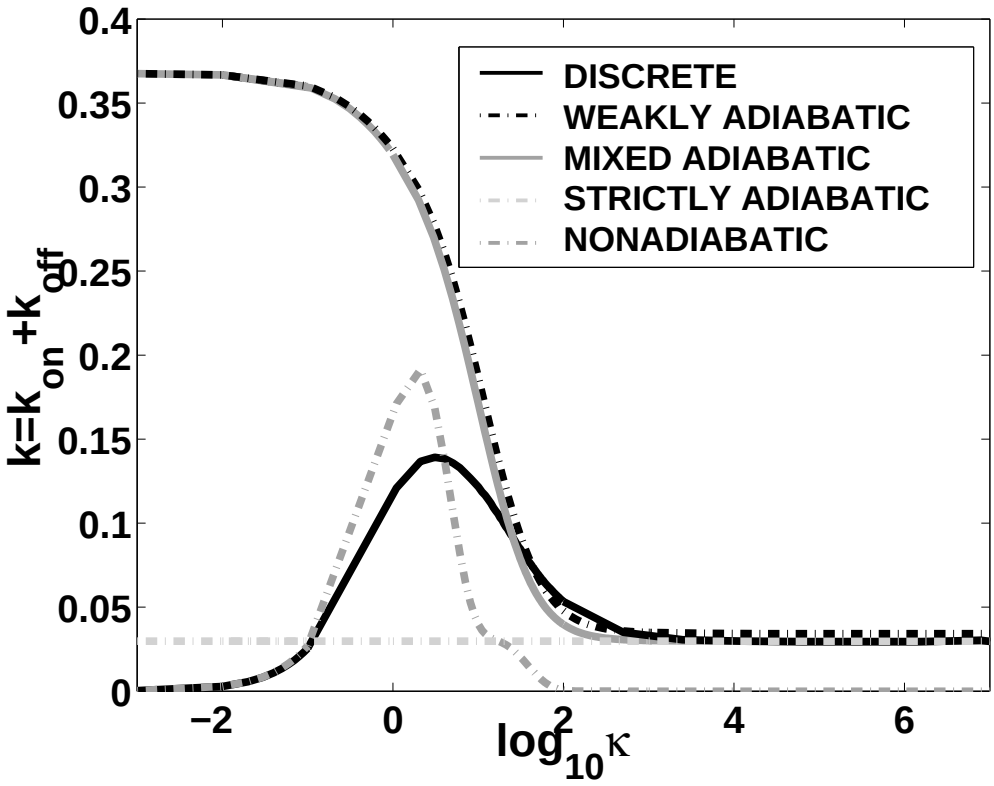

Figure 6.1: The sum of the escape rates $k = k_{on} + k_{off}$ as a function of the adiabiaticity parameter $\kappa = \frac{hg_{\uparrow}^2}{2k^3}$ for a self-activating switch with $g_{\uparrow} = 100, g_{\downarrow} = 8, k = 1, n_N^{\dagger} = 53.4$. Comparison of the exact discrete n numerical calculation based on the mean free passage time (black solid line), with approximate methods: in the nonadiabatic limit (small κ) (gray dashed line, Eqs 6.1 and 6.3), in the weakly adiabatic regime (black dashed line, Eqs 6.5 and 6.C) and mixed crossover regime (gray solid line). The adiabatic results tend asymptotically to the strictly adiabatic limit (large κ-flat escape rate) (light gray dashed line, Eqs 6.6 and 6.C). In the strictly adiabatic limit the binding of transcription factors to the DNA binding site is equilibrated. In the nonadiabatic and weakly adiabatic limits the escape rates show a dependence on the adiabaticity parameter- the process in influenced by the DNA binding state fluctuations.

It is helpful to understand the "eddy-induced" diffusion in an intuitive way. The effective production rate $g_{eff}(n) = (fg_{\downarrow} + h(n)g_{\uparrow})/(f + h(n))$ is the production rate averaged over the binding and unbinding states, if they were in equilibrium. The diffusion expected solely from the birth-death processes would just be $D_{BD}(n) = g_{eff}(n) + kn$. This fluctuation mechanism is augmented by diffusion in the orthogonal two state "binding-space", that is the eddy motion. The difference in the mean distance in protein number that would be travelled in the two DNA states during a typical eddy cycle will be $\Delta n = (g_{\uparrow} - g_{\downarrow})/(f + h(n))$. It is the typical difference in protein number expected after a full cycle of an eddy has been traversed. It is given by the difference in velocity in protein number space in a given binding state, $\Delta v = |g_{\uparrow} - g_{\downarrow}|$, times the mean time before the binding state changes $\Delta t = (f + h(n))^{-1}$, such that $\Delta n = \Delta v \Delta t$. The mean free time, or the eddy mixing time, is given by the sum of the characteristic times for binding and unbinding, both of which must occur to return to the original binding state, $\tau = f^{-1} + (h(n))^{-1}$. The rate which describes the eddy cycling thus becomes $\tau^{-1} = fh(n)/(f + h(n))$. The diffusion coefficient $D = \Delta n^2/\tau$ is the square of the mean change in protein number divided by the characteristic time spent within a given eddy. The latter depends on both binding and unbinding events. One thus finds $D_{binding}(n) = fh(n)(g_{\uparrow} - g_{\downarrow})^2/(f + h(n))^3$.

The mean number of proteins of a given type produced in the active state is of the order of $g_{\uparrow}/k \sim 10^2$. The degradation rate of proteins gives lifetimes of the order of a bacterial generation $k \sim 10^{-3}s^{-1}$. Dissociation rates from the DNA vary from $f \sim 1 - 10^{-3}s^{-1}$ and typical equilibrium constants may be taken to be $K^{eq}V^2 \sim 10^2 - 10^4$, which results in association constants $h/2 = f/(K^{eq}V^2) \sim 10^{-2} - 10^{-7}s^{-1}$ (based on λ phage data as assembled in [87] and references therein). We therefore see that typical adiabaticity parameters scan a wide range: $\kappa = hg_{\uparrow}^2/(2k^3) \sim 10^0 - 10^5$. The diffusion coefficient from churning, which depends on the DNA occupancy dynamics, typically influences the escape rate over four orders of magnitude of the adiabaticity parameter $\kappa \in (10^0 - 10^4)$, nearly

covering the biologically relevant regime. For escape processes the DNA binding dynamics cannot be neglected until the adiabaticity parameter becomes extremely large ultimately yielding the strongly adiabatic regime. As shown in Figure 6.2, the eddies due to the influence of the DNA binding state become smaller with faster binding, until the motion becomes dominated by simple birth-death diffusion along the effective potential, giving the steady state probabilities, averaged over the DNA binding states (panel c of Figure 6.2).

In the adiabatic limit, the escape rate is governed largely by the fraction of systems at the separatrix $N_A^\dagger$ compared to the fraction residing near the original attractor N_{attr}: $N_A^\dagger/N_{attr} = P(n_A^\dagger)/p_<^s(n_A^\dagger)$, where $p_<^s(n_A^\dagger) = \sum_{n<n_A^\dagger} P(n)$ and $P(n)$ is the steady state probability density for a state with n proteins. Clearly $p_<^s(n_A^\dagger) = P(n_{in})\delta n_{in}$, where δn_{in} is the width of the attractor. It is important to understand the spatial variation of $P(n)$, described by the "potentials" in Fig. 6.2. The spatial variation depends on the balance of the local mean flow against the flow due to diffusion. We can understand this balance by considering the motion pictured in Figure 6.2 b. The mean local velocity by which the protein number changes is $\bar{v} = g_{eff} - kn$. In addition to this drift the protein number changes by diffusive motion from places of low to high probability, with a velocity of $v_{diffusion} = D^i(n)/(2l_c)$, where l_c is a characteristic "lengthscale" over which the steady state probability changes by roughly one e-fold. $D^i(n)$ refers to the diffusion coefficient, which governs the motion in a particular regime. It is equal to $D_{binding}(n)$ in the weakly adiabatic regime, $D_{BD}(n)$ in the strictly adiabatic regime and is roughly $D_{BD}(n) + D_{binding}(n)$ in the small crossover region in between. To traverse this scale the local velocity has to be at least as large as the velocity of the diffusive motion $\bar{v} \geq v_{diffusion}$. The equality $\bar{v} = v_{diffusion}$ sets a characteristic length scale of the problem $l_c = D^i(n)/|2\bar{v}(n)|$, over which locally the probability in a steady state should change by a factor of e. This relation is valid both in the adiabatic and nonadiabatic regimes. The quantity l_c is analogous to the "scale height" in the equilibrium barometric problem. How many of these characteristic

steps of length l_c are needed in order for the system to reach $n_A^\dagger$ from its steady state value? Bearing in mind that the length of each step, depends on n, we must concatenate these steps to give the probability to be at the separatrix relative to being near the initial state. The probability exponentially depends on the number of scale heights of varying length l_c needed to reach the improbable separatrix starting from the most probable situation at the basin center, $\exp\left[-\int_{n_A^\downarrow}^{n_A^\dagger} dn l_c^{-1}\right]$.

To find the rate, we finally need the width δn_{in}. The size of the attractor δn_{in} is analogous to l_c at the bottom of the basin, but quadratic order effects must be included. To compare the velocities of the motion near the basin center due to drift and diffusion, the drift velocity must be computed as the "drift frequency" in the initial state $\omega(n_{in}) = (\partial v(n)/\partial n)|_{n_{in}} = fhn_{in}(g_\uparrow - g_\downarrow)/(f + h(n_{in}))^2 - k$ times the distance from the stationary point. Comparing drift and diffusion velocities in the same region $\omega(n_{in})\delta n_{in} = D^i(n_{in})/\delta n_{in}$, gives the size of the attractor $\delta n_{in} = \sqrt{|D^i(n_{in})/\omega(n_{in})|}$. The exponential term counts the paths from all possible position within the attractor. We must therefore divide the by the width of the attractor.

To determine the epigenetic escape rate we need also the transmission factor. In the adiabatic limit, reaching the separatrix does not yet guarantee a successful escape. Once the protein number reaches the vicinity of the stochastic separatrix the system may directly cross the separatrix, or recross it many times before committing to the new attractor. The number of escapes per unit time rate is thus proportional to the velocity with which the system moves over the separatrix, divided by the number of attempts before it successfully commits to the new attractor $k = \delta v/NP(n_{in} \to n \sim peak)$. The velocity around the peak is determined by a mean free path for number fluctuations l_{mfp} and a mean free time τ relevant to that region, $\delta v = l_{mfp}/\tau$. Only in the crossover region is it necessary to take all processes into account on equal footing when evaluating the mean free path l_{mfp} and the associated mean free time τ. In the weakly and strongly adiabatic limits the results simplify. In the weakly adiabatic region, the mean free path

is dominated by the DNA churning cycles and is given by the typical eddy size $l_{mfp} \approx (g_\uparrow - g_\downarrow)/(f + h(n))$ and $\tau = f^{-1} + (h(n))^{-1}$. In the strictly adiabatic limit, the motion is determined by the birth and death of proteins. Effectively, the protein number changes by l_{mfp} equal to one protein in the mean free time $\tau = (g_{eff}(n) + k)^{-1}$. Once the mean free path has been determined, the number of crossings is then the number of steps of the size of the mean free path needed to cross the transition state region l_{TST}. Like the basin size, the size of the transition state region is $l_{TST} = \sqrt{D^i(n_A^\dagger)/\omega(n_A^\dagger)}$. The escape rate from the left attractor, $n_A^\downarrow$, is $k_{on}(n_A^\downarrow) = l_{mfp}^2/(l_{TST}\tau)(\delta n_A^\downarrow)^{-1} e^{-\int_{n_A^\downarrow}^{n_A^\dagger} dn l_c^{-1}}$, where $l_{mfp}^2/\tau = D_i(n_A^\dagger)$ and i indicates BD, *binding* and *mixed* in the appropriate regimes. This gives the rate of the escape from the low protein number state in the weakly adiabatic regime:

$$k_{on}(n_A^\downarrow) = \frac{1}{2\pi} D^i(n_A^\dagger) \sqrt{\frac{|\omega(n_A^\downarrow)\omega(n_A^\dagger)|}{D^i(n_A^\dagger) D^i(n_A^\downarrow)}} e^{-\int_{n_A^\downarrow}^{n_A^\dagger} dn l_c^{-1}} \tag{6.5}$$

where $n_A^\dagger$ is the number of proteins corresponding to the minimum of the total steady state probability distribution. In the adiabatic regime the separatrix is given as the fixed point of the average flow: $g_{eff}(n_A^\dagger) = k n_A^\dagger$.

In the strictly adiabatic limit, the eddy motion may be neglected. So $l_c^{\kappa \to \infty}$ is determined solely by the equilibrated diffusion in protein number space $l_c^{\kappa \to \infty} \approx (g_{eff} + kn)/(2(g_{eff} - kn))$. All the components in Eq 6.5 can be obtained using quadrature, in this case, yielding a complex expansion. A more simplified result, explicit in terms of chemical rate constants, follows if we linearize l_c^{-1} in the region, which contributes most to the result of the integral. In this situation equation 6.5 becomes:

$$k_{on}(n_A^\downarrow) = \tilde{f}_1(K^{eq}) e^{-\frac{|l_c^{-1}(n_{min})|}{4(n_A^\dagger - n_{min})}(n_A^\dagger - n_A^\downarrow)^2} \tag{6.6}$$

where n_{min} number of proteins for which l_c^{-1} has the largest value. The largest value of l_c^{-1} corresponds to the the smallest characteristic length scale in the region of integration. The value of $l_c^{-1}(n_{min})$ scales as $n_{min} \sim V\sqrt{K^{eq}/2}$. The pre-exponential factor has the form $\tilde{f}_1(K^{eq}) = kV/(2\pi)\sqrt{K^{eq}/(a_0 n^{\uparrow 6})(n^{\uparrow 4} - (K^{eq}V^2)^2 - 2K^{eq}V^2(n^\uparrow)^2)}$,

where $a_0 = g_\downarrow/g_\uparrow$. The escape rate decreases with the equilibrium constant and system size. Using the dependence of the minimum of the integrand as a function of the equilibrium constant $K^{eq}V^2$, one finds the escape rate scales as $e^{-\alpha_1 n^{\uparrow -2}(K^{eq}V^2 - 3a_0 n^{\uparrow 2})^{3/2}}$, where α_1 is a numerical factor of the order of 1/2. The rate of escaping from the off state attractor exponentially decreases with increasing of the equilibrium constant.

How the escape rate depends on the molecular parameters, can be seen by assuming, for simplicity, a highly cooperative variation of the equilibrium DNA occupancy with protein concentration. In this case the effective production rate can be approximated by the production rate in the off state attractor, $g^{eff}(n) \approx g_\downarrow$. Now, the protein dynamics will be determined by the rates characteristic of the attractors, until the system reaches the separatrix. This approximation is like the threshold picture of Metzler and Wolynes [39]. In this approximation one finds:

$$k_{on}(n_A^\downarrow) = \tilde{f}_1(K^{eq}) e^{-\frac{1}{2}\frac{k(n_A^\dagger - n_A^\downarrow)^2}{k n_A^\dagger + g_\downarrow}} \tag{6.7}$$

When the cell is sufficiently small the separatrix merges with both attractors. In such a regime, this simple formula correctly predicts the functional dependence of the escape rate on the equilibrium constant and the protein production rates. When the separatrix begins to merge the attractor, the exponential term approaches unity. Thus stability is compromised. When the attractors merge with the separatrix the pre-exponential factor becomes important for quantitative analysis [122, 104].

In the κ dependent weakly adiabatic region, the probability distributions look qualitatively similar to those in the strictly adiabatic limit: the extrema do not change as κ increases. In the escape rate calculation, however, one compares the ratios of the probabilities near the minimum and the saddle regions. This ratio is significantly different in the weak and strong adiabatic regimes and strongly affects the spontaneous switching rates, as seen in Figure 6.1. In the weak adiabatic regime one finds the escape rates depend exponentially on the adiabaticity parameter κ.

The escape rate therefore is approximately dominated by

$$kV/(2\pi)K^{eq}{n^{\uparrow}}^3/a_0\sqrt{{n^{\uparrow}}^2-2K^{eq}V^2}/({n^{\uparrow}}^2+K^{eq}V^2)^3\cdot$$
$$\exp(-f/kK^{eq}V^2a_0/n^{\uparrow}(n_A^{\dagger}-n_A^{\downarrow})/(n_A^{\downarrow}n_A^{\dagger}))$$

In the weakly adiabatic regime, the effective growth rate can be well approximated as that with a fixed DNA occupancy, as in the Metzler-Wolynes threshhold model [39].

The transition can be treated from the high protein number state to the low protein number state much as above in the adiabatic limit. The rate of escape from high protein number to the low protein number depends on the relative probability that the system is to the right of the separatrix, characterized by a mean protein number $n^{\uparrow}$ compared to the steady state probability of being at the separatrix $n_A^{\dagger}$, $k(n^{\uparrow}\rightarrow n\sim peak)=P(n_A^{\dagger})/p_{>}^{s}(n_A^{\dagger})$. $p_{>}^{s}(x)=\sum_{n=x}^{n=\infty}P(n)$. The escape rate turns out to be:

$$k_{off}(n^{\uparrow})=\frac{1}{2\pi}D^i(n_A^{\dagger})\sqrt{\frac{|\omega(n^{\uparrow})\omega(n_A^{\dagger})|}{D^i(n_A^{\dagger})D^i(n^{\uparrow})}}e^{-\int_{n_A^{\dagger}}^{n^{\uparrow}}dnl_c^{-1}}$$

We can approximate l_c^{-1} in the strictly adiabatic limit as for the k_{on} calculation. Then the strictly adiabatic escape rate becomes:

$$k_{off}(n^{\uparrow})=\tilde{f}_2(K^{eq})e^{-\frac{l_c^{-1}(n_{max})}{4(n_{max}-n_A^{\dagger})}(n_A^{\uparrow}-n_A^{\dagger})^2} \tag{6.8}$$

where n_{max} is the number of proteins at the maximum of l_c^{-1}, which scales as $n_{max}\sim= V\sqrt{K^{eq}}$. The pre-exponential fatcor has the form

$$\tilde{f}_2(K^{eq})\approx kV/(2\pi)({n^{\uparrow}}^4-(K^{eq}V^2)^2-2K^{eq}V^2{n^{\uparrow}}^2)\sqrt{K^{eq}}/{n^{\uparrow}}^5$$

More explicitly the escape rate from the on state scales as $\sim e^{-\alpha_2{n^{\uparrow}}^{-2}\sqrt{(\zeta{n^{\uparrow}}^2-K^{eq}V^2)}^3}$, where $\alpha_2\approx 2$ and $\zeta\approx 1/4+a_0/2$ are constant numerical factors. The escape rate from the on state attractor exponentially increases with the increase of the equilibrium constant. A simple result is also obtained by replacing the effective production rate by the value of the effective production rate in the on state attractor

$g^{eff}(n) \approx g^{eff}(n_A^{\uparrow})$:

$$k_{off}(n_A^{\uparrow}) = \tilde{f}_2(K^{eq}) e^{-\frac{1}{2}\frac{k(n_A^{\uparrow}-n_A^{\dagger})^2}{kn_A^{\dagger}+g^{eff}(n_A^{\uparrow})}} \tag{6.9}$$

The equilibrium constant for the dressed genetic switch state in the strongly adiabatic limit is

$$\bar{K}^{GS} = k_{off}/k_{on} \sim f_r(K^{eq}) e^{-{n^{\uparrow}}^{-3}(\beta_1(\sqrt{(\zeta {n^{\uparrow}}^2 - K^{eq}V^2)}^3 - \beta_2\sqrt{(K^{eq}V^2 - 3a_0{n^{\uparrow}}^2)}^3)}$$

which sharply depends on the proteomic atmosphere.

$$f_r(K^{eq}) = \sqrt{a_0({n^{\uparrow}}^4 - (K^{eq}V^2)^2 - 2K^{eq}V^2{n^{\uparrow}}^2)/{n^{\uparrow}}^4} \tag{6.10}$$

$\beta_1 \approx 2$, $\beta_2 \approx 1/4$ are numerical factors.

In the weakly adiabatic regime the exponential term in the off escape rate becomes $\exp(-h/(2k){n^{\uparrow}}^{-2}/(K^{eq}V^2)(1/6(n_A^{\dagger})^6 - (n_A^{\uparrow})^6 - g^{eff}(n_A^{\uparrow})/(5k)((n_A^{\dagger})^5 - (n_A^{\uparrow})^5)))$. So in the weakly adiabatic limit the equilibrium coefficient for the dressed genetic switch states $\bar{K}^{GS} = k_{off}/k_{on}$ scales as

$$\bar{K}^{GS} \sim a_0{n^{\uparrow}}^3/\sqrt{K^{eq}V^2}^3 e^{-h/(2k)\xi_1(n^{\uparrow})^{-2}/(K^{eq}V^2)((n^{\uparrow})^6 - \xi_2(K^{eq}V^2)^3)} \tag{6.11}$$

where the coefficients are determined by the positions of the on and off state attractors and are of the order of $\xi_1 \approx 0.01$ and $\xi_2 \approx 100$.

Whether the switch is nonadiabatic or adiabatic can be determined by comparing the mean free path to the size of the transition region. If $l_{TST}/l_{mfp} > 1$ many crossings are required and the transition is adiabatic. If $l_{TST}/l_{mfp} < 1$ the system commits to the new attractor once it reaches the separatrix, hence the transition is nonadiabatic. In the strictly adiabatic regime the diffusion of the system is governed by protein diffusion induced by the birth-death process, as opposed to the weakly adiabatic regime, where diffusion due to churns dominates. A phase diagram showing the different escape mechanisms in parameter space for fixed $K^{eq}V^2$ is shown in Figure 6.3.

6.D Comparison with Numerically Exact Results

While the mechanism of spontaneous switching or epigenetic escape is different in the various regimes, we understand the rates in all regimes using the notion of a stochastic separatrix. We can compare these approximations with numerical calculations due to Kepler and Elston [76, 81] and our own full numerical results.

In the nonadiabatic limit (small $\kappa = hg_{\uparrow}^{2}/(2k^{3})$) the escape process is determined by the rate of DNA state fluctuations. In this regime the rates are given by equations 6.1 and 6.3 (gray dashed line) (Figure 6.1). These agree with the discrete numerical calculation of the mean free passage time from each basin. Our numerical calculations confirm that only in the extremely adiabatic limit (large κ - flat escape rate) can the DNA fluctuations safely be neglected. Only for this extreme limit does the lifetime become determined by protein synthesis/ degradation fluctuations alone (light gray dashed line). Estimates of the input parameter would suggest that the weakly adiabatic regime is common for biological switches. In the weakly adiabatic regime the escape rate does not just depend on occupancy averaged growth rates, but still depends on the adiabaticity parameter, as shown in Figure 6.1. Neglecting the influence of DNA fluctuations in this limit, as many treatments have done would give the extreme adiabatic asymptotic value of the escape rate also pictured on the graph. Both the strictly and weakly adiabatic regimes can be obtained from the more general calculation using the full diffusion coefficient. The full treatment is only required in a small crossover regime (gray solid line).

6.E Summary

Spontaneous transitions between attractors of genetic systems are caused by coupled stochastic fluctuations in the DNA state and protein number. Even in parameter regimes where the DNA state locally would appear to reach a steady state much more rapidly than the protein number state, the fluctuations due to

binding and unbinding of transcription factors greatly influence the protein number fluctuations and hence modify the rate of spontaneous transitions between epigenetic states. We call such a regime the weakly adiabatic by contrast to the strongly adiabatic limit, where the DNA binding state may be taken to be in equilibrium. The mechanism of spontaneous switching between stable attractors in the weakly adiabatic regime is graphically explained by a churning process, which causes protein numbers to fluctuate from the mean flow. How the escape rates k_{on} and k_{off} depend on molecular parameters in the nonadiabatic, weakly and strongly adiabatic should allow one to understand the evolutionary constraints necessary to achieve stable yet responsive switches, a topic we hope to return to. By considering both the DNA and protein degrees of freedom, the rate theories we have presented provide an intuitive description of spontaneous switching events, in terms of the molecular parameters that determine the functioning of a genetic switch.

6.F Acknowledgements

The text and data of Chapter 6, in full, has been published in "Absolute rate theories of epigenetic stability" by A. M. Walczak, J. N. Onuchic and P. G. Wolynes in Proc. Natl. Acad. Sci. USA (**102**), 18926, (2005). The dissertation author was the primary investigator and author of this article.

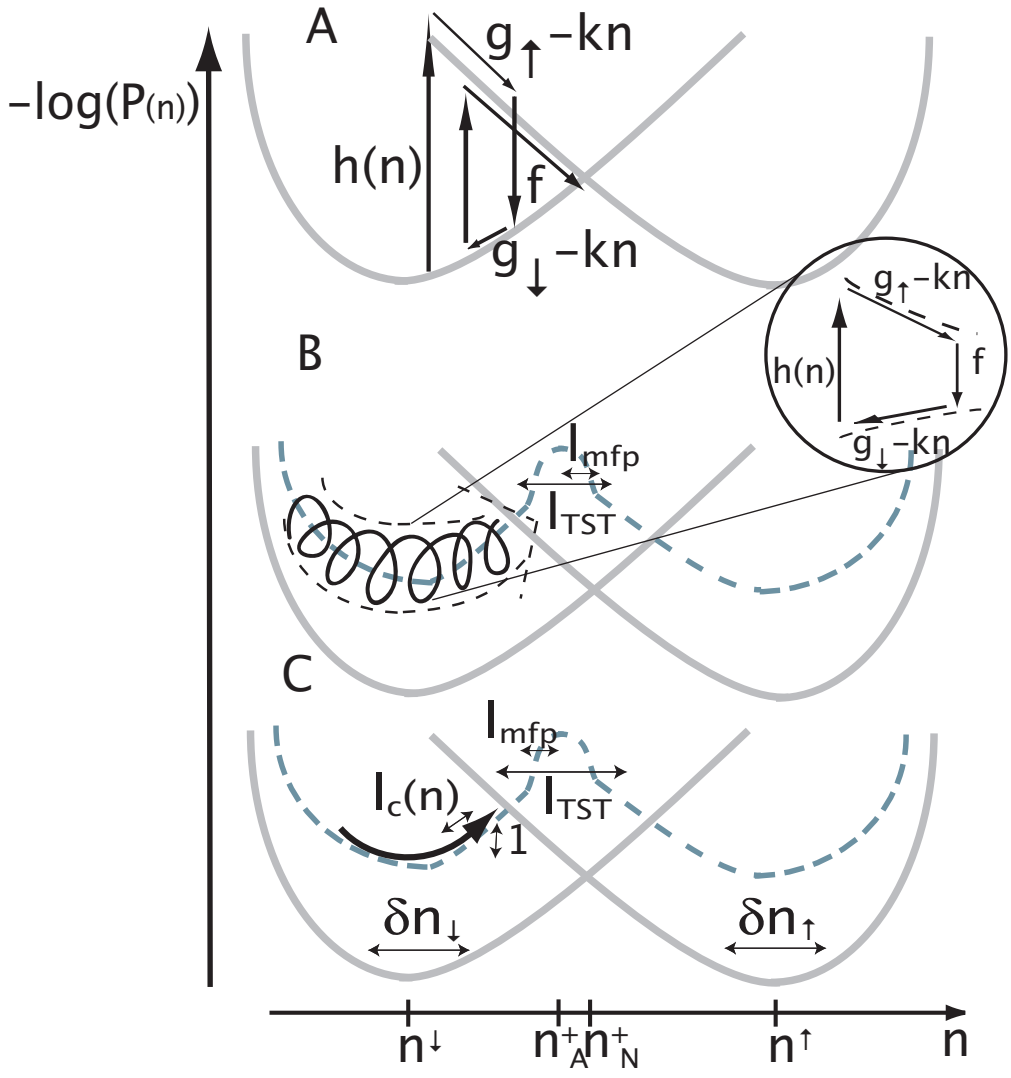

Figure 6.2: A schematic diagram of the difference in the character of the transitions from the state with a small number of proteins to the state with a large mean steady state number of proteins in the nonadiabatic, adiabatic and extremely adiabatic regimes. In the nonadiabatic limit (A) the escape rate is given by Eqs 6.1 and 6.3, in the adiabatic (B) and extremely adiabatic (C), the escape rate is given by Eqs 6.5 and 6.C. The dark gray line marks the effective potential for protein number change. The horizontal arrows signify binding and unbinding events.

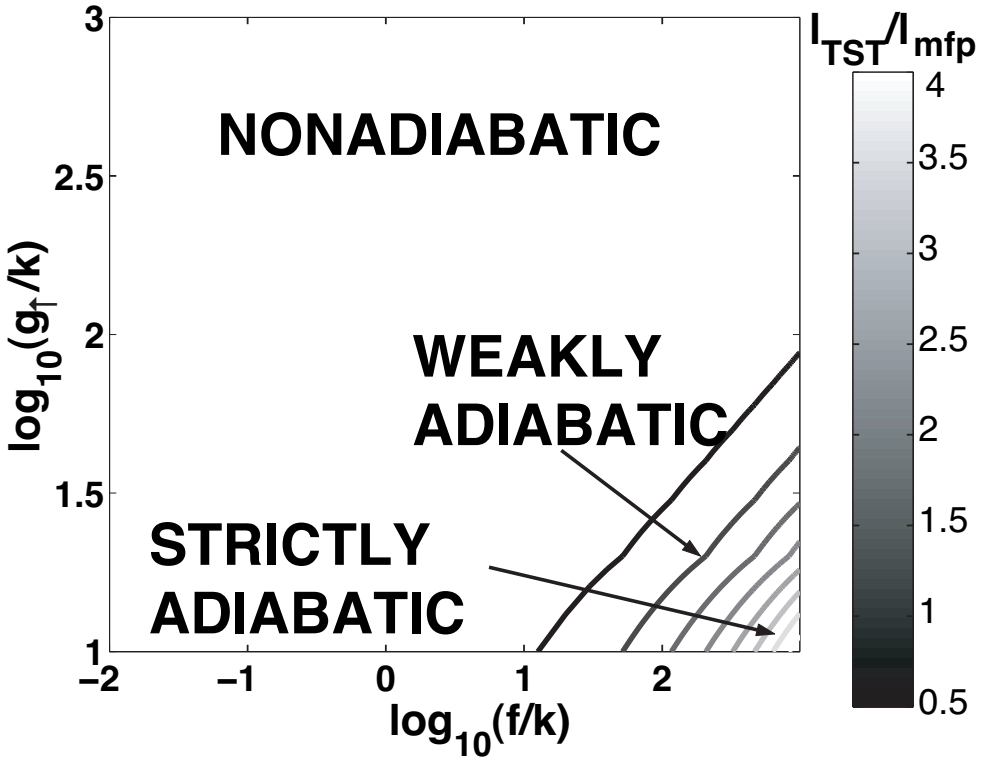

Figure 6.3: A phase diagram as a function of the activated production rate $g_\uparrow$ and the unbinding rate f for constant $K^{eq}V^2$, showing the areas of parameter space where a given escape mechanism dominates based on the ratio of the size of the transition state region l_{TST} to the mean free path l_{mfp}. If the number of crossings of the separatrix is large $\frac{l_{TST}}{l_{mfp}} > 1$, the transition is adiabatic. If the system commits to a new attractor after one crossing $\frac{l_{TST}}{l_{mfp}} < 1$ the transition is nonadiabatic.

7

Conclusions

Gene expression regulation networks are examples of systems of nonlinear units coupled together, which result in emergent steady states. A gene network may therefore be studied as a many-body system, which may function out of equilibrium. In this thesis I have tried to present how methods of theoretical many-body and nonequilibrium physics may be used to gain a better understanding of the interactions between genes and proteins which are the building blocks of these circuits. These methods have proven fruitful.

I set out to investigate the effect of the DNA operator occupancy fluctuations in genetic switches. I found that for amplified signals, such as when proteins are produced in bursts the mean number of proteins is influenced by the DNA binding state fluctuations. Furthermore, if we consider dynamical systems, even if the steady state probability distributions are weakly dependent on the DNA state, the escape rate in a large part of parameter space depends on DNA fluctuations. Recent experimental work by Juan Pedraza and Alexander van Oudenaarden [20] shows how noise generated in the expression of one gene is propagated to other genes, by binding of the product proteins of upstream genes. The fluctuations in the number of proteins produced in turn by these downstream genes are larger: noise in the expression of downstream genes is amplified. I therefore conclude that DNA state fluctuations could play a significant role in large networks. However our findings also lead to a lot of new questions.

The results presented here touch only upon a few problems and do not exploit all analogies to the maximum of their capacities. There are many ways to continue the research which is described in this thesis. From the biological perspective, it seems that questions of separation of timescales on which the gene expression state degrees of freedom and protein number state change, are more relevant in eukaryotes. Therefore incorporating spatial elements into gene regulation models, which would account for compartmentalization of processes and more complicated transport effects seems to be important [125, 126]. The spatially constrained expression of certain genes in development in eukaryotic systems is a fascinating example of how such models could be used to shed new light on experiments. Other elements of eukaryotic gene expression are also of interest in light of slow and fast timescales, such as chromatin unravelling [127, 128].

This is not to say that these effects are not important for prokaryotes. More and more experimental indications show [49] show slow binding kinetics of transcription factors in these organisms. Yet a question, which remains unanswered is why so many prokaryotes have evolved to function in the regime of fast binding and unbinding of transcription factors to the DNA.

Methods of theoretical physics can also be used to propose better approximations for the interaction of small genetic circuits in large networks. These questions, especially in the context of temporal expressions of genes, should help us develop a language needed when looking at large scale interactions. Especially if these interactions cannot be predicted from the behaviour of the particular circuits. The behaviour of the large network may also not have a mean field character. Thinking about such systems may result in new ways of describing other nonlinear coupled systems.

To conclude I would like to state that the integration of molecular systems biology and theoretical nonequilibrium many-body physics will continue to bring new ideas to both fields and should be pursued further.

Appendix A

Appendix

A.A Appendix A

In this appendix we derive the explicit form of the moment equations for the switch discussed in the section "The Toggle switch" of Chapter 5. In the operator formalism developed for classical diffusion by Doi (Doi, 1976) and Zeldovich' and Ovchinikov (Zeldovich and Ovchinikov, 1978), the number operator may be written in terms of number state creation $a^\dagger$ and annihilation a operators, as $n = a^\dagger a$. It is then particularly easy to write down the equations for the a moments instead of the n moments. Setting the left hand side to zero one obtains the steady state equations:

$$0 = -\omega_i[\frac{F(3-i)}{X_i^{eq}}C_1(i) - C_2(i)]$$

$$\begin{aligned} 0 &= k[(X_i^{ad} + (-1)^j \delta X_i^{sw}) < a_{ji}^{k-1} > - < a_{ji}^k >]C_j(i) + \\ &\quad +(-1)^j \omega_i[\frac{F(3-i)}{X_i^{eq}} < a_{1i}^k > C_1(i) - < a_{2i}^k > C_2(i)] \end{aligned}$$

Using the probability conservation relation $C_1(i) + C_2(i) = 1$, the zeroth order equations become:

$$C_1(i) = \frac{X_i^{eq}}{X_i^{eq} + F(3-i)} \qquad C_2(i) = \frac{F(3-i)}{X_i^{eq} + F(3-i)} \tag{A.1}$$

Dividing the higher order $a_j(i)$ moment equations by $C_j(i)$ and using the relation $C_1(i)/C_2(i) = F(3-i)/X_i^{eq}$ from the zeroth order equations one can calculate

$$< a_{1i}^k - a_{2i}^k > \;=\; \frac{(X_i^{ad} + \delta X_i^{sw}) < a_{1i}^{k-1} > -(X_i^{ad} - \delta X_i^{sw}) < a_{2i}^{k-1} >}{\omega_i + kC_j(i)} kC_j(i)$$

which depends only on a moments of lower order than the k^{th} moment. This allows one to obtain the following form for the higher order a moments

$$\begin{aligned} < a_{1i}^k > \;&=\; (X_i^{ad} + \delta X_i^{sw})(1 - \frac{\omega_i C_2(i)}{\omega_i + kC_1(i)}) < a_1^{k-1} > + \\ &\quad +(X_i^{ad} - \delta X_i^{sw}) \frac{\omega_i C_2(i)}{\omega_i + kC_1(i)} < a_2^{k-1} > \\ < a_{2i}^k > \;&=\; (X_i^{ad} - \delta X_i^{sw})(1 - \frac{\omega_i C_1(i)}{\omega_i + kC_1(i)}) < a_2^{k-1} > + \\ &\quad +(X_i^{ad} + \delta X_i^{sw}) \frac{\omega_i C_1(i)}{\omega_i + kC_1(i)} < a_1^{k-1} > \end{aligned}$$

Going back and forth between the two types of moments is straightforward. The n-moment equations have however more complicated forms:

$$
\begin{aligned}
< n_{1i}^k > \;=\; & \frac{1}{k}\Big[\sum_{s=0}^{k-1}\Big[\frac{k!}{s!(k-s)!}(X_i^{ad}+\delta X_i^{sw})(1-\frac{\omega_i C_2(i)}{\omega_i+C_1(i)k}) < n_{1i}^s > + \\
& +(X_i^{ad}-\delta X_i^{sw})\frac{\omega_i C_2(i)}{\omega_i+C_1(i)k} < n_{2i}^s > \Big] + \\
& +\sum_{s=0}^{k-2}\frac{k!}{s!(k-s)!}(-1)^{k-s}\Big[(1-\frac{\omega_i C_2(i)}{\omega_i+C_1(i)k}) < n_{1i}^{s+1} > + \\
& +\frac{\omega_i C_2(i)}{\omega_i+C_1(i)k} < n_{2i}^{s+1} > \Big]\Big] \\
< n_{2i}^k > \;=\; & \frac{1}{k}\Big[\sum_{s=0}^{k-1}\Big[\frac{k!}{s!(k-s)!}(X_i^{ad}-\delta X_i^{sw})(1-\frac{\omega_i C_1(i)}{\omega_i+C_1(i)k}) < n_{2i}^s > + \\
& +(X_i^{ad}+\delta X_i^{sw})\frac{\omega_i C_1(i)}{\omega_i+C_1(i)k} < n_{1i}^s > \Big] + \\
& +\sum_{s=0}^{k-2}\frac{k!}{s!(k-s)!}(-1)^{k-s}\Big[(1-\frac{\omega_i C_2(i)}{\omega_i+C_1(i)k}) < n_{2i}^{s+1} > + \\
& +\frac{\omega_i C_2(i)}{\omega_i+C_1(i)k} < n_{1i}^{s+1} > \Big]\Big]
\end{aligned}
$$

A.B Appendix B

In the case when proteins are produced in bursts of N and repressors bind as dimers the master equation has the form:

$$\begin{aligned}
\frac{\partial P_1(n_i)}{\partial t} &= g_1(i)[P_1(n_i - N) - P_1(n_i)] + k_i[(n_i + 1)P_1(n_i + 1) - n_i P_1(n_i)] + \\
&\quad -h_i n_{3-i}^2 P_1(n_i) + f_i P_2(n_i) \\
\frac{\partial P_2(n_i)}{\partial t} &= g_2(i)[P_2(n_i - N) - P_2(n_i)] + k_i[(n_i + 1)P_2(n_i + 1) - n_i P_2(n_i)] + \\
&\quad +h_i n_{3-i}^2 P_1(n_i) - f_i P_2(n_i)
\end{aligned}$$

for $n \geq N$. For $n < N$ the equations have the form.

$$\begin{aligned}
\frac{\partial P_1(n_i)}{\partial t} &= -g_1(i)P_1(n_i) + k_i[(n_i + 1)P_1(n_i + 1) - n_i P_1(n_i)] + \\
&\quad -h_i n_{3-i}^2 P_1(n_i) + f_i P_2(n_i) \\
\frac{\partial P_2(n_i)}{\partial t} &= -g_2(i)P_2(n_i) + k_i[(n_i + 1)P_2(n_i + 1) - n_i P_2(n_i)] + \\
&\quad +h_i n_{3-i}^2 P_1(n_i) - f_i P_2(n_i)
\end{aligned}$$

Following the same procedure as for the the single protein production case, we get the following equations of motion for the first three moments:

$$
\begin{aligned}
\frac{\partial C_1(i)}{\partial t} &= -h_i F(3-i) C_1(i) + f_i C_2(i) \\
\frac{\partial C_2(i)}{\partial t} &= h_i F(3-i) C_1(i) - f_i C_1(i) \\
\frac{\partial C_1(i) < n_1(i) >}{\partial t} &= [N g_1(i) - k_i < n_1(i) >] C_1(i) + \\
&\quad -h_i F(3-i) < n_1(i) > C_1(i) + f_i < n_2(i) > C_2(i) \\
\frac{\partial C_2(i) < n_2(i) >}{\partial t} &= [N g_2(i) - k_i < n_2(i) >] C_2(i) + \\
&\quad +h_i F(3-i) < n_1(i) > C_1(i) - f_i < n_2(i) > C_2(i) \\
\frac{\partial C_1(i) < n_1^2(i) >}{\partial t} &= g_1(i)[2N < n_1(i) > +N^2] C_1(i) + \\
&\quad +k_i[-2 < n_1^2(i) > + < n_1(i) >] C_1(i) \\
&\quad -h_i F(3-i) < n_1^2(i) > C_1(i) + f_i < n_2^2(i) > C_2(i) \\
\frac{\partial C_2(i) < n_2^2(i) >}{\partial t} &= g_2(i)[2N < n_2(i) > +N^2] C_2(i) + \\
&\quad +k_i[-2 < n_2^2(i) > + < n_2(i) >] C_1(i) \\
&\quad +h_i F(3-i) < n_1^2(i) > C_1(i) - f_i < n_2^2(i) > C_2(i)
\end{aligned}
$$

where $F(i) = C_1(i) < n_1^2(i) > + C_2(i) < n_2^2(i) >$ as before. Writing out $N^2 = N(N-1) + N$ and subtracting the $< n_j(i) >$ equations from $< n_j^2(i) >$ we get the equations of motion for the previously defined creation operators a. Due to the form of $F(i)$ for the dimer binding case only the first three moments are relevent. However generally this procedure can be carried out for higher moments, yielding

an expression for the m^{th} creation operator moment in the steady state of the form:

$$\begin{aligned}
< a_{1i}^m > &= (NX_i^{ad} + N\delta X_i^{sw})(1 - \frac{\omega_i C_2(i)}{\omega_i + mC_1(i)}) < a_1^{m-1} > + \\
&\quad +(NX_i^{ad} - N\delta X_i^{sw})\frac{\omega_i C_2(i)}{\omega_i + mC_1(i)} < a_2^{m-1} > + \\
&\quad +\frac{N^{m-1} - 1}{2}(NX_i^{ad} - N\delta X_i^{sw}(1 - \frac{\omega_i C_2(i)}{\omega_i + mC_1(i)})) \\
< a_{2i}^m > &= (X_i^{ad} - \delta X_i^{sw})(1 - \frac{\omega_i C_1(i)}{\omega_i + mC_1(i)}) < a_2^{m-1} > + \\
&\quad +(X_i^{ad} + \delta X_i^{sw})\frac{\omega_i C_1(i)}{\omega_i + mC_1(i)} < a_1^{m-1} > + \\
&\quad +\frac{N^{m-1} - 1}{2}(NX_i^{ad} - N\delta X_i^{sw}(1 - \frac{\omega_i C_1(i)}{\omega_i + mC_1(i)}))
\end{aligned}$$

To consider the binding of higher order oligomers when proteins are produced in bursts one simply accounts for the changed form of $F(i)$ as discussed in the "The Case when Proteins bind as Higher Order Oligomers" section.

Bibliography

[1] J. D. Watson and F. H. C. Crick, Nature **171**, 737 (1953).

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell, Fourth Edition* (Garland, New York NY, 2002).

[3] B. Lewin, *Genes IX* (Jones and Bartlett Publishers, Boston Ma, 2007).

[4] R. Wagner, *Transcription Regulation in Prokaryotes* (Oxford University Press, London, 2000).

[5] E. Davidson, J. P. Rast, P. Olivieri, A. Ransick, C. Calestani, C. H. Yuh, T. Minokawa, V. H. G. Amore, C. Arenas-Mena, and et al., Science **295**, 1669 (2002).

[6] S. Kauffman, *The Origins of Order* (Oxford University Press, London, 1993).

[7] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi, Nature **407**, 651 (2000).

[8] Y. Levy, J. N. Onuchic, and P. G. Wolynes, **129**, 738 (2007).

[9] Y. Levy, S. S. Cho, J. N. Onuchic, and P. G. Wolynes, JMB **346**, 1121 (2005).

[10] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North Holland, NL, 2001).

[11] M. Delbrueck, J Chem Phys **8**, 120 (1940).

[12] J. M. Raser and E. K. O'Shea, Science **307**, 1669 (2005).

[13] H. H. McAdams and A. Arkin, Trends in Genetics **15**, 65 (1999).

[14] J. Hasty, D. McMillen, F. Issacs, and J. J. Collins, Nature Review Genetics **2**, 268 (2001).

[15] D. Sprinzak and M. B. Elowitz, Nature **24**, 438 (2005).

[16] M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins, Nature Rev Genetics **6**, 451 (2005).

[17] M. B. Elowitz and S. Leibler, Nature **403**, 335 (2000).

[18] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain, Science **297**, 1183 (2002).

[19] M. Acar, A. Becskei, and A. van Oudenaarden, Nature **435**, 228 (2005).

[20] J. M. Pedraza and A. van Oudenaarden, Science **307**, 1965 (2005).

[21] N. Rosenfeld, M. B. Elowitz, and U. Alon, JMB **323**, 785 (2002).

[22] A. Becskei, B. Seraphin, and L. Serrano, EMBO Journal **20**, 2528 (2001).

[23] J. Paulsson, O. G. Berg, and M. Ehrenberg, Proc. Nat. Acad. Sci. **97**, 7148 (2000).

[24] J. Hasty, F. Issacs, M. Dolnik, D. McMillen, and J. J. Collins., Chaos **11**, 207 (2001).

[25] J. Hasty, J. Pradines, M. Dolnik, and J. J. Collins, Proc. Nat. Acad. Sci. **97**, 2075 (2000).

[26] K. G. McLure and P. W. Lee, EMBO J **17**, 3342 (1998).

[27] L. Cai, N. Friedman, and X. S. Xie, Nature **440**, 358 (2006).

[28] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz, Science **307**, 1962 (2005).

[29] A. Becskei, B. B. Kaufmann, and A. van Oudenaarden, Nature Genetics **37**, 937 (2005).

[30] J. T. Mettetal, D. Muzzey, J. M. Pedraza, E. M. Ozbudak, and A. van Oudenaarden, Proc. Nat. Acad. Sci. **103**, 7304 (2006).

[31] H. N. Lim and A. van Oudenaarden, Nature Genetics **39**, 269 (2007).

[32] J. Tsang and A. van Oudenaarden, Molecular Systems Biology **2**, 0025 (2006).

[33] P. Swain, JMB **344**, 965 (2004).

[34] J. Paulsson, Nature **427**, 415 (2004).

[35] N. E. Buchler, U. Gerland, and T. Hwa, Proc. Nat. Acad. Sci. **100**, 5136 (2003).

[36] M. Thattai and A. van Oudenaarden, Proc. Nat. Acad. Sci. **98**, 8614 (2001).

[37] P. S. Swain, M. B. Elowitz, and E. D. Siggia, Proc. Nat. Acad. Sci. **99**, 12795 (2002).

[38] W. Bialek, Advances in Neural Information Processing **13**, 159 (2001).

[39] R. Metzler and P. G. Wolynes, Chemical Physics **284**, 469 (2002).

[40] M. Sasai and P. G. Wolynes, Proc. Nat. Acad. Sci. **100**, 2374 (2003).

[41] P. B. Warren and P. R. ten Wolde, J Phys Chem B **109**, 6812 (2005).

[42] M. Scott, T. Hwa, and B. Ingalls, accepted to PNAS.

[43] G. Fritz, N. Buchler, T. Hwa, and U. Gerland, submitted 2006.

[44] N. Buchler, U. Gerland, and T. Hwa, Proc. Nat. Acad. Sci. **102**, 9559 (2005).

[45] C. V. Rao and A. Arkin., Annu. Rev. Biomed. Eng. **3**, 391 (2001).

[46] D. L. Cook, A. N. Gerber, and S. J. Tapscott., Proc. Nat. Acad. Sci. **95**, 15641 (1998).

[47] T. Lu, J. Hasty, and P. G. Wolynes, Biophys J **91**, 84 (2006).

[48] J. M. G. Vilar, H. Y. Kueh, N. Barkai, and S. Leibler, Proc. Nat. Acad. Sci. **99**, 5988 (2002).

[49] S. J. Maerkl and S. R. Quake, Science **315**, 233 (2007).

[50] I. Golding and E. C. Cox, Current Biology **16**, R371 (2006).

[51] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox, Cell **123**, 1025 (2005).

[52] A. M. Walczak, M. Sasai, and P. G. Wolynes, Biophys J **88**, 828 (2005).

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

[54] A. M. Walczak, J. N. Onuchic, and P. G. Wolynes, Proc. Nat. Acad. Sci. **102**, 18926 (2005).

[55] S. T. Gardner, C. R. Cantor, and J. J. Collins, Nature **403**, 339 (2000).

[56] H. C. Berg, *Random Walks in Biology* (Princeton University Press, Princeton, 1993).

[57] I. Golding and E. C. Cox, Phys. Rev. Lett. **96**, 098102 (2006).

[58] Y. M. Wang, J. O. Tegenfeldt, and W. R. et al., Proc. Nat. Acad. Sci. **102**, 9796 (2005).

[59] F. Jacob and J. Monod, JMB **3**, 318 (1961).

[60] E. Ozbudak, M. Thattai, I. Kurster, A. D. Grossman, and A. van Oudenaarden, Nature Genetics **31**, 69 (2002).

[61] E. Levine, T. Kuhlman, Z. Zhang, and T. Hwa, submitted 2006.

[62] T. Kuhlman, Z. Zhang, M. S. Jr, and T. Hwa, Proc. Nat. Acad. Sci. **104**, 6043 (2007).

[63] R. Tsien, Annual Review of Biochemistry **67**, 509 (1998).

[64] J. M. Raser and E. K. O'Shea, Science **304**, 1811 (2004).

[65] E. Ozbudak, M. Thattai, H. Lim, B. Shraiman, and A. van Oudenaarden, Nature **427**, 737 (2004).

[66] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, Proc. Nat. Acad. Sci. **101**, 17102 (2004).

[67] S. Kauffman, C. Peterson, B. Samuelsson, and C. Troein, Proc. Nat. Acad. Sci. **100**, 14796 (2003).

[68] J. M. G. Vilar, C. C. Guet, and S. Leibler, Journal of Cell Biology **161**, 471 (2003).

[69] G. K. Ackers, A. D. Johnson, and M. A. Shea, Proc. Nat. Acad. Sci. **79**, 1129 (1982).

[70] A. Becskei and L. Serrano, Nature **405**, 590 (2000).

[71] R. Hermsen, S. Tans, and P. R. ten Wolde, PLOS Computational Biology **12**, 1552 (2006).

[72] L. Bintu, N. E. Buchler, H. G. Garcia, and et al, Current Opinion in Genetics and Development **15**, 116 (2005).

[73] L. Bintu, N. E. Buchler, H. G. Garcia, and et al., Current Opinion in Genetics and Development **15**, 125 (2005).

[74] C. C. Guet, M. B. Elowitz, W. Hsing, and S. Leibler, Science **296**, 1466 (2002).

[75] A. Arkin, J. Ross, and H. H. McAdams, Genetics **149**, 1633 (1998).

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

[77] R. J. Allen, P. B. Warren, and P. R. ten Wolde, Phys. Rev. Lett. **94**, 018104 (2005).

[78] P. B. Warren, S. Tanase-Nicola, and P. R. ten Wolde, J Chem Phys **125**, 144904 (2006).

[79] S. Tanase-Nicola, P. B. Warren, and P. R. ten Wolde, Phys. Rev. Lett. **97**, 068102 (2006).

[80] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, Oxford UK, 2004).

[81] C. W. Gardiner, *Handbook of Stochastic Methods* (Springer-Verlag, Berlin, 1990).

[82] R. Kubo, M. Toda, and N. Hashitsume, *Statistical Mechanics II* (Springer-Verlag, Berlin, 1992).

[83] I. Bose, B. Ghosh, and R. Karmakar, Physica A **346**, 49 (2005).

[84] R. J. Allen, D. Frenkel, and P. R. ten Wolde, J Chem Phys **124**, 194111 (2006).

[85] R. J. Allen, D. Frenkel, and P. R. ten Wolde, J Chem Phys **124**, 024102 (2006).

[86] T. Ushikubo, W. Inoue, and M. Y. et al, Chemical Physics Letters **430**, 139 (2006).

[87] E. Aurell, S. Brown, J. Johanson, and K. Sneppen, Phys. Rev. E. **65**, 051914 (2002).

[88] H. H. McAdams and A. Arkin, Proc. Nat. Acad. Sci. **94**, .

[89] J. L. Cherry and F. R. Adler, Journal of Theoretical Biology **203**, 117 (2000).

[90] J. R. Pirone and T. Elston, Journal of Theoretical Biology **226**, 111 (2004).

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

[92] M. Doi, J Phys A **9**, 1479 (1976).

[93] M. Doi, J Phys A **9**, 1465 (1976).

[94] Y. B. Zeldovich and A. A. Ovchinikov, Sov. Phys. JETP **74**, 1588 (1978).

[95] A. S. Mikhailov, Physics Letters **85A**, 214 (1981).

[96] L. Peliti, Journal de Physique **46**, 1469 (1985).

[97] L. Peliti, Journal of Physics A- Mathematical and General **19**, L365 (1986).

[98] D. C. Mattis and M. L. Glasser, Rev Mod Phys **70**, 979 (1998).

[99] G. L. Eyink, Phys. Rev. E. **54**, 3419 (1996).

[100] H. A. Kramers, Physica A **7**, 284 (1940).

[101] M. Smoluchowski, Phys Z **17**, 557 (1916).

[102] P. Hanggi, P. Talkner, and M. Borkovec, Rev Mod Phys **62**, 251 (1990).

[103] R. S. Maier and D. L. Stein, Phys. Rev. E. **48**, 931 (1993).

[104] E. Aurell and K. Sneppen, Phys. Rev. Lett. **88**, 048101 (2002).

[105] P. Ao, Journal of Physics A-Mathematical and General **37**, 25 (2004).

[106] A. J. Leggett, S. Chakravarty, A. T. Dorsey, M. P. A. Fisher, A. Garg, and W. Zwerger, Rev Mod Phys **67**, 725 (1995).

[107] R. A. Marcus, Rev Mod Phys **65**, 599 (1993).

[108] J. Onuchic and P. G. Wolynes, Journal Phys Chem **92**, 6495 (1988).

[109] M. Ptashne, *A genetic switch* (Cell Press and Blackwell Science, Cambridge MA and Palo Alto CA, 1992).

[110] M. L. Simpson, C. D. Cox, and G. S. Sayler, Journal of Theoretical Biology **229**, 383 (2004).

[111] P. J. Darling, J. M. Holt, and G. K. Ackers, JMB **302**, 625 (2000).

[112] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, Applied Mathematics Series - 55, 1972).

[113] D. K. Hawley and W. R. McLure, JMB **157**, 493 (1982).

[114] J. Majewski and J. Ott, Genome Research **12**, 1827 (2002).

[115] M. Ptashne and A. Gann, *Genes and Signals* (Cold Spring Harbor Laboratory Press, New York, 2002).

[116] S. Huang, G. Eichler, Y. Bar-Yam, and D. E. Ingber, Phys. Rev. Lett. **94**, 128701 (2004).

[117] J. W. Little, D. P. Shepley, and D. W. Wert, EMBO J **18**, 4299 (1999).

[118] C. Kwon, P. Ao, and D. J. Thouless, Proc. Nat. Acad. Sci. **102**, 13029 (2005).

[119] P. G. Wolynes, Lectures in Science and Complexity, SFI Studies in the Sciences of Complexity, ed. D. Stein, **355** (1989).

[120] E. Wigner, Phys Rev **40**, 749 (1932).

[121] H. Frauenfelder and P. G. Wolynes, Science **229**, 337 (1985).

[122] D. M. Roma, R. A. O'Flanagan, A. E. Ruckenstein, A. M. Sengupta, and R. Mukhopadhyay, Phys. Rev. E. **71**, 011902 (2005).

[123] J. Wang, J. Onuchic, and P. G. Wolynes., Phys. Rev. Lett. **76**, 4861 (1996).

[124] J. Ninio, Proc. Nat. Acad. Sci. **84**, 663 (1987).

[125] J. S. van Zon, M. J. Morelli, and S. Tanase-Nicola. et al, Biophys. J **91**, 4350 (2006).

[126] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, Science **402**, C47 (1999).

[127] L. Y. Chen and J. Widom, Cell **120**, 37 (2005).

[128] W. Mobius, R. A. Neher, and U. Gerland, Phys. Rev. Lett. **97**, 208102 (2006).