Fluctuations in Genetic Networks: a Computational Study

Marco Morelli

Fluctuations in Genetic Networks: a Computational Study

Academisch Proefschrift

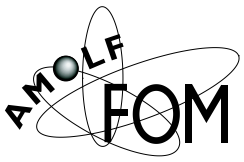
ter verkrijging van de graad van doctor aan de Universiteit van Amsterdam op gezag van de Rector Magnificus prof. dr. D. C. van den Boom ten overstaan van een door het college voor promoties ingestelde commissie, in het openbaar te verdedigen in de Aula der Universiteit op donderdag 18 oktober 2007, te 12:00 uur.

door

Marco Morelli

geboren te Bergamo, Italië

Promotor:	Prof. Dr. D. Frenkel
Co-promotor:	Dr. P. R. ten Wolde
Faculteit:	Natuurwetenschappen, Wiskunde en Informatica
Overige leden:	Prof. Dr. P. G. Bolhuis
	Prof. Dr. R. van Driel
	Prof. Dr. K. J. Hellingwerf
	Prof. Dr. F. C. MacKintosh
	Dr. S. J. Tans
	Dr. R. J. Allen



The work described in this thesis was performed at the FOM Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ, Amsterdam, The Netherlands. The work is part of the research program of the Stichting voor Fundamenteel Onderzoek der Materie (FOM) and was made possible by financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

Cover: Angelo Cacciuto

Subject headings: / computer simulations / biochemical networks / fluctuations / diffusion / genetic switches / bacteriophage λ

Alla memoria di mio padre,
Luigi

The work in this thesis is based on the following publications:

Chapter 2:
Computing stationary distributions in equilibrium and non-equilibrium systems with forward flux sampling
Valeriani C, Allen RJ, Morelli MJ, Frenkel D, ten Wolde PR
in press on J. Chem. Phys.

Reaction coordinates for the flipping of genetic switches
Morelli MJ, Allen RJ, Tănase-Nicola S, ten Wolde PR
submitted to Biophys. J.

Chapter 3:
Eliminating fast reactions in stochastic simulations of biochemical networks: a bistable genetic switch
Morelli MJ, Allen RJ, Tănase-Nicola S, ten Wolde PR
submitted to J. Chem. Phys.

Chapter 4:
A computational model for the bacteriophage λ genetic switch
Morelli MJ, Tănase-Nicola S, ten Wolde PR, Allen RJ
in preparation

Chapter 5:
Exact Reaction-Diffusion Brownian Dynamics
Morelli MJ, ten Wolde PR
in preparation

Chapter 6:
Diffusion of transcription factors can drastically enhance the noise in gene expression
van Zon JS, Morelli MJ, Tănase-Nicola S, ten Wolde PR
Biophys. J., **91**, 4350-4367 (2006)

Other publications of the author not based on this Thesis:

Pricing financial derivatives with path integral and neural networks
Montagna G, Morelli MJ, Nicrosini O, Farina M and Amato P
Physica A, **324**, 189-195 (2003).

Pricing financial derivatives with neural networks
Morelli MJ, Montagna G, Nicrosini O, Treccani M, Farina M and Amato P
Physica A, **338**, 160-165 (2004).

Contents

Chapter 1

Introduction

The lack of real contact between mathematics and biology
is either a tragedy, a scandal or a challenge.
Giancarlo Rota

Even the simplest organisms must be able to detect and respond to changes in their environment. In unicellular organisms, these decisions are performed by networks of proteins and DNA that chemically and physically interact with each other. In this Thesis, I will study gene regulatory networks in bacterial cells. I will use computer simulations to study the design principles of network motifs: overrepresented patterns of interactions that serve as the building blocks of gene regulatory networks. The central question is how these network motifs can operate reliably even in the presence of biochemical noise.

In this Introduction, I will first summarise the history of genetics, and discuss the principal findings that lead to the seminal work of Jacob and Monod on gene regulation. Since the concept of regulation of gene expression lies at the heart of this Thesis, I will discuss the principles of gene regulation in Section 1.2. In the next Section, I will discuss the statistical properties of gene regulatory networks, and subsequently address the various sources of biochemical noise they are prone to. The most widely-used descriptions to model gene regulation networks will be presented in Section 1.5, while Section 1.6 will discuss common and novel computational techniques that I used to simulate the systems. A brief outline of the thesis is printed at the end of Section 1.6.

1.1 Genetics: an historical background

The XIX century was preparing for another scientific revolution, as important and as contested as Copernicus' new vision of the universe. It is exactly from astronomy that some fundamental ideas about the dynamic nature of the universe were borrowed: the astronomer William Herschel understood that the solar system not only is currently moving through space, but also did all the planets form from a primitive state, a gas nebula, thus "downgrading" further the status of Earth from a religious viewpoint. The evidence for these discoveries came mostly from geological data. When Charles Darwin studied geology at the University of Edinburgh, he surely was challenged to get in touch with ideas. Perhaps his imagination was tickled by a distic written by Charles Lyell in his textbook "Principles of Geology":

> He that such quest would go must know not fear or failing
> To coward soul or fatihless heart the search were unavailing.

Lyell was referring to the problem of the origin of species and called it the "mystery of mysteries". On the 3rd of June 1836, during his trip on the Beagle, Charles Darwin visited the mathematician and astronomer John Herschel in Cape Town. His host "verbally aggressed" him on the question of the presence of extinct species in nature.

Darwin was profoundly influenced by Herschel's words, and started bridging the gap between his background in geology and his interest in the life sciences. This seminal meeting between two bright scientists, held almost two centuries ago, questioned one of the ascertained paradigms in science, namely that the species are created by God and do not change in time. The seeds of the spectacular successes of modern genetics are rooted in the questions these two english gentlemen brought to attention, and in the history of the answers that the scientific community slowly accumulated. It is then interesting to retrace this history, which has benefitted from contributions of scientists coming from a plethora of different backgrounds, and that was strongly influenced by the concomitant evolution of the *zeitgeist*.

Two decades later, in 1859, Charles Darwin published his book "The Origin of Species", where he summarised his investigations and formulated the revolutionary hypothesis that animal and vegetal species can change, driven by the fight for survival and competition for resources. Another echo of these ideas can be found in the theories of the economist Thomas Malthus, who believed that populations fight to survive by competing for food resources. As in every scientific major turning point, Darwin's ideas created a lot of scandal, and involved a radical re-thinking of the Lamarckian paradigm "the use creates the organ". Darwin's second famous book, "The Descent of Man"(1871) challenged the origin of human life on earth, and therefore its role: if human beings were also subject to the rules of evolution, it was no longer possible to believe they were collocated by God to rule over the world, but rather the result of an accidental process in common with all the other animals. Eventually, the "brute" nature of life, based on rough competition rather than on a divinely-planned harmony, was accepted by most intellectuals, yet leaving a crucial

question unanswered: how are the "characters" of individuals transmitted?

During this time, the Bohemian monk Gregor Mendel patiently cultivated and cross-bred generations of peas (*Pisum sativum*), and collected comprehensive statistics of how their hereditary characters are transmitted. He was aware of Darwin's results and driven by the same curiosity towards the roots of the manifest variety of nature. On the 8th of February 1865, during a meeting of the Natural History Society of Brünn (now Brno, Moravia), he illustrated the results of his meticulous records: in every plant, during mating the characteristics of the parents are combined according to precise statistical rules, and transmitted to the following generations. Mendel's studies in physics clearly influenced his reductionistic views: the simple "hereditary elements" which combine in fixed proportions bear much resemblance to the atomic entities recently discovered in physics and chemistry. With the physicists, he also shared a statistical approach to the scientific problems. Unfortunately, his results did not spread beyond the German-speaking scientific community, and the manuscript Mendel sent to Darwin was completely ignored by the English scientist. The nature of the "hereditary elements" stayed mysterious. However, since the observations at the microscope of the German biologists Theodor Schwann and Matthias Schleiden in the first half of the XIX century, it became clear that plants and animals are made of basic structural and functional units, called *cells*: the hereditary elements must then be searched for within the cells.

Mendel's results were re-discovered by the German-speaking scientists Hugo de Vries, Carl Correns and Erich von Tschermak, at the beginning of the XX century. In the same years, the German physicist Max Planck showed that energy is quantized and therefore is not a continuous quantity. The first advances of the quantum theory brought the idea of discreteness under the spotlight: the theory of elementary hereditary elements was reconsidered, and a new name was created for them: *genes*, from the greek word γενος (race, offspring). The science studying genes was called *genetics*.

The seminal experiments that shed light on the mystery of genetics in the second half of the XX century were primarily conducted on very simple organisms. It is therefore instructive to make a short digression and briefly explain the main features of such organisms. *Bacteria* are tiny, unicellular forms of life that were first seen by the Dutch tradesman Anthony van Leeuwenhoek with his hand-crafted microscope in the XVII century. They were extensively studied by the French chemist Louis Pasteur and the German physician Robert Koch in the second half on the XIX century. The outbreak of the French-Prussian conflict made the two scientists enemies; luckily, before that event they had already identified bacteria as the pathogenic micro-organisms responsible for many diseases. Bacteria are *prokaryotes*, that is their cells do not have a clearly defined nucleus and are much less spatially organised than cells showing a clear nucleus, called *eukaryotes*. Up to current times, prokaryotic cells are the preferred object of experimental investigation, due to their simplicity. There exists diseases that are carried by sub-microscopical

particles that can affect biological cells: viruses. Invisible under most microscopes and able to penetrate any known filter, these small, elusive agents were first identified on tobacco plants by the Russian biologist Dmitri Ivanovsky in 1892. Viruses are composed of genetic material encapsulated in a protein shell or in a membrane, and they cannot reproduce alone: they need to get into a host cell, and exploit its biochemical machinery to multiply as much as possible; eventually, the progeny is released, and the host is killed. Viruses lie therefore at interface between life and the inanimate world. Viruses infecting bacteria, called *bacteriophages*, were extensively studied in early molecular biology.

Genetics made one of its first steps in 1903 with the American biologist Walter Sutton who saw that when sex cells are formed, small filaments reorganize and randomly divide in the two daughter cells. Such structures were first observed by the German physician Walther Flemming and earned their name (chromosomes) by virtue of their capacity to be stained very strongly by many dyes. By cross-breeding mutated fruit flies *Drosophila melanogaster*, the American biologist Thomas Morgan was able to demonstrate in 1915 that genes are carried on chromosomes. Despite all this progress and acquisition of importance in science, genetics was still permeated by a deep mystery: what were genes made of, exactly? How do they reproduce? How do they transmit hereditary characters?

A bacteriologist, the American Oswald Avery, made in 1944 a breakthrough discovery, complementary to Morgan's observations: he demonstrated that all hereditary information is contained in a unique chemical species, deoxyribonucleic acid, or DNA, which was first isolated from the cellular nucleus by the Swiss physician Friedrich Miescher in 1863. In his experiments, Avery followed the seminal work of the British geneticist Frederick Griffith. He systematically removed various organic compounds from virulent bacteria, and checked if the remaining compounds were still able to infect new bacteria: when DNA was removed, the bacteria lost their infecting power.

The German physicist Max Delbrück, together with Carl Zimmer and Nikolai Timofeef-Ressovsky, proposed in 1935 that the genes must carry some kind of information, and that it should be written in a specific "alphabet". This work was profoundly influenced by the physics background of its authors, and tried to interpret the phenomenon of mutations on the basis of collision theory. Unfortunately, their mathematically challenging article was ignored by the biological community. His ideas nevertheless exerted an influence on Erwin Schrödinger's famous book of 1945, "What is Life?", arisen from a collection of lectures in Dublin. The idea that the transmission of the piece of information carried by a gene happens by the recursion of a limited number of symbols, whose distribution represent a message, spread into the scientific community, and the quest for the genetic code intensified.

During the XIX and the XX centuries, another piece of evidence had become clear in biology: the multiplicity of functions necessary for the development of life are sustained by a single type of molecule. These molecules, called *proteins* by the Swedish chemist Jöns Berzelius in 1838, are composed of chains of smaller elements, called amino acids,

which are now known to be of 20 different types.

Since 1929, it was known that the DNA molecule was also composed of repeating units, first recognised by the American biochemist Phoebus Levene with X-ray techniques. These units, called *nucleotides*, are formed by a phosphate, a base and a sugar, and are linked together through phosphate groups, which form the "backbone" of the molecule. Amusingly, Levene thought DNA was chemically too simple to store the genetic code.

The presence of repeated elements in two of the main molecules of life led the scientific community to hypothesise a correspondence between the two sequences. Along these lines, in 1941, the American geneticists George Beadle and Edward Tatum found that when the bread mold *Neurospora crassa* is irradiated with X-rays its capacity to produce enzyme changes. It was clear since the experiments on *Drosophila* performed by the American geneticist Hermann Muller in 1926 that the exposition to X-rays created mutations in organisms that could be inherited: Beadle and Tatum's experience thus demonstrated a profound link between genes and proteins, and suggested a one-to-one relationship. By the middle of the XX century, then, the unanswered questions of genetics had become: how did the DNA replicate in cells? How did the information contained in the DNA, which never leaves the cell nucleus, pass to proteins?

In order to shed light on these issues, the simple chemical composition of the DNA was not sufficient, and the quest for the three-dimensional structure of the molecule started. In 1953, the American biologist James Watson, together with the English physicist Francis Crick, using unpublished X-ray images and preliminary results from Rosalind Franklin, finally proposed the famous double helix model for the DNA. The helixes are formed by two phosphate backbones which lay at the outside of the structure, and are formed by 4 different kinds of nucleotides, that vary in their base part (adenosine, thymine, guanine, cytosine). Nucleotides are repeated aperiodically. The discovery that the amount of guanine was equal to cytosine and the amount of adenine was equal to thymine, led to the hypothesis of a specific base-pairing. This already suggests a simple, template-based mechanism that could be exploited to faithfully duplicate DNA. However, the relationship between the molecule carrying the information, DNA, and the proteins, was not yet completely understood: how does the information stored in the DNA pass to proteins, which then make the specificity of a form of life? How does the periodic structure of DNA, based on 4 different bases, map onto the protein chains, formed by 20 elements?

The situation started getting clearer in 1956, when the German biologist Vernon Ingram analysed the hemoglobin molecules in patients affected by the sickle cell disease. He determined that in all the sickle hemoglobins the same amino acid (the 6th) was mutated. It was already known that the origin of this disease was genetic, as it was transmitted according to Mendel's laws. These two indications together highlighted that not only did genes control the nature of amino acids in proteins, but also their positions. It was known that every cell contained another nucleic acid, called RNA (ribonucleic acid). Later, it

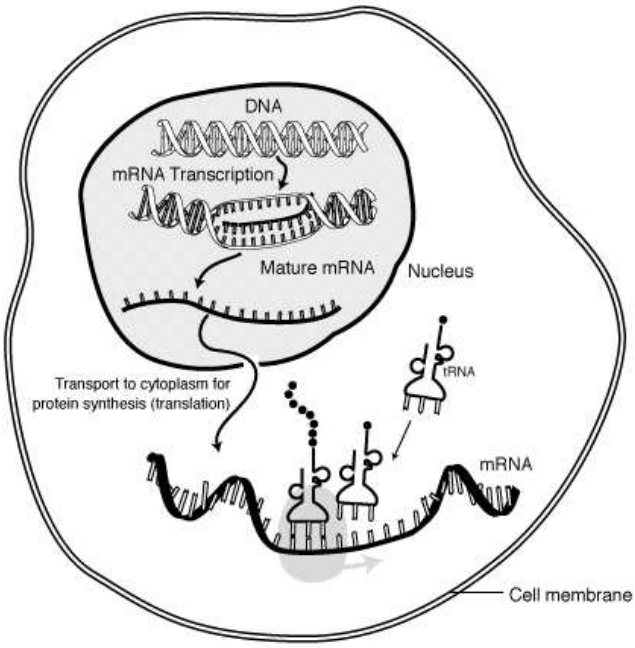


Figure 1.1: Central dogma of molecular biology. Genetic information is encoded on stretches of DNA called genes. Genes are copied to messenger RNAs by a process called transcription. In eukaryotic cells, mRNA is processed and migrates from the nucleus to the cytoplasm. Ribosomes "read"the information content of mRNAs and use it for protein synthesis (translation).

became clear that this molecule has a similar structure to DNA, but it is much smaller, and composed of a single chain of nucleotides, and does not necessarily have a helical structure. The Belgian biologist Hubert Chantrenne managed to isolate RNA from rabbits (*Oryctolagus cuniculus*), inject it into frogs of the *Xenopus* genus, and obtain rabbit hemoglobin proteins from these latter animals, thus elucidating the role of RNA in protein synthesis.

A correspondence between the codes of nucleic acids and proteins was almost certain. However, the different number of basic units posed a problem. Once more, the key idea towards the solution came from an outsider: the Russian physicist Georg Gamow, who proposed the idea that short sequences of bases could form a “code”, capable of carrying necessary information for the synthesis of proteins. Finally, in 1958 Crick stated the so-called *central dogma* of molecular biology, where he systematised the known state of the art of molecular biology. As it is depicted in Figure 1.1, the dogma states that genetic information can be transmitted between nucleic acids, by a process called *transcription* and from a nucleic acid to a protein, by a process called *translation* but not vice-versa, from protein to nucleic acids. The DNA needs not to directly intervene in the synthesis of proteins: the relevant part of the code (the genes) is copied to RNA molecules. When the RNA acts as an information carrier, it is called *messenger* RNA (mRNA). The mRNA can leave the nucleus, and the information it carries can be translated to proteins by other biochemical machines called *ribosomes*. Ribosomes are able to couple triplets of bases on the mRNA to transfer RNAs (tRNAs): specific molecules, each carrying an amino acid. tRNAs are recruited by ribosomes that add amino acids to a nascent protein chain, thus

faithfully translating the information present on the DNA. The genetic code is redundant, as out of the 64 possible triplets of basis, only 20 different amino acids are obtained. In 1961 the biochemists Marshall Nirenberg and Heinrich Matthaei demonstrated experimentally the central dogma: they created a synthetic mRNA, formed of a single repeated base, inserted it into an *E. coli* extract, and obtained a pseudoprotein composed of a single amino acid, the phenylalanine. This was a key step in deciphering the genetic code, which was quickly completed in the following years. This genetic code is universal, conserved in every living organism.

Since the discovery of the central dogma, molecular biology has taken off and genetics is now considered one of the most promising research areas in the XXI century. With new experimental techniques, the double helix is no longer a mere object of theoretical investigation, but rather a material that can be modified to make new mutant organisms that fulfill required tasks. Collaborations with industries are leading to applications which are revolutionising our everyday life.

However, the seeds of the *scientific revolution* that led to the birth of the molecular viewpoint and its successes were planted during a long period of time, largely benefitting from the influence of other disciplines, that contributed to modify the intellectual atmosphere of the time. As in the famous words of Bernard of Chartres, reported by John Salisbury in his Metalogicon (1159):

> We are like dwarfs sitting upon the shoulders of giants,
> and so able to see more and see farther than the ancients.

1.2 Principles of Gene Regulation

In 1961, the French biologists François Jacob and Jacques Monod explored the idea that the control of enzyme expression levels in cells is a result of feedback on the transcription of DNA sequences. After the determination of the structure and central importance of DNA, it became clear that the production of proteins might form a key control point. Jacob and Monod studied lactose metabolism in *E. coli*, and demonstrated that there are specific proteins that are devoted to repressing the transcription of the DNA to its mRNA product.

Accordingly, some genes code for proteins having a specific function in the cell, and other genes code for proteins whose task is to regulate the expression of other proteins. These regulatory interactions can govern key biochemical and cellular mechanisms, and achieve a notable level of complexity.

In this Thesis, I will study some features of gene regulation. However, the process of protein production and its regulation can be extremely complicated in eukaryotic cells. I will therefore focus on the regulatory interactions in simple organisms consisting of single prokaryotic cells, typically the bacterium *E. coli*. In these simple organisms, gene

transcription and mRNA translation lacks most of the additional complexity of higher forms of life, and allows a reductionistic approach, with the aim to extract the universal, elementary features of the process.

In prokaryotic cells, gene regulation happens mainly at the level of transcription [1]. The fundamental molecule involved in gene transcription is called *RNA polymerase* (RNAp): formed by the assembly of several subunits, it can grab the DNA with its crab-claw structure, and form a bound state called *closed complex*. The RNAp binds specifically to DNA: it can recognize specific sequences, called *promoters*, which are found immediately upstream of every gene. When the RNAp is bound to a promoter, it can separate the two DNA strands and open a bubble of about 13 base-pairs in the double helix. RNAp starts then walking along the DNA, reading the sequence of base pairs and simultaneously synthesising an mRNA molecule which bears a faithful copy of the same sequence (a process called *elongation*). During the process the bubble grows to about 17 base-pairs, moves together with the RNAp and it is eventually closed when a palindromic sequence causes the formation of a hairpin structure of the mRNA. At this point, RNAp detaches from the DNA and releases the mRNA molecule.

Proteins that regulate the transcription of a gene are called *transcription factors* (TFs). They typically exert their role by binding to the DNA on promotor regions, located immediately upstream of a gene. These proteins can either repress a gene by blocking the binding of RNAp, or activate a gene by enhancing the the open complex formation, or the affinity of RNAp for the promoter. Clearly, the abundance of TFs on the DNA depends on their concentration. In this way, the cell has a method to regulate the expression of a gene as a function of the concentration of another regulating molecule. This strategy can be used to express genes only when they are needed, thus realising a substantial saving of the available energetic resources.

It was already mentioned that the study of gene regulation started with the experiments by Monod on the *lac* genes in *E. coli*. This system represents an excellent example of the complex behaviours that a simple organism can achieve via gene regulation: it is instructive to briefly analyse it as an illustrative example (see Figure 1.2).

The three *lac* genes share the same promoter, and are therefore transcribed together in one operon: *lacZ* codes for an enzyme cleaving the lactose molecules into simpler sugars, *lacY* codes for the β-galactoside permease, a membrane-bound transport protein that pumps lactose into the cell, and *lacA*, an enzyme that transfers an acetyl group to β-galactosides.

The cell requires these genes for the transport and metabolism of lactose, a sugar that can be used by the bacterium as a carbon and energy source. However, there is a preferred energy source for the cell: glucose, which is easier to convert into ATP. In the presence of both sources of energy, it is therefore more advantageous for *E. coli* to primarily use glucose.

The promoter for the *lac* genes overlaps with a binding site for the lacI protein, also known as the lac repressor. lacI has a high affinity for the DNA: it can bind strongly,

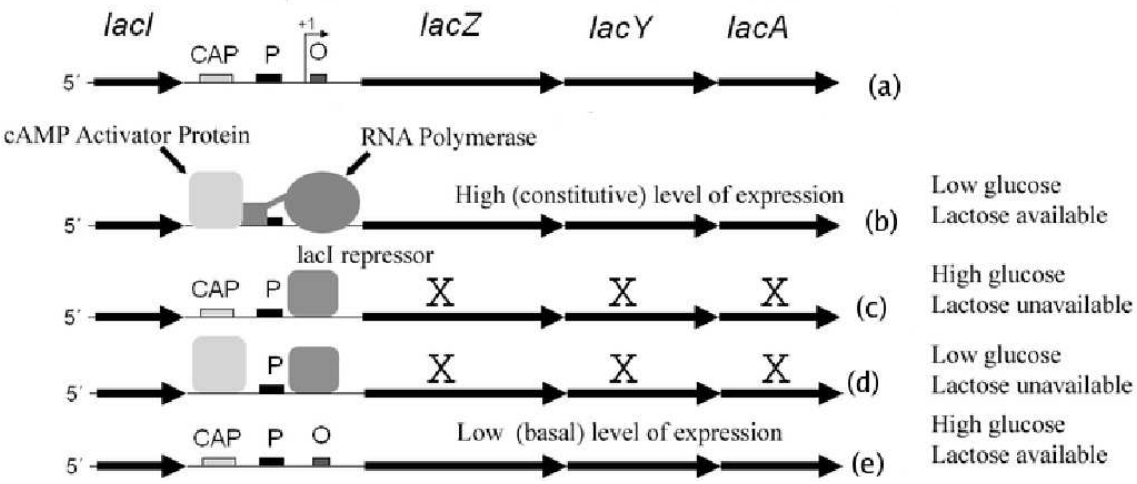


Figure 1.2: The *lac* operon regulates lactose metabolism in *E. coli*. Lactose is a secondary energy source, and it is used only when there is no glucose in the environment. (a) The genes coding for the LacZ, LacY and LacA proteins are regulated and transcribed together. Regulation happens by binding of several molecules to DNA, on a region upstream of the operon: LacI binds on the operator O, RNA polymerase to the promoter P, and CAP to its own binding site. (b) In case of low glucose and high lactose, the CAP-cAMP complex binds to the DNA and helps RNA polymerase to bind: the *lac* genes are strongly expressed. (c) For high glucose and low lactose, the lac repressor strongly binds to the DNA and blocks the transcription of the *lac* genes (large crosses indicate a repressed gene). (d) For low lactose and low glucose, both the lac repressor and the CAP-cAMP complex are bound to the DNA, and the *lac* genes are turned off. (e) For high glucose and high lactose, the DNA is free and the *lac* genes are weakly transcribed at the basal rate.

block the association of RNAp to the promoter, and turn off the transcription of the *lac* operon. When lactose is present however, one of its metabolites can bind to the lac repressor, which undergoes a conformational change and drastically lowers its affinity to DNA. RNAp can therefore bind to the *lac* promoter, and start transcribing the genes.

The intrinsic affinity of RNAp to the *lac* promoter is low, and the level of gene expression may not be sufficient. Another mechanism helps then the bacterium to enhance the production of the *lac* operon beyond the basal level, by means of another protein, called CAP (Catabolyte Activator Protein), which can bind upstream of the *lac* promoter. When CAP is bound, some of its residues can interact with some subunits of the RNA polymerase, strongly facilitating its binding and thus the frequency of the gene transcription. CAP is therefore an *activator* for the *lac* genes. CAP only works when it is bound to another molecule, called cyclic AMP (cAMP). This in turn is produced only when glucose is absent from the cell.

The *lac* genes are then both positively and negatively regulated, being under the control of both an activator and a repressor molecule. Using a genetic network based on these simple mechanisms, the bacterium is able to choose between two different nourishments, with one being preferred.

1.3 Large-scale properties of Genetic Networks

Since a few years, high-throughput techniques allow the simultaneous execution of millions of biochemical, genetic or pharmacological assays. The most successful technique is based on microarrays: plastic or glass plates featuring a grid of small spots, containing an antibody, or a small piece of DNA or RNA, that can be used to measure, respectively, a large number of protein-protein (as well as other specific properties like post-translational modifications), or protein-DNA interactions simultaneously. Each spot is used for a single experiment, whose qualitative output can be automatically revealed by a dedicated machine. Protein-protein interactions can also be screened with the yeast two-hybrid method.

With this qualitative experimental evidence, it is possible to visualise and study the properties of a generic transcriptional regulatory network [2]: one represents the elements genes and regulating proteins as *nodes*, and connects pairs of them with (possibly directed) *edges* when there is an interaction between the two nodes. The graphs obtained with this procedure are well known to mathematicians (probably the first proof in this field was given by Euler in 1735 when he solved the famous problem of the Köninsberg bridges). However, in the past few decades the exponentially increasing amount of available data has made it possible to characterise networks of huge dimensions, for which a direct visual analysis is not informative. As an example, Figure 1.3 shows a representation of the *E. coli* transcriptional regulatory network. It is then necessary to define statistical quantities that characterise the network, and that can be numerically measured.

The most elementary characteristic of a node is its *degree* k: the number of edges

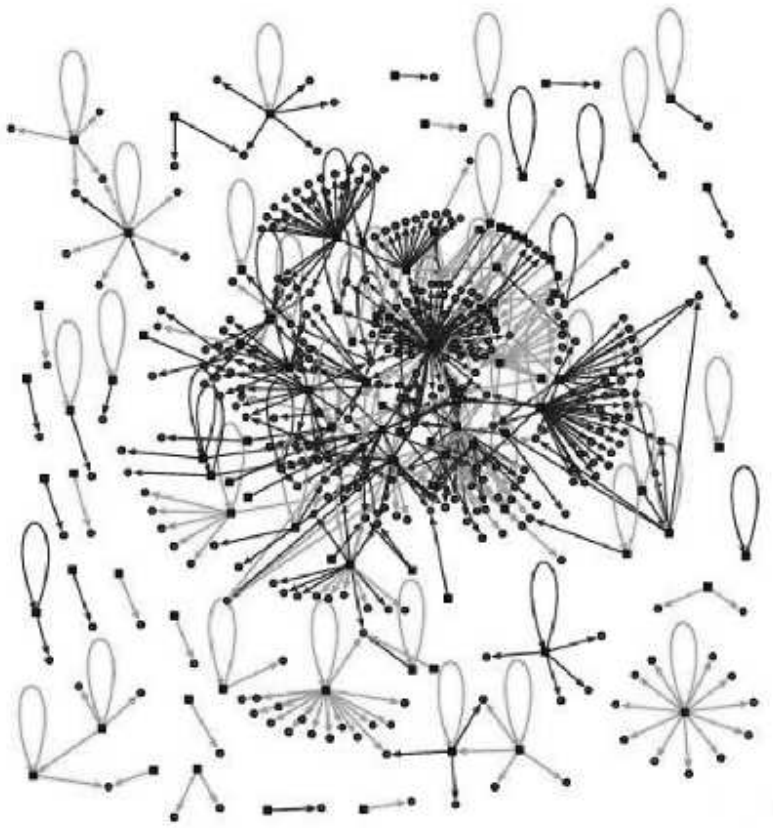


Figure 1.3: Graphical representation of the *E. coli* transcriptional regulatory network. Nodes represent transcription factors (TF), or regulated operons. Links refer to the activating or repressing function of the TFs. Figure taken from [3].

connecting it to other nodes (supposing undirected edges). Collecting the number of nodes with the same degree yields the degree distribution $P(k)$, which represents the probability that an arbitrary node has exactly k edges. If the network is *directed*, *i.e.* the edges are directional, one can define an incoming degree distribution $P_{\text{in}}(k)$ and an outgoing degree distribution $P_{\text{out}}(k)$. For a random network, where each pair of nodes has the same probability of being connected, $P(k)$ decays exponentially and therefore the number of nodes with high k is very low.

The gene regulating network of *E. coli* is directed: focussing on nodes representing transcription factors, the distribution $P_{\text{out}}(k)$ decays exponentially: most genes are regulated by one to three transcription factors. However, $P_{\text{in}}(k)$ decays instead with a power law: $P_{\text{in}}(k) \sim k^{-\alpha}$, with $2 < \alpha < 3$ [2]. Such networks are highly non-uniform: most of the nodes have only a few links, while a few nodes (called hubs) have a large number of links and hold the network together. In the case of *E. coli*, this means that there are a few proteins that regulate a large number of genes.

When $P(k)$ decays with a power law, the variance of $P(k)$ is infinite, hence the network does not have a typical node: such networks are called *scale-free*. Scale-free networks arise spontaneously when new nodes are preferably attached to other nodes with a high number of links. It has been speculated that, in the case of genetic networks, this might

have arisen from gene duplication during evolution [2]. In scale-free networks, the shortest path between two nodes (that is, the path composed by the smallest number of edges between the two nodes) increases much more slowly with the total number of nodes in the network as compared to randomly connected networks [4, 5]. Therefore, in scale-free networks it is very easy to "travel" between nodes (an effect often called *small world*) and local perturbations spread quickly. Scale-free networks are much less vulnerable to failures of random nodes than random networks, since most probably the failure will involve a relatively unimportant low-degree node. Yet their reliance on hubs leaves them prone to targeted attacks.

Cellular functions are carried out in a modular manner [6]. A module refers to a group of physically or functionally linked molecules that work together to achieve a distinct function. The signature of modularity can be traced even for a genetic network, where a module can be identified by the fact that a subset of nodes is connected in a specific wiring diagram. A complex network is likely to display all sorts of distinct subgraphs, from triangles to squares, to higher-order figures. However, recent work indicates that transcription networks contain a small set of recurring regulation patterns, called *motifs* [7, 8]. Such patterns occur much more often in a transcription network than would be expected in a random network. Investigating these motifs will reveal whether their frequent presence could be related to a functional role.

The transcription regulatory network of *E. coli* shows several motifs composed of a few nodes. The topology of these nodes is usually quite simple, and every motif can carry out an elementary operation. Complex behaviours can then be achieved by wiring simple components together, like in an electronic circuit. A cartoon representing the principal motifs found in the *E. coli* genetic network is shown in Figure 1.4.

The simplest motif found in this network is the simple regulation: a transcription factor activates or represses a gene X. If X encodes for the TF, the mechanism is called autoregulation, as shown in Figure 1.4(a-b). Without any regulation, the concentration of X rises in response to a stimulus and reaches a steady state equal to the ratio of the production and decay rates of X. In the case of negative autoregulation, the protein X represses its own gene: if the promoter for X is strong, this mechanism can be exploited by the cell to have a faster rise in the concentration of the protein, which will stop when it has reached the repression threshold and its production rate starts to decrease. This is the case for the SOS DNA-repair system of *E. coli*, in which the master regulator, LexA, represses its own promoter [9]. Moreover, negative autoregulation can be used to reduce cell-cell variation in protein concentrations. Conversely, in the case of positive autoregulation, a TF enhances its own rate of production. The dynamics of this system shows an initial slow rise of the concentration of X, followed by a quick increase when the enhancement takes off. It can be used to shift the response of a network to shorter timescales. Moreover, it tends to enhance cell-to-cell variability. The effect of positive and negative autoregulation on the response time of a gene X is shown in Figure 1.5.

A more complicated motif involves three genes: a regulator X, which regulates Y, and a gene Z, which is regulated by both X and Y, as it is depicted in Figure 1.4(c-d). This

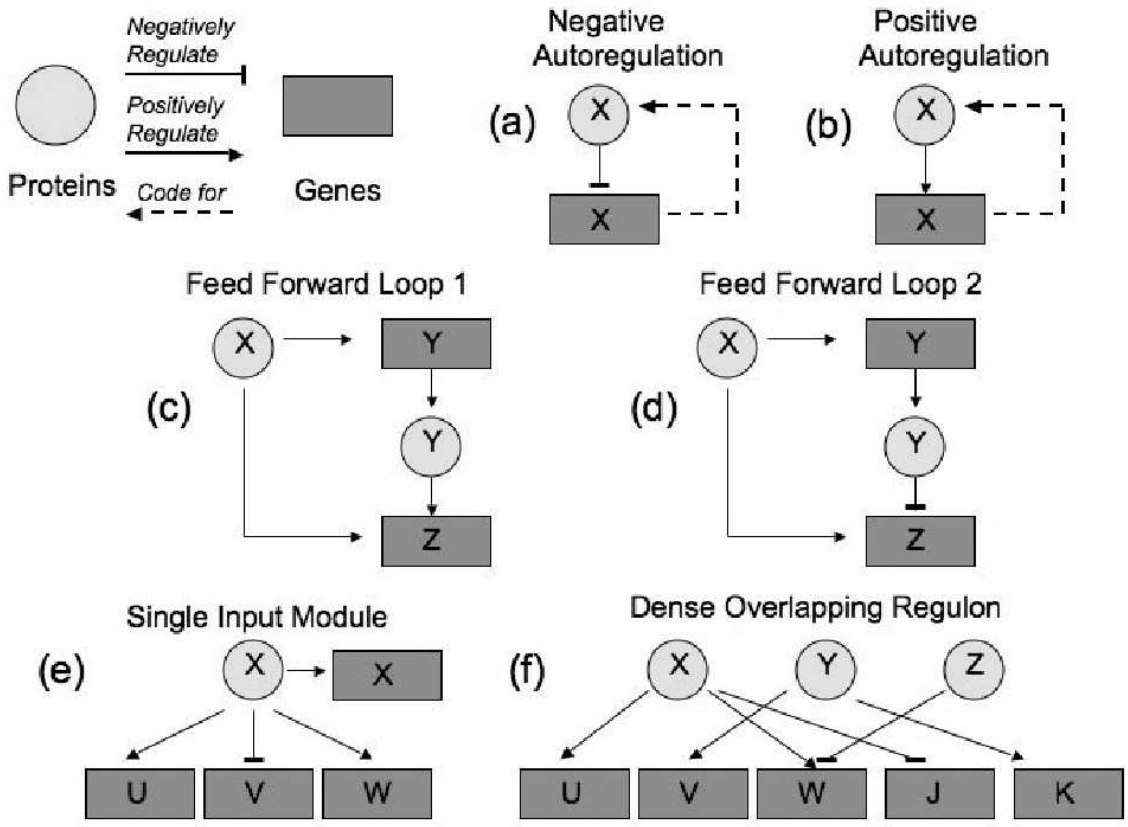


Figure 1.4: Simplest motifs in prokaryotic transcription regulatory networks. In a transcription regulatory network, nodes are either genes or proteins regulating them, while edges can represent coding or regulatory relationships. Many of the X proteins are activated specifically by a signal S (not shown in Figure). Motifs are subnetworks occurring in nature much more frequently than in random networks. (a-b) Single-gene motifs, involving negative or positive autoregulation. (c-d) Feed Forward loops: three-gene motifs used to filter signals or introduce delays. (e) Single Input Module: a single transcription factor controls the expression of several genes, including itself. (f) Dense Overlapping Regulon: a set of regulators that combinatorially control a set of output genes.

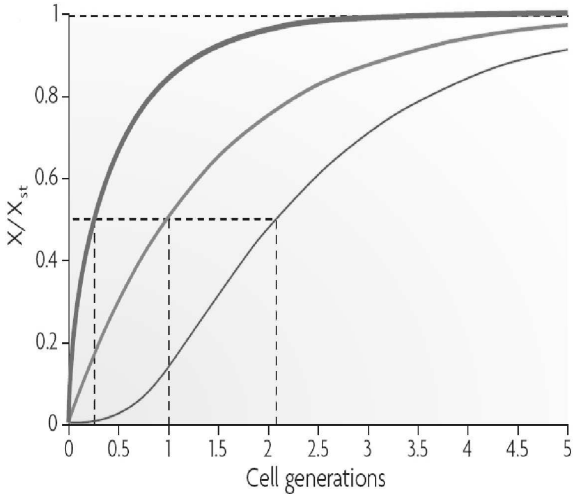


Figure 1.5: Response time comparison for a simply-regulated, negatively autoregulated and positively autoregulated system. Negative autoregulation (top curve) speeds up the response time of a protein X (the time needed to reach halfway to the steady state concentration of X, X_{st}) relative to a simply-regulated system (middle curve) that reaches the same steady state expression. Positive autoregulation (bottom curve) slows down the response time. Figure taken from [10].

motif is called Feed-Forward Loop (FFL), and can exist in 8 different combinations of positive and negative regulation. In the *E. coli* network, two versions occur much more frequently than the others: in the first case (let us call it FFL1, Figure 1.4(c)), both X and Y are transcriptional activators. With FFL1, if the simultaneous presence of X and Y is necessary to activate Z, the system will show a delay in the activation of Z (due to time necessary to accumulate Y), but no delay when the stimulation stops. Such a system can be exploited by the cell to filter brief spurious pulses. This is the case for the arabinose-utilisation system in *E. coli*: a delay of 20 minutes occurs when the input signal cAMP is added to the system [11]. The timescale of the delay is similar to the timescale of random pulses of cAMP in the environment: with FFL1, the network is able to filter them. Conversely, if only the presence of at least one of the two activators is needed at Z, the delay will not be present in the activation, but only at the end of the stimulation (due to the time necessary switch Y off). This behaviour was experimentally demonstrated in the flagella system in *E. coli*, where a delay is observed when the input signal stopped, but no delay occurred when the input signal appeared [12]. In the second feed-forward type loop (FFL2, Figure 1.4(d)), X activates Y and Z, but Y represses Z. As a result, when X is activated, Z is rapidly produced, falling back down when Y has accumulated. This results in pulse-like dynamics, whose rise is very fast, as it is demonstrated for a synthetic FFL2 [13] constituted by the activator LuxR (X), a GFP reporter (Z) and the λ repressor cI (Y).

More complicated motifs involve a regulator X regulating a group of target genes (Single-Input Module, or SIM), typically including itself (Figure 1.4e). This motif can generate a temporal expression program or block several reactions that occur simultaneously when they are needed, like in the case of the arginine-biosynthesis pathway in *E. coli* [14]. Finally, the most complex motif found in the genetic networks of bacteria involves a set of regulators that combinatorially control a set of output genes (Dense Overlapping Regulons, DOR, Figure 1.4(f)). DORs are used when complex computations need to be carried out. DORs have a single layer of regulation in *E. coli*: there is no other DOR at

the output of a DOR, as the need of rapid response forbids the presence of long regulatory cascades. An example of a DOR in *E. coli* is the set of genes regulated by RpoS that are expressed upon entry into stationary growth phase [15].

The large-scale analysis of transcription networks highlights the modularity of such networks and the involvement of small subgraphs that occur frequently and can carry out independent functional tasks. The top-down approach presented so far to characterise these networks can be complemented by a detailed analysis of some of the recurrent fundamental motifs. A similar attempt cannot be extended to the whole network because of the intrinsic complexity of the problem. Moreover, high-throughput techniques can not easily provide quantitative details of the interaction between proteins, or between proteins and DNA: in general quantities such as binding affinities, reaction rates, diffusion constants, etc. must be obtained by specific, laborious procedures. In the rest of this Thesis, I will zoom in on some statistically significant building blocks of transcription networks, and analyse them individually.

1.4 Fluctuations in Gene Expression

Cells deriving from the same common bacterial ancestor are genetically largely identical. They can nonetheless display notable differences in their phenotypes. These variations can be traced back to stochastic fluctuations to which cells are subject. Such fluctuations can originate in the process of gene expression, or they can derive from variations in the external environment.

Recent advances in genetic engineering have made it possible to build small synthetic gene networks in bacterial cells. These networks are typically much simpler than those naturally occurring in the cell, and they can therefore be used to investigate the influence of stochastic effects.

One of the first networks to be constructed was composed of three genes, positioned in a small piece of circular DNA (a *plasmid*), and inserted into *E. coli* cells [16]. Each gene represses another gene, in a circular arrangement, as illustrated in Figure 1.6a. The first gene, *tetR* is regulated by the protein lacI; the protein tetR represses the *cI* gene, which can in turn down-regulate the *lacI* gene. The tetR protein repressed also the expression of a green fluorescent protein (GFP, encoded in a reporter plasmid, Figure 1.6b), whose amount in the cell was directly related to the amount of TetR. The concentration of GFP was monitored by fluorescent microscopy. A theoretical analysis predicted oscillations in TetR, with a period of several hours, which were indeed observed. However, when sibling cells were monitored, they had the same amount of TetR at division, but they lost synchronisation within a few hours (Figure 1.6b).

The emergence of these oscillations, and in general the functioning of every genetic network, relies on protein-protein and protein-DNA interactions. Chemical reactions are stochastic in nature, and fluctuations in the number of molecules of a chemical species,

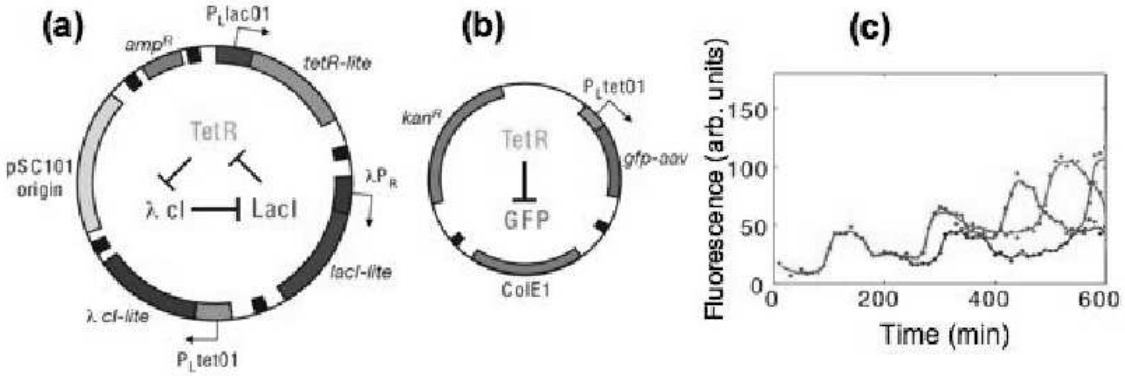


Figure 1.6: The repressillator system [16]. (a) Three genes, each repressing another one in a circular arrangement, are engineered on a plasmid and inserted into a bacterial cell. The stability of the three repressors is reduced by the presence of destruction tags, and they are called "lite". (b) The reporter plasmid contains the gene for a green fluorescent protein (GFP), which is repressed by TetR. GFP concentration can be monitored via fluorescence microscopy. (c) TetR shows an oscillatory behaviour. However, synchronisation between sibling cells is lost after a few hours due to fluctuations in gene expression.

or in their position, can introduce elements of variability in every cell. When only a few copies of a molecular species exist in a cell (as it is often the case for mRNAs or transcription factors), stochastic effects can become prominent. Gene expression is then a *noisy* process, and genetic networks are *noisy* devices. However, some key decisions that must be taken by the cell rely on these networks. Therefore, cells must have adapted to deal with the noise, and perhaps to exploit it as a source of variation.

Fluctuations in genetic networks can have a *temporal* and/or a *spatial* origin, both dramatically increasing when chemical species are found in a low copy number.

Temporal fluctuations are related to the stochastic behaviour of the chemical reactions governing gene expression: the times between subsequent reactions are not regularly spaced, but rather follow a particular probability distribution, depending on the parameters of the system. Assuming a constant probability per unit time for first-order processes, such as dissociations of oligomers or associated complexes, a Poissonian statistics of the reaction times is obtained. The case of second-order reactions is different, because reactants must first find each other in space: the distribution of reaction times can be considered poissonian only if the concentration of both chemical species is high and the system is well-stirred. Low concentrations of reactants can lead to substantial variations in this distribution.

Spatial fluctuations are related to the erratic behaviour of molecules, can produce strong variations in local concentrations, and give rise to effects like fast rebinding of dissociated species, or formation of spatial patterns. They can become predominant when

molecules move slowly or when the system is far form well-mixed. Prokaryotic cells are less compartmented than eukaryotic, and molecules move primarily by free diffusion: spatial fluctuations are therefore particularly important in these systems. In Chapter 6, it will be shown that spatial fluctuations are the dominant source of noise for a genetic network, representing a repressed gene when the concentration of the repressor is below 50nM.

The fluctuations in the biochemical reactions leading to the production of a protein have an *intrinsic* source: they introduce variations in protein levels even in a population of cells with identical genotype and concentrations and states of cellular components. However, noise in protein levels can also derive from fluctuations in the amount or activity of molecules involved in the expression of a gene, like RNA polymerase or ribosomes. These fluctuations depend on the particular state of individual cells, and are sources of *extrinsic* noise.

An elegant experiment [17] allowed the decomposition of noise in protein production into these two internal and external contributions: a plasmid with two different genes under the control of identical promoters was inserted in *E. coli* cells. The two genes encoded fluorescent proteins of different colour, which allowed simultaneous detection of their concentration. In the absence of intrinsic noise, the abundance of the two proteins should be perfectly correlated, and points should align on the diagonal of a scatter plot: when the level of the first protein increases or decreases by a certain amount, the level of the second follows the same pattern (Figure 1.7A). The results of a numerical simulation representing this situation are collected in the upper panels of Figure 1.7. However, since the biochemical steps in the expression of the two genes are independent, gene-intrinsic noise causes the number of expressed proteins to differ, spreading off of the diagonal in the scatter plot, as depicted in the lower panels of Figure 1.7. The spread along the diagonal represents then the extrinsic contribution to the noise, while the spread on the orthogonal direction measures the intrinsic noise. This calculation, reproducing the experiment of Ref. [17], shows that both sources of noise contribute to variation within cells, although the extrinsic part generally dominates [18, 19]. In eukaryotic cells, the slow remodelling of the packed structure of DNA can turn on or off genes by exposing or concealing the genes, adding another source of extrinsic noise to gene expression [20].

The production of proteins depends on a cascade of processes, starting with the binding of an RNA polymerase to a gene promoter, and ending with the folding of the protein into its active state. Stochastic effects can be manifested in each step and be propagated to the final product. Recently, the propagation of noise along this cascade has been subject of theoretical studies.

For prokaryotic cells, transcription generally is the dominant source of noise in protein levels, as demonstrated by monitoring the fluctuations of reporters in *Bacillus subtilis* cells for several transcriptional and translational efficiencies [21]. As the lifetime of mRNA is much shorter than the lifetime of proteins, a cell would like to fully exploit an mRNA

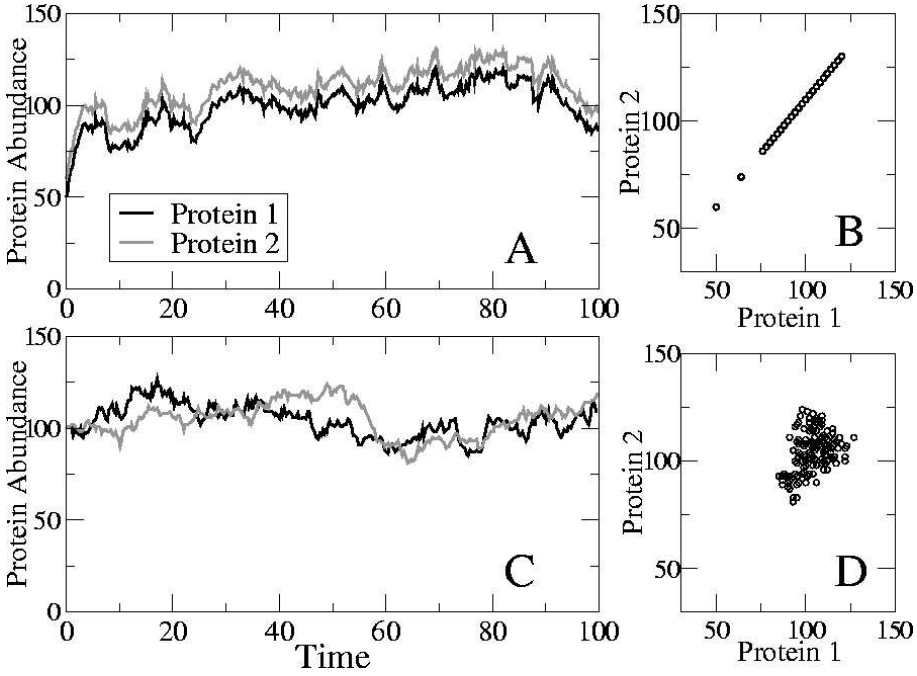


Figure 1.7: Computational equivalent of the two-protein assay experiment [17]. Protein dynamics was simulated by generating stochastic trajectories according to the Chemical Master Equation of the system. (A) If noise in gene expression comes only from extrinsic sources, the stochastic expression of two proteins under the control of identical promoters in a single cell is correlated and on a scatter plot points align on the diagonal (B). (C) As the biochemical steps in the expression of the two genes are independent, gene-intrinsic noise causes the number of proteins to differ (protein productions were allowed to independently fluctuate around the same mean level), giving rise to a scatter plot containing off-diagonal points (D).

molecule and produce as many copies of a protein as possible from a single transcript. However, a high translation rate results in large bursts of protein production, which introduces a large noise. Therefore, in order to keep the protein noise low, a prokaryotic cell needs to have a high transcription rate (to produce many mRNAs molecules), followed by a low translation frequency. Yet, this strategy has a high energetic cost, and can be used only for a few key genes. An example is the *cya* gene in *E. coli*, which codes for Adenylate cyclase. Adenylate cyclase is involved in the regulation of many genes via its enzymatic product cAMP, and its expression display a very low translation frequency.

In Chapter 6, I consider a gene under the control of a repressor, and investigate how noise propagates through all the substeps that lead to production of the protein product.

The situation is similar to gene expression in the eukaryote *Saccaromyces cerevisiae*: transcription is mainly responsible for the noise in protein products and for their possible bursty production [22]. However, in eukaryotes, slow promoter fluctuations due to the

remodelling of chromatin, followed by efficient transcription can reduce the noise.

Genetic networks are then noisy devices working in noisy environments. They must however be able to carry out the specific task they are designed for in a wide range of conditions, *i.e.* they must be robust. The robustness property was investigated in the case of the chemotactic network of *E. coli*: the cascade of reactions that allows the bacterium to respond to changes in the food distribution and swim towards food sources. A computational model [23] and subsequent experiments [24] showed that the network is able to perform even when a large number parameters such as reaction rates or protein concentrations were strongly varied within the physiological range. Thanks to the robustness of the chemotactic network, the bacterium is able to cope with many different environmental conditions and survive in a wide range of situations. Robustness could then have arisen by evolution: a highly optimized, fine-tuned network would not be the most advantageous strategy in a noisy environment.

As it is widely known in the field of engineering, negative feedback can be used to operate robustly in an uncertain environment [25]. It is mentioned in Section 1.3 that negative autoregulation can lower the noise in gene expression. In a recent experiment [26] the distribution of fluorescence of a TetR-EGFP fusion protein was found to become much narrower when the gene is under the control of a negatively autoregulated promoter. About 40% of the genes of *E. coli* are negatively autoregulated, suggesting that this mechanism of noise-reduction is widely utilised by prokaryotic cells.

Noise can also be exploited and amplified by cells to create heterogeneity in a population. When bacteriophage λ infects a bacterium, the decision to commit to either of two alternative pathways (*i.e.* the lytic or the lysogenic states, described in detail in Chapter 4) depends on the level of two proteins, cI and cro. By means of a stochastic fluctuation, one of the two proteins may reach a concentration sufficient to repress the expression of the other, and establish a stable epigenetic state [27].

Positive feedbacks are typical devices that can be used to generate a bistable behaviour by amplifying the noise and partition a genetically identical population into different phenotypes. This effect can be understood by thinking of networks showing a steep response curve, as in Figure 1.8: an increased variability of the input signal in the region of high sensitivity can cause a transition from a unimodal to bimodal population distribution.

Exploiting stochasticity to populate multiple steady states may play an important role in differentiation in multicellular organisms. In the fruit fly, *Drosophila melanogaster*, a GFP reporter subject to repeated rounds of stochastic activation and inactivation of gene expression resulted in patches of fluorescent cells [28].

Microbial cells monitor very specifically their environment in order to adapt efficiently to sudden changes in environmental conditions. Fluctuations in gene expression can provide a mechanism for sampling distinct physiological states and therefore increase the probability of surviving during times of stress, without the need for a genetic mutation. Heterogeneous bacterial populations of isogenic cells might achieve higher survival prob-

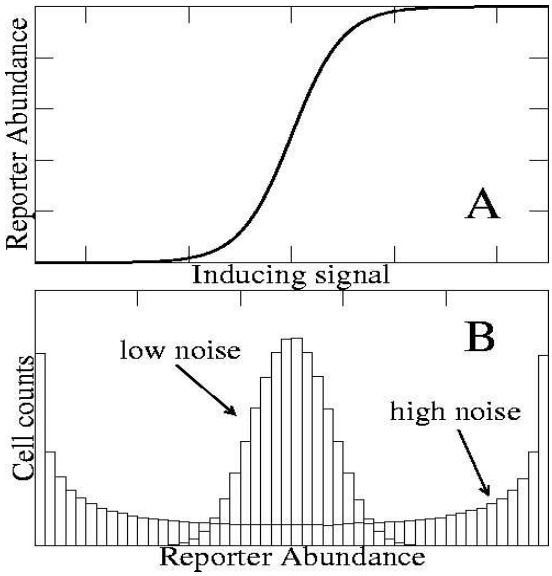


Figure 1.8: Positive feedback can amplify noise and generate a bistable behaviour in a cell population. (A) The response curve of a generic reporter gene to an inducing signal displays the highest sensitivity for intermediate values of the signal, where the slope is highest. (B) When fluctuations in the incoming signal around its mean values are moderate, the response curve yields a single-peaked distribution of reporter genes. Positive feedback in the signal can increase the amplitude of stochastic fluctuations, leading to a bistable distribution of reporters.

abilities than a homogeneous population [29]. Heterogeneity can also be exploited by bacteria to survive antibiotic treatment: while most of the bacterial population will be quickly killed, a genetically identical minority may persist because of a dormant state determined by an alternative gene expression pattern.

Recent measurements of gene expression in single *E. coli* cells over long time periods have provided insights into the relative amplitude and relevant time scales of intrinsic and extrinsic noise [18]. These results confirm that the dominating contribution to noise comes from extrinsic factors, and provides a dynamical explanation of this effect. The autocorrelation time of the extrinsic noise is in fact about 40 minutes, much longer than the autocorrelation time of intrinsic fluctuations ($\leq$ 10 min, consistent with the rapid fluctuations of mRNA numbers). Therefore, perturbations due to extrinsic sources can accumulate over a cell cycle, and be transmitted to daughter cells, affecting their phenotypes.

The frequency components of the noise in protein production were determined by monitoring green fluorescence proteins in *E. coli* cells [30]. Intriguingly, when the protein negatively autoregulates its own expression, the noise frequency distribution broadens, and shifts to higher frequencies. Autoregulation frequency response is limited by protein decay and dilution, and therefore has a larger effects on slower fluctuations than on faster fluctuations. This results in a remodeling of the frequency range of the noise, pushing the time scale of extrinsic noise towards faster dynamics. Fast fluctuations can be more easily filtered out by downstream gene circuits in a regulatory cascade, and therefore have a smaller regulatory impact.

In Chapter 6, I perform a frequency analysis on the noise propagated in a genetic network, aimed to repress a gene, and I show that high-frequency noise, due to fluctuations in promoter occupancy, is filtered out by the slow protein dynamics.

1.5 Models

Models and experiments on biochemical networks benefit from concepts used in several branches of science: physics, (bio)chemistry, biology, mathematics and engineering. An interdisciplinary approach is needed when one wants to analyse a biological system trying to recognize its universal features beyond the specific details of components. This approach is typical of *systems biology* which brings together scientists coming from different backgrounds.

In this Section, I will briefly review the most common mathematical descriptions of genetic networks; computational techniques will be described in the next Section.

1.5.1 The Macroscopic Rate Equation

Biochemical networks are traditionally described in terms of kinetic rates that describe how the concentrations of the various species change with time. The case of a simple *birth-and-death* process, where a species A is created with rate k and degraded with rate μ will be used as an illustrative example:

$$\emptyset \xrightarrow{k} A \xrightarrow{\mu} \emptyset. \tag{1.1}$$

The macroscopic rate equation for the process is the following ordinary differential equation (ODE)

$$\frac{dA(t)}{dt} = k - \mu A(t), \tag{1.2}$$

and its solution is a simple exponential relaxation to the steady state value $A_\infty = k/\mu$:

$$A(t) = A(0)e^{-\mu t} + \frac{k}{\mu}(1 - e^{-\mu t}). \tag{1.3}$$

This formulation of the problem is *deterministic*: if the starting conditions are fixed, the future evolution of the system is also precisely fixed. In the case of the rate equations described in Section 1.5.1, the set of ODE can be solved numerically by standard integration techniques, even for complex networks. This way of modelling biochemical networks assumes that the cell is well-mixed and homogeneous, and the number of copies of A is high enough to justify the use of a continuous variable for the chemical species. Temporal or spatial fluctuations are therefore not taken into account (see Section 1.4). A linear stability analysis can be performed on rate equations to identify the number of steady states of complicated reaction networks, and their stability.

1.5.2 The Chemical Langevin Equation

Molecular fluctuations can be incorporated explicitly by including random variables in the macroscopic model. The easiest approach is to append a noise term to the rate equation.

Eq. (1.2) becomes then:

$$\frac{dA(t)}{dt} = k - \mu A(t) + \xi(t), \tag{1.4}$$

where $\xi(t)$ is a stochastic variable. Typically, $\xi(t)$ is assumed to be a white noise term: $\langle \xi(t) \rangle = 0$, $\langle \xi(t)\xi(t') \rangle = q\delta(t-t')$. Eq. (1.4) is a stochastic differential equation (SDE), often referred to as a chemical Langevin equation. In this framework, species concentrations are now fluctuating stochastic variables; it is possible to reformulate a Langevin equation to an equivalent form [31], called Fokker-Planck equation, that describes the time evolution of the probability density of a chemical species A. The Fokker-Planck equation is a partial differential equation (PDE), whose general form is:

$$\frac{\partial p(A,t)}{\partial t} = \frac{\partial}{\partial A}[c_1(A,t)p(A,t)] + \frac{1}{2}\frac{\partial^2}{\partial A^2}[c_2(A,t)p(A,t)]. \tag{1.5}$$

However, for systems involving more than a few species, it is impossible to solve the Fokker-Planck equation, even numerically. An alternative possibility is to generate many trajectories of the system with the Langevin equation, and use their statistics to estimate the probability density function at a given time t. Implicit in the Langevin and Fokker-Planck approach is the continuous description of molecular species. However, when the copy number of some species is very low, the discreteness can give rise to dynamical states that cannot be captured by a continuous model [32, 33].

1.5.3 The Chemical Master Equation

Genetic networks often involve chemical species present is very low copy numbers. One can then model every single reaction event in the framework of probability theory, and adopt a discrete, particle-based, event-driven approach. Every chemical species has then a probability per unit amount of time of undergoing a certain reaction. The equation that describes how the probability $p(n_A,t)$ of having n_A molecules of a species A varies in time is called the chemical Master Equation (ME). The master equation for the system described in Eq. (1.1) is:

$$\begin{aligned} \frac{\partial p(n_A,t)}{\partial t} &= -\left(k\,p(n_A,t) - \mu n_A\, p(n_A,t)\right) + \\ &\quad k\,p(n_A-1,t) + \mu\,(n_A+1)\,p(n_A+1,t). \end{aligned} \tag{1.6}$$

The master equation is linear, the moments of the distribution $p(n_A,t)$ can be obtained in steady state by using the moment generating functions [31].

In order to quantitatively characterise the fluctuations of a chemical species, a concise parameter is introduced: the *noise* coefficient η_X^2 relative to a species X, defined as the variance of X over its squared mean: $\eta_X^2 = (\langle X^2 \rangle - \langle X \rangle^2)/\langle X \rangle^2$. η^2 can be computed and measured in experiments; however, it does not give any information about the frequency

distribution of the fluctuations in X. Spectral techniques provide instead a more detailed insight into the stochastic behaviour of a species, and can be used to understand more deeply how a genetic network processes the noise.

For the simple reaction (1.1), the moments of the steady state solution $p_\infty(n_A)$ of the master equation (1.6) can be obtained analytically:

$$\langle n_A \rangle = k/\mu, \qquad \langle n_A^2 \rangle = k^2/\mu^2 + k/\mu. \tag{1.7}$$

The noise is easily computed:

$$\eta_A^2 = \mu/k = 1/\langle n_A \rangle. \tag{1.8}$$

Intuitively, the higher the copy number of species A in steady state, the lower its noise. As production and degradation of A are independent events, η_A^2 shows the $1/N$ dependence typical of Poissonian processes. The power spectrum $S_A(\omega)$ of the noise can be obtained by calculating the Fourier transform of the autocorrelation function of A:

$$S_A(\omega) = 2\int_0^\infty dt\, \cos(\omega t)\, C_{ii}(t) = \frac{2k}{\mu^2 + \omega^2}, \tag{1.9}$$

where

$$C_{ii}(t) = \langle (A(0) - \langle A \rangle)(A(t) - \langle A \rangle) \rangle . \tag{1.10}$$

The noise and the power spectrum of a fluctuating signal are linked by the following relation: $\eta_A^2 = (2\pi)^{-1} \int_{-\infty}^{\infty} d\omega S_A(\omega)$. $S_A(\omega)$ is shown in the Inset of Figure 1.9: and displays, on a logarithmic scale, a corner frequency, representing the time scale on which fluctuations relax back to steady state. In this case, the corner frequency is just the decay rate of the protein: $\omega_{\text{corner}} = \mu$.

Fluctuations in the number of protein molecules can be strongly influenced by fluctuations in the low number of mRNA molecules and their power spectra were recently measured [30]. To spectrally analyse the propagation of noise in gene expression, a more complicated model of protein production is introduced: a protein P is randomly produced from an mRNA M, which is stochastically created and removed according to a birth-and-death process. Its fluctuations are therefore an extrinsic source of noise to protein production. The reaction scheme is the following:

$$\emptyset \xrightarrow{k_M} M \xrightarrow{\mu_M} \emptyset, \qquad M \xrightarrow{k_P} M + P, \qquad P \xrightarrow{\mu_P} \emptyset. \tag{1.11}$$

In this case, the average number of P molecules depends on the average number of proteins produced during the lifetime of M, $b = k_P/\mu_M$ [34]:

$$\langle n_M \rangle = \frac{k_M}{\mu_M}, \qquad \langle n_P \rangle = \frac{k_P}{\mu_P} \langle n_M \rangle = b\,\frac{k_M}{\mu_P}. \tag{1.12}$$

The noise for the two species can be computed analytically:

$$\eta_M^2 = \frac{\mu_M}{k_M} = \frac{1}{\langle n_M \rangle}, \qquad \eta_P^2 = \frac{1}{\langle n_P \rangle}\left(1 + \frac{k_P}{\mu_M + \mu_P}\right). \tag{1.13}$$

While η_M^2 is typical of a Poisson process, η_P^2 is given by two terms: the first represents the intrinsic protein fluctuations and is similar to the noise for a birth-and-death process (Eq. (1.8)); the second is an extrinsic contribution coming from mRNA fluctuations. The power spectrum of the noise can be computed analytically for the two species [34]:

$$\begin{aligned} S_M(\omega) &= \frac{2k_M}{\mu_M^2 + \omega^2}, \\ S_P(\omega) &= \frac{2k_M}{\mu_M^2 + \omega^2}\frac{k_P^2}{\mu_P^2 + \omega^2} + \frac{2k_P\langle M \rangle}{\mu_P^2 + \omega^2} = S_{\text{ext}}(\omega)g(\omega) + S_{\text{int}}(\omega). \end{aligned} \tag{1.14}$$

Similarly to the noise, the power spectrum of proteins can be decomposed into an intrinsic contribution and an extrinsic term, which is processed by the network with a transfer function $g(\omega)$. As it is presented in Figure 1.9, the power spectrum of P is dominated by extrinsic contributions at low frequencies: the power spectrum of the extrinsic noise is much higher than the power spectrum of intrinsic noise at these frequencies.

1.5.4 Reaction-diffusion systems

Spatial fluctuations can be significant in bimolecular reactions. A network motif can then be modelled as a reaction-diffusion system, where a chemical species i can diffuse with a diffusion coefficient D_i, and react with a partner located in its close proximity. The evolution of this system is described by a reaction-diffusion equation, which correctly accounts for spatial and temporal fluctuations.

In the case of the elementary reaction $A + B \xrightarrow{k} C$, the interaction between particles can be modelled with a potential $U(|\mathbf{r}_A - \mathbf{r}_B|)$, which gives rise to a force $\mathbf{F}(\mathbf{r}) = -\nabla_B U(\mathbf{r})$ acting on B, and an opposite force acting on A. If the particles diffuse with coefficients D_A and D_B respectively, the reaction-diffusion equation can be written as:

$$\begin{aligned} \frac{\partial}{\partial t} p(\mathbf{r}_A, \mathbf{r}_B, t | \mathbf{r}_{A0}, \mathbf{r}_{B0}, t_0) &= [D_A \nabla_A^2 + D_B \nabla_B^2 - D_B \beta \nabla_B \cdot \mathbf{F}(\mathbf{r}) + D_A \beta \nabla_A \cdot \mathbf{F}(\mathbf{r})] \\ &\quad p(\mathbf{r}_A, \mathbf{r}_B, t | \mathbf{r}_{A0}, \mathbf{r}_{B0}, t_0), \end{aligned} \tag{1.15}$$

This equation can be solved exactly for pairs of particles, but not in the case of a many-body interaction. Therefore, an analytical solution for the spatial fluctuations is not feasible.

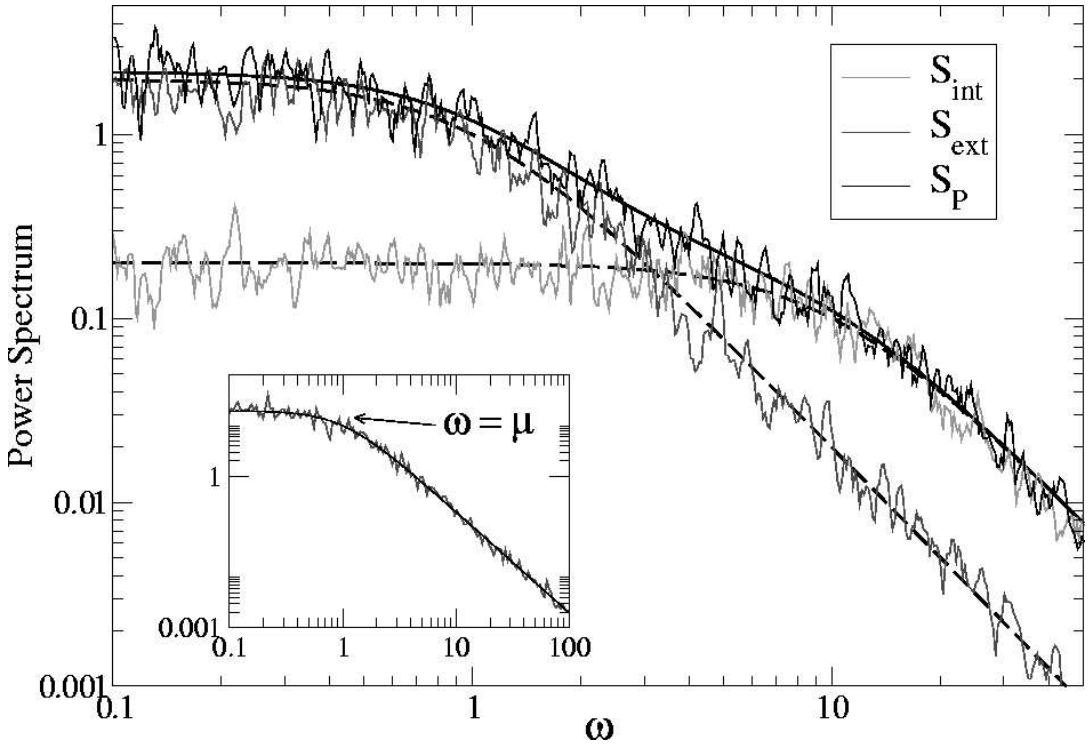


Figure 1.9: Power spectrum of the model (1.11), for $k_M = \mu_M = 1$, $k_P = \mu_P = 10$. Computational data fluctuate around the analytical curves (thicker lines). The power spectrum for the noise in protein production S_p is given by the combination of an intrinsic contribution and an extrinsic term, due to fluctuations in the number of mRNAs. The extrinsic term increases the spectrum at low frequencies, and is therefore responsible for the largest contribution to the noise. The inset represents the power spectrum of a simple birth and death process (1.1), for $k = 10$ and $\mu = 1$, and it displays a corner frequency (the inverse of the decorrelation time) at $\omega_{\text{corner}} = \mu$.

1.6 Computational Techniques

Computational techniques are ideal instruments to investigate the stochastic behaviour of network motifs: they allow us to express the structure and composition of a network formally, and to explore its dynamical behaviour. Numerical techniques can be used to test and generate hypotheses about the fundamental operating principle of a network and the sources and consequences of intracellular noise. Such models are occasionally in a cross-talk with experiments: they often incorporate experimentally-measured parameters, and can unravel new effects suggesting new experiments.

1.6.1 The Stochastic Simulation Algorithm

In general, when one models a system at the level of the master equation (Section 1.5.3), one rarely works directly with expressions like Eq. (1.6), because a huge number of equations are necessary to model systems involving more than a few reactions of species; analytical solutions are then not feasible. Alternatively, one can simulate the random evolution of the system using a Monte Carlo approach. A Stochastic Simulation Algorithm (SSA) was applied by Gillespie in the field of biochemical networks [35] and it is now widely used in the literature.

The SSA, or "Gillespie algorithm" consists of an event-driven kinetic Monte Carlo scheme [36] applied to a set of chemical reactions. The algorithm evolves the system in time consistently with the corresponding master equation. All the chemical reactions in the system must be specified, together with their rate constants. It is assumed that all reactions are Poisson processes: the probability that a reaction i occurs at time $t+\tau$, given that the previous reaction occurred at time t, is

$$P_i(t+\tau)d\tau = a_i(t)\exp\left\{-\tau\sum_{j=1}^{N_r} a_j(t)\right\}, \tag{1.16}$$

where N_r is the total number of reactions in the system and $a_i(t)$ is the "propensity" of reaction i at time t. For a zeroth order reaction like $\emptyset \rightarrow A$, $a_i = k_i V$, where V is the volume of the system, assumed here to be constant, and k_i is the rate constant per unit volume of reaction i. For a first-order reaction like $A \rightarrow B$, $a_i = k_i n_A$, *i.e.* the propensity depends linearly on the number of molecules of the reacting species. In the case of a second-order reaction involving different species,like $A + B \rightarrow C$, $a_i = k_i n_A n_B / V$, while for a second order reaction involving identical species, such as the dimerisation reaction $2A \rightarrow A_2$, the propensity depends on the number of possible reaction pairs: $a_i = k_i n_A (n_A - 1)/V$. One could include a factor of two in the last equation. Here, we choose not to use this factor of two.

At each simulation step, one calculates the propensity of all the reactions, based on the current state of the system. The probability that the next reaction happens between time $t+\tau$ and $t+\tau+d\tau$, given that no reactions occurred between t and $t+\tau$ is

$$P(t+\tau)d\tau = \sum_{i=0}^{N_r} P_i(t+\tau)d\tau = \sum_{j=1}^{N_r} a_j(t)\exp\left\{-\tau\sum_{j=1}^{N_r} a_j(t)\right\}, \tag{1.17}$$

while the probability that the ith reaction occurs at time $t+\tau$ is

$$P_i = \frac{a_i(t)}{\sum_{j=1}^{N_r} a_j(t)}. \tag{1.18}$$

Using two random numbers, the next reaction time $t+\tau$ is chosen from the distribution (1.17) and a reaction s is chosen with probability (1.18). The simulation time is then advanced to $t+\tau$ and the numbers of molecules of all species are updated according to the

stoichiometry of reaction s.

When the number of reactions is high, this procedure is time consuming: although a few strategies have been proposed to increase the efficiency [37], there exist currently no satisfactory approach for simulating processes across multiple time scales. In Chapter 3, I design a general coarse-graining strategy with the aim to integrate out physiologically relevant fast reactions from a biochemical reaction set. I then apply our scheme to a biochemical network displaying bistable behaviour (a *genetic switch*).

The master equation, and thus the SSA, correctly accounts for the temporal fluctuations and low copy number of biochemical species and can therefore be used to study propagation of noise in biochemical networks. However, the SSA assumes a well-stirred system, where many non-reactive collisions happen between two reactive events. The distribution of the species is assumed to be homogeneous in the whole cell. This is often not the case, as proteins display a diffusive behaviour in prokaryotic cells, which can strongly enhance local concentrations, especially for low copy numbers. Moreover, spatial effects such as pattern formation cannot be described by SSA, which tells us only *how many* particles are in the system at time t, but not *where* they are.

Several attempts have been made to extend the SSA to the reaction-diffusion system described in Section 1.5.4: in the most common, the space is divided into small volume elements, and the diffusion of species from one volume element to the next is added as extra reactions [38, 39]. However, this method often creates huge reaction lists that need special routines to be efficiently updated. Moreover, in every volume element the homogeneity assumption of the SSA must be obeyed.

The output of a simulation is usually a time track representing the fluctuations of a chemical species. The noise η^2 can easily be computed from such data. It is also possible to numerically compute the Fourier transform of the signal, and then its power spectrum. If signals are not evenly sampled, the usual Fast Fourier Transform techniques cannot be used. In Appendix C a special algorithm used to obtain Fourier Transforms of unevenly-sampled tracks is presented.

The numerical power spectrum of the process (1.1) is displayed in the Inset of Figure 1.9, while the power spectra of the species in the reaction set (1.11) are presented in the main panel: in both cases, the data obtained from numerical simulation fluctuate around the analytical results of Eqs. (1.9) and (1.14).

In the case of more complicated networks, analytical results are rarely available. Moreover, if the intrinsic and the extrinsic contributions to the noise in a chemical species are correlated, the noise addition rule as in Eq. (1.13) is not valid. Noise-propagating and spectral techniques will be applied in Chapter 6 to investigate the dynamics of a gene under the control of a repressor.

1.6.2 Brownian Dynamics

A reaction-diffusion model of a genetic network can be simulated by a Brownian Dynamics (BD) algorithm. With this method, individual particles move in space and experience stochastic fluctuations in their position as a result of the numerous interactions with an implicit solvent. The random walk of particles in space can bring two reaction partners in close proximity: if so, the probability of a reaction is evaluated. Hence, the algorithm correctly accounts for spatial and temporal sources of fluctuations.

Usually, particles are considered to have a finite probability to react when a BD move brings them to overlap. However, such a choice violates the detailed balance rule, and the algorithm is no longer able to reproduce the equilibrium properties of the system. It is not trivial to incorporate reactions in a BD scheme in a manner that correctly obeys detailed balance. In Chapter 5, I design a rigorous reaction-diffusion BD algorithm, and I illustrate the systematic errors introduced by violation of detailed balance.

Even if this rule is perfectly obeyed, BD simulations can be extremely inefficient, as the simulation requires very small time steps to resolve the fast reaction events. A particle can typically move only a small fraction of its radius in such time steps. Therefore, if the system is dilute (as it is often the case in genetic networks), most of the computational time will be spent on simulating the uninteresting diffusion of particles, and only few reaction events will be sampled.

1.6.3 Green's Function Reaction Dynamics

The most commonly used computational methods for studying biochemical networks fail to efficiently model spatial and temporal sources of fluctuations. In this thesis, a novel algorithm, called Green's Function Reaction Dynamics (GFRD) [40, 41] was implemented and used. With GFRD it is possible to efficiently speed up the simulation of reaction-diffusion systems.

The algorithm relies on the analytical solution of the reaction-diffusion equation for the reaction $A+B \rightleftharpoons C$ (1.15). In order to find this solution, the first step is to express the positions of the A and B particles $(\mathbf{r}_A, \mathbf{r}_B)$ as functions of a new set of coordinates $(\mathbf{r}, \mathbf{R})$ defined as:

$$\mathbf{R} = \sqrt{D_B/D_A}\,\mathbf{r}_A + \sqrt{D_A/D_B}\,\mathbf{r}_B, \qquad \mathbf{r} = \mathbf{r}_B - \mathbf{r}_A. \tag{1.19}$$

In this new reference frame, Eq. (1.15) can be decomposed into two independent equations for $\mathbf{R}$ and $\mathbf{r}$:

$$\frac{\partial}{\partial t}\, p_{\mathbf{R}}(\mathbf{R}, t|\mathbf{R}_0, t_0) = (D_A + D_B)\nabla^2_{\mathbf{R}}\, p_{\mathbf{R}}(\mathbf{R}, t|\mathbf{R}_0, t_0), \tag{1.20}$$

$$\frac{\partial}{\partial t}\, p_{\mathbf{r}}(\mathbf{r}, t|\mathbf{r}_0, t_0) = (D_A + D_B)\nabla_{\mathbf{r}} \cdot (\nabla_{\mathbf{r}} - \beta\mathbf{F}(\mathbf{r}))\, p_{\mathbf{r}}(\mathbf{r}, t|\mathbf{r}_0, t_0). \tag{1.21}$$

The equation for $\mathbf{R}$ describes the simple diffusion of this coordinate, with diffusion coefficient $D_A + D_B$. The equation for $\mathbf{r}$ is not trivial, as the reaction is incorporated as a

boundary condition for Eq. (1.21):

$$-j(\sigma,t|\mathbf{r}_0,t_0) \equiv 4\pi\sigma^2 D\left(\frac{\partial}{\partial r} - \mathbf{F}(\mathbf{r})\right) p_{\mathbf{r}}(\mathbf{r},t|\mathbf{r}_0,t_0)|_{|\mathbf{r}|=\sigma} = k p_{\mathbf{r}}(|\mathbf{r}|=\sigma,t|\mathbf{r}_0,t_0), \quad (1.22)$$

where σ is the sum of the radii of the two particles and k is the forward reaction rate. Eq. (1.21), with the reacting boundary condition (1.22) and the usual behaviour at infinity and at initial time

$$p_{\mathbf{r}}(\mathbf{r},t_0|\mathbf{r}_0,t_0) = \delta(\mathbf{r}-\mathbf{r}_0), \qquad p_{\mathbf{r}}(|\mathbf{r}|\to\infty,t|\mathbf{r}_0,t_0) = 0, \quad (1.23)$$

admits an analytical solution in terms of cylindrical Bessel functions.

The analytical solution for the interparticle distance gives then two pieces of information: 1) what is the probability that two particles have reacted at a certain time, and 2) at which relative distance $\mathbf{r}$ the two particles are located if they have not reacted. The first quantity is given by the *survival probability*:

$$S(t|\mathbf{r}_0,t_0) = \int_{|\mathbf{r}|>\sigma} d\mathbf{r}\, p_{\mathbf{r}}(\mathbf{r},t|\mathbf{r}_0,t_0), \quad (1.24)$$

which yields the probability that two particles have *not* reacted at time t, given that they were at a distance $\mathbf{r}_0$ at time t_0. The survival probability is monotonically decreasing in time because of the reactions between A and B. Therefore, the rate at which $S(t)$ changes in time is a measure of the probability that a reaction between A and B happens at time t. This quantity is the next-reaction time distribution:

$$q(t|\mathbf{r}_0,t_0) \equiv -\frac{\partial}{\partial t} S(t|\mathbf{r}_0,t_0). \quad (1.25)$$

The distribution $q(t)$ measures the probability per unit amount of time that a pair, initially separated by $\mathbf{r}_0$, will have a next reaction event at time t.

The dissociation from the bound state C is easy to treat, because it is a first-order process which does not depend on space. Poissonian statistics will be assumed for these events, thus yielding an exponential next-reaction time distribution.

An efficient algorithm to simulate a system composed by *two particles* that diffuse and can react until time t_{sim} is then the following:

1. If the system is in the dissociated state $A+B$, draw a next association time t_{ass} from the distribution $q(t|\mathbf{r}_0,t_0)$, where t_0 is the current simulation time.

   (a) If $t_{\text{ass}} < t_{\text{sim}}$, the next reaction will happen within the simulation time. Particles A and B will then be removed from the system, and a new position for C is obtained from the distribution $p_{\mathbf{R}}(\mathbf{R},t_{\text{ass}}|\mathbf{R}_0,t_0)$. Time is advanced to t_{ass}.

   (b) If $t_{\text{ass}} > t_{\text{sim}}$, the reaction will not happen within simulation time. The system will then be propagated until t_{sim} by drawing new positions for A and B from the probability distributions $p_{\mathbf{R}}(\mathbf{R},t_{\text{sim}}|\mathbf{R}_0,t_0)$ and $p_{\mathbf{r}}(\mathbf{r},t_{\text{sim}}|\mathbf{r}_0,t_0)$.

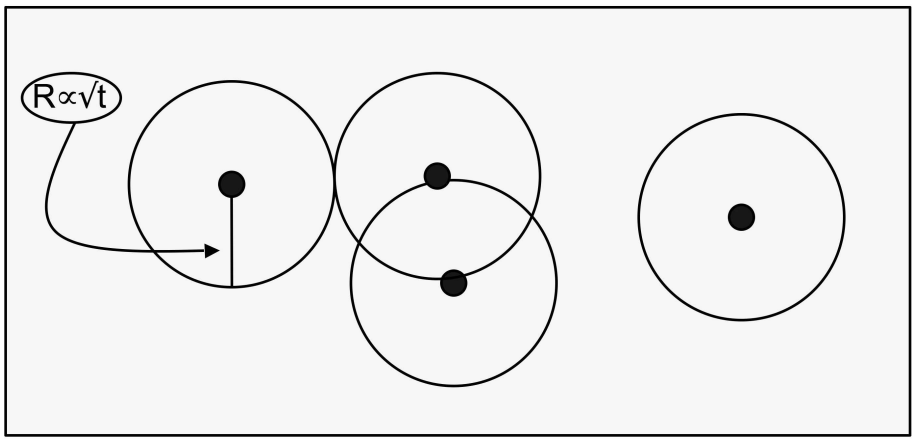


Figure 1.10: The GFRD algorithm automatically reduces a many-body reaction-diffusion system to a sum of one-body and two-body problems. In a time step t_{step}, 99.7% of particles diffusing with coefficient D will remain within a sphere of radius $3\sqrt{6Dt_{\mathrm{step}}}$ centered on the particle. In order to avoid three-body interactions, the maximum time step of the system is determined by the smallest second nearest neighbor distance between particles. In the Figure, the four particles are divided in a pair and two singles. The pair can undergo a reaction, whereas the single particles will be propagated diffusively. The leftmost particle sets the maximum time step.

2. If the system is in the associated state C, draw a next dissociation time t_{diss} from an exponential distribution.

   (a) If $t_{\mathrm{diss}} < t_{\mathrm{sim}}$, the C particle diffused until t_{diss} and then suddenly dissociated. Particle C is then removed from the system, and particles A and B will be positioned at contact, with a random orientation; the position of the center of mass of the system will be drawn from a three-dimensional gaussian distribution with mean $\mathbf{r}_{C0}$ and width $\sqrt{2D_C(t_{\mathrm{diss}} - t_0)}$.

   (b) If $t_{\mathrm{diss}} > t_{\mathrm{sim}}$, the dissociation will not happen within the simulation time. The particle C will freely diffuse until time t_{sim}.

The time step of the algorithm is adaptive, and it can become very large when the particles are far apart. The propagation of particles is yet exact, because it makes use of an exact analytical result. This approach can be extended to a many-body system, provided that it is broken down to a sum of one-body and two-body problems. This is possible by fixing a maximum time step t_{max}, as it can be seen in Figure 1.10. In a time step, particles will stay with a probability of 99.7% within spheres with radius $3\sqrt{6Dt_{\mathrm{step}}}$. A maximum radius for the spheres is set by the smallest second neighbour distance; the maximum radius is translated to a maximum time step t_{max}. Therefore, for a general system of species that can diffuse and react, the GFRD algorithm is the following:

1. Determine $t_{\max}$ as the minimum time that would create a many-body interaction in the system. In this way, the system is partitioned in a set of pairs and single particles and a list of potential reactions is created.

2. Draw for each element of the list of potential reactions a time $t_{\text{react},i}$ from the distributions $q_i(t)$. If the smallest of these times $t_{\text{react},j}$ is smaller than $t_{\max}$, the corresponding reaction will occur and $t_{\text{step}} = t_{\text{react},j}$. Otherwise $t_{\text{step}} = t_{\max}$.

3. Propagate single particles by free diffusion for a time t_{step}. Apply the reaction procedure described above for the selected pair of particles. Propagate the other pairs according to the solution of the diffusion-reaction equation.

4. Update the identities of the particles according to the selected reaction.

The algorithm is event-driven, and it can take very long time steps when particles are far apart and they cannot interact. Conversely, when particles are close together, the time step decreases automatically and resolves the large number of reactive events that may happen. In contrast with the SSA, the spatial fluctuations are rigorously accounted for; with respect to BD, the efficiency is enormously increased, especially for dilute systems. I apply this algorithm in Chapter 6 to study the effect of diffusion of repressors in the expression of a gene.

1.6.4 Forward Flux Sampling

A major challenge in the simulation of genetic networks is the sampling of events that are rare, yet important. A prominent example is given by genetic switches, *i.e.* genetic networks displaying a multistable behaviour. With these networks, a cell can commit to one among several epigenetic states (states that are not directly encoded in the DNA) and maintain it through many generations. The well-known example of the bacteriophage λ bistable genetic switch will be discussed in depth in Chapter 4.

Fluctuations in a genetic switch can cause the system to spontaneously flip from one stable state to another. The switching events are very rare, and, when they happen, occur on time scales much shorter than the mean residence time in the basins of attraction of every stable state. With conventional techniques like the SSA, most CPU time is wasted on simulating the uninteresting waiting times in between the switching events.

A novel method, called Forward Flux Sampling (FFS) [42, 43, 44], has recently been developed for sampling spontaneous transitions between two regions in phase space A and B, and for computing the rate constant for such transitions. A and B are defined by an order parameter λ, such that $\lambda < \lambda_{\text{A}}$ in A and $\lambda > \lambda_{\text{B}}$ in B. A series of nonintersecting surfaces $\lambda_0, \ldots, \lambda_n$ are defined in phase space, such that $\lambda_0 = \lambda_{\text{A}}$ and $\lambda_n = \lambda_{\text{B}}$. Any path from A to B must cross each interface, without reaching λ_{i+1} before λ_i.

The transition rate k_{AB} from A to B is the average flux of trajectories reaching B from A $\bar{\Phi}_{\mathrm{A,n}}$, and it can be decomposed in the following way:

$$k_{\mathrm{AB}} = \bar{\Phi}_{\mathrm{A,n}} = \bar{\Phi}_{\mathrm{A},0} P(\lambda_n|\lambda_0) = \bar{\Phi}_{\mathrm{A},0} \prod_{i=0}^{n-1} P(\lambda_{i+1}|\lambda_i). \tag{1.26}$$

Here, $\bar{\Phi}_{\mathrm{A},0}$ is the average flux of trajectories leaving A in the direction of B and $P(\lambda_n|\lambda_0)$ is the probability that a trajectory that crosses λ_0 in the direction of B will eventually reach B before returning to A. On the right-hand side, $P(\lambda_{i+1}|\lambda_i)$ is the probability that a trajectory which reaches λ_i, having come from A, will reach λ_{i+1} before returning to A. In Eq. (1.26), the flux of trajectories from A to B is split into the flux across the first interface λ_0, multiplied by the probability of getting from that interface to B, without returning to A. This last term is then factorized in a product of conditional probabilities of reaching the next interface (before returning to A), having arrived at a particular interface from A.

The flux $\bar{\Phi}_{\mathrm{A},0}$ is obtained by running a simulation of the system in the "basin of attraction" of A and counting how many times the trajectory in phase space crosses λ_0 coming from A. At the same time, one generates a collection of phase space points that correspond to the moments that the trajectory reached λ_0, moving in the direction of B. This ensemble of points is then used as the starting point for a calculation of $P(\lambda_1|\lambda_0)$. A point from the collection is chosen at random and used to initiate a new trajectory, which is continued until either A or λ_1 is reached. If λ_1 is reached, the trial is designated a "success". This is repeated many times, generating an estimate for $P(\lambda_1|\lambda_0)$ (the number of successes divided by the total number of trials), plus a new collection of points at λ_1 that are the end points of the successful trial runs. This collection of points is used to initiate trial runs to λ_2, generating an estimate for $P(\lambda_2|\lambda_1)$ and a new collection at λ_2, etc. The procedure is schematised in Figure 1.11.

The interfaces are used to drive the system over the barrier. While the efficiency of the method will depend upon the precise positioning of the interfaces and thus upon the choice of the order parameter λ, λ does not have to be the true reaction coordinate: the transition paths are generated according to the underlying dynamics of the system and are free to follow any possible path between state A and B. The choice of a specific λ will not affect the rate constant, nor the ensemble of transition paths.

FFS can be applied to systems out of equilibrium and allows sampling of the trajectories corresponding to the transition (the transition path ensemble) by tracing back to A paths that eventually arrive in B. The trajectories allow an estimation of the steady state probability distribution as a function of the order parameter λ.

The thesis is organized as follows: in Chapter 2, I will analyse a genetic network modelling a simple genetic switch. The fluctuations that cause a spontaneous flip of the switch come from different biochemical processes, such as oligomerization of transcription factors or binding to DNA. I will elucidate which sources of fluctuations are actively exploited by the system, and characterise the switching trajectories.

In Chapter 3, I will study the effect of several dynamical coarse-graining techniques

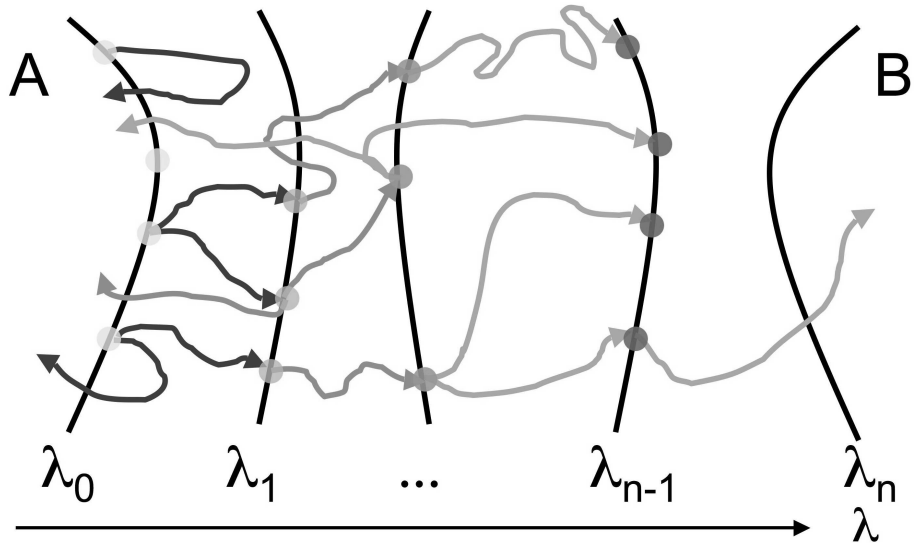


Figure 1.11: Scheme of the Forward Flux Sampling method for sampling rare events in non-equilibrium systems. A parameter λ monotonically increases when the system moves in phase space from the stable state A towards B. A series of interfaces is set between A and B, such that a trajectory from A to B must cross the interface λ_i before λ_{i+1}. A brute-force run in the A basin is used to store a collection of points on the first interface λ_0. From these points, many trial trajectories are fired and stopped when they reach either A or the next interface. The number of successful trajectories over the number of trials yields an estimate of $p(\lambda_1|\lambda_0)$. From the new collection of points on λ_1 the procedure is repeated until B is reached. Partial paths can be connected to obtain switching trajectories from A to B.

on the model switch: I find that one can safely integrate out protein-protein interactions, because they are fast and their fluctuations are not exploited by the system to flip the switch. The coarse-graining procedure must, however, be conducted at the level of the master equation, since methods based on the macroscopic rate equation give incorrect values of the switching rates.

The work carried out in Chapter 2, will then be applied to a realistic model of the bacteriophage λ genetic switch, where dimerisation of transcription factors will be integrated out, in Chapter 4. Using FFS, I will be able to measure a spontaneous switching rate between the two stable states of the system that agrees with experimental predictions. I will then investigate the effects of the crowded cellular environment and of a possible DNA loop that could stabilise one stable state.

As discussed above, Brownian Dynamics algorithms that treat second order reactions naively may introduce systematic errors originating from violations of the detailed balance rule. In Chapter 5, I will design a rigorous reaction-diffusion BD algorithm and extensively test it. As an illustrative example, I apply the algorithm to a network consist-

ing of a substrate under the action of two antagonist enzymes, and evaluate the effects of the spatial fluctuations on the response of the system.

Finally, in Chapter 6, I will apply the GFRD method to study the dynamics of a gene under the control of a repressor. I will design a detailed model of protein production and see that diffusion of repressor molecules is the dominant source of noise in the protein output. The enhancement of noise comes from the immediate rebinding of repressors to the DNA, a purely spatial effect that could not be captured with techniques like the SSA. I will compute the power spectrum of the noise to demonstrate that the slow protein decay filters out the high-frequency fluctuations.

Chapter 2

Barriers and reaction coordinates for the flipping of genetic switches

The things we fear most in organizations –
fluctuations, disturbances, imbalances
are the primary sources of creativity
Margaret Wheatley

We have applied Forward Flux Sampling to elucidate the switching dynamics of a genetic toggle switch, which consists of two genes that mutually repress each other. In analogy with the spontaneous transitions between two stable equilibrium states, the switching rate can be factorised in a kinetic prefactor times the probability of being at the separatrix between the two states. The analysis reveals that the switching rate increases with the rate of dimerisation, while it decreases with the rate of operator binding. These reactions affect the switching rate in a fundamentally different manner: while increasing the rate of dimerisation only increases the kinetic prefactor, increasing the rate of operator binding decreases both the kinetic prefactor and the probability of being at the dividing surface (separatrix). We elucidate these differences computing the paths as a function of the probability the system has to commit the transition. These reveal that varying the rate of operator binding can drastically change the pathway of switching, while changing the rate of dimerisation predominantly changes the speed the transition paths are travelled, but not the location of the transition state ensemble. The implications for the simulation of other rare events in non-equilibrium systems are discussed.

2.1 Introduction

Multistable biochemical networks are omnipresent in living cells. Multistability can provide cellular memory, it can enhance the sharpness of the response to intra- and extracellular signals, it can make the cell robust against biochemical noise, and it allows cells to differentiate into distinct cell types. The steady states of a multistable biochemical network are often very stable and the network typically only switches from one state to the next under the influence of an external signal [45]. A key question, therefore, is what determines the stability of these steady states. To understand the stability of multistable networks, we have to elucidate the pathways of switching from one steady state to the next. These switching events are, however, intrinsically difficult to study experimentally, because the switching event itself can be much faster than the typical life time of the steady state. Computer simulations are a natural tool to study rare switching events. But, precisely because the switching events are rare, special numerical techniques are required. In this Chapter, we show how Forward Flux Sampling, described in Section 1.6.4 and in Refs. [42, 43, 44] can be combined with committor distributions to characterize the pathways of switching and to elucidate the stability of multistable biochemical networks.

If a biochemical network is bistable, with two stable states A and B, respectively, then it will show a bimodal steady-state probability distribution, $P(q)$, of some order parameter q. This order parameter can be the concentration of a species, or a combination of the concentrations of a number of species. It is usually interpreted as a reaction coordinate that measures the progress of the "reaction" from state A to B. Recently, such bimodal distributions have been measured experimentally for biochemical networks [46, 47, 48]. These distributions are potentially useful, because they are linked to the rate of switching from one state to the other. For equilibrium systems, widely studied in the field of soft condensed matter physics, as well as in other branches of physics and chemistry, the transition rate k_{AB} from state A to state B is given by

$$k_{AB} = R\exp\left[-\beta\Delta F^*(q^*)\right] = RP(q^*). \tag{2.1}$$

Here, β is the inverse temperature, q^* denotes the location of the top of the free-energy barrier that separates the two (meta)stable states, $\Delta F^*(q^*)$ is the height of the free-energy barrier, and R is a kinetic prefactor that gives the average flux over the barrier. The above expression shows that the height of the free-energy barrier is related to the probability of finding the system at the top of the barrier, $P(q^*)$. While Eq. (2.1) has mostly been used for equilibrium systems, we have recently shown that an identical relation can also be derived for rare switching events in non-equilibrium systems [49]. This is useful, because, as Eq. (2.1) shows, the rate of switching from one steady state to the next, can be written as the probability of being at the dividing surface, the separatrix [49, 50], times a kinetic prefactor that describes the average flux of trajectories crossing the dividing surface.

However, while the rate k_{AB} is insensitive to the choice of the order parameter q as long as it connects the states A and B, the location of the barrier, q^* (the probability $P(q^*)$ of being at the top of the barrier) and the kinetic prefactor R all depend upon the choice for q. Only if the order parameter q is the true reaction coordinate that accurately describes the

progress of the reaction, does q^* give the location of the transition state, which separates the two steady states. And only then are $P(q^*)$ and R accurate measures for the probability of being at the top of the barrier and the flux over the barrier, respectively. Hence, an important question is: what is the reaction coordinate q that describes the progress of the transition?

To answer this question, we need to study the pathways of switching. However, pathways of switching are intrinsically difficult to study experimentally. The reason is that the average waiting time in between the switching events is typically orders of magnitude longer than the time scale of the event itself. The flipping of biochemical switches is thus both infrequent and fast, and this makes it very difficult to obtain good statistics on these rare events.

Computer simulations are a powerful technique to study the flipping of biochemical switches. However, conventional simulations, using brute-force Kinetic Monte Carlo schemes [36] such as the Gillespie algorithm [35], are highly inefficient, because most of the CPU time is waisted on simulating the uneventful waiting time. In the field of soft-condensed matter physics, a number of numerical techniques have been developed that makes it possible to simulate rare events efficiently. These techniques, however, usually exploit the fact that the system obeys detailed balance and microscopic reversibility. Biochemical networks, however, are always out of equilibrium (as long as the cell is alive), and this means that these techniques cannot be used.

We have recently developed a new class of techniques, called Forward Flux Sampling, which make it possible to efficiently simulate rare events in both equilibrium and non-equilibrium systems [42, 43, 44]. FFS allows us to generate switching pathways and obtain the rate of switching k_{AB}. Moreover, we have recently shown that from a single FFS calculation one can simultaneously determine the steady-state probability distribution $P(q)$ [51]. Using Eq. (2.1), one can then also obtain the kinetic prefactor R.

In this Chapter, we apply FFS to study a toggle switch that consists of two genes that mutually repress each other [52, 53, 54, 55, 49, 42, 56, 57, 58]. We show that both the rate of operator binding and the rate of dimerisation of the gene products can strongly affect the switching rate: while the switching rate decreases with an increasing rate of operator binding, it increases as the rate of dimerisation is increased. In both cases, we vary the forward and backward rate constants keeping their ratio, *i.e.* the equilibrium constant unchanged. Interestingly, varying these rate constants changes the switching rate in a fundamentally different manner. Changing the rate of operator binding can have a profound effect on the mechanism of switching [42, 50]; it thereby changes the switching rate by affecting both the kinetic prefactor and the probability of being at the separatrix. In contrast, changing the rate of dimerisation only has a minor affect on the location of the switching pathways; the mechanism of switching is indeed fairly insensitive to the dimerisation rate. Concomitantly, the probability of being at the separatrix is only marginally affected by changes in the dimerisation rate. However, the rate at which the system crosses the dividing surface, is strongly reduced as the dimerisation rate is decreased. Dimerisation thus changes the switching rate via the kinetic prefactor.

To elucidate the effect of dimerisation and operator binding on the switching rate, we discuss the pathways of switching and show how committor distributions can be used to identify the reaction coordinate [59, 42]. Every configuration x has a commitment probability or "committor", $P_B(x)$. This is the probability that a trajectory, fired at random from that configuration, will reach state B before state A. The transition state ensemble (TSE) is formed by the ensemble of configurations that have a committor of $P_{\mathrm{B}}(x) = 0.5$. After generating transition paths by FFS, the committor values can be computed for the configurations these trajectories visit. Each transition path will cross the dividing surface, the separatrix where $P_{\mathrm{B}} = 0.5$, one or more times, and will therefore generate one or more members of the TSE. The objective is then to find an order parameter, or a combination of order parameters, that characterizes the states of the TSE. To test whether a (combination of) order parameter(s) constitutes the reaction coordinate, one can compute the probability distribution of this (combination of) parameter(s) for the harvested configurations of the TSE [59, 42]. While poorly chosen parameters will exhibit a broad distribution or even a bimodal one [60], the combination of order parameters that describe the reaction coordinate will have a narrow distribution of values in the TSE.

We have applied this analysis to the toggle switch. It reveals that the reaction coordinate does not only involve the difference in the total copy number of the two species, but also the state of the operator [42]. In contrast, we find no evidence that dimerisation is an important reaction coordinate. This explains why the rate of operator binding affects both the probability of being at the separatrix and the kinetic prefactor, while dimerisation only affects the kinetic prefactor.

In the next Section, we first describe the model of the genetic switch. We then present the results on the switching rate, the kinetic prefactor and the probability of being at the separatrix. We also show how they depend upon the rate of operator binding and the rate of dimerisation. To elucidate these dependencies, we then discuss the pathways of switching in the next Section. We end with a discussion of the implications of our findings for the modelling of genetic switches.

2.2 Model: The Exclusive Switch

We consider a genetic toggle switch in which the two genes mutually repress each other [61, 52, 55, 49, 42, 56]. In particular, we study the 'exclusive switch' introduced by one of us [55, 49]. In this system, two transcription factors mutually exclude each other's binding

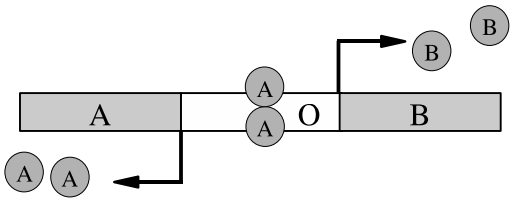


Figure 2.1: Pictorial representation of our model switch, corresponding to Eq. (2.2). Two divergently-transcribed genes are under the control of a shared regulating binding site on the DNA, called the operator. Both proteins can bind, in the homodimer form, to the operator and block the production of the other species.

to the operator. The switch is represented by the following set of reactions [55, 49, 42]:

$$\mathrm{A}+\mathrm{A} \underset{k_\mathrm{b}}{\overset{k_\mathrm{f}}{\rightleftharpoons}} \mathrm{A}_2, \qquad \mathrm{B}+\mathrm{B} \underset{k_\mathrm{b}}{\overset{k_\mathrm{f}}{\rightleftharpoons}} \mathrm{B}_2 \tag{2.2a}$$

$$\mathrm{O}+\mathrm{A}_2 \underset{k_\mathrm{off}}{\overset{k_\mathrm{on}}{\rightleftharpoons}} \mathrm{OA}_2, \qquad \mathrm{O}+\mathrm{B}_2 \underset{k_\mathrm{off}}{\overset{k_\mathrm{on}}{\rightleftharpoons}} \mathrm{OB}_2 \tag{2.2b}$$

$$\mathrm{O} \xrightarrow{k_\mathrm{prod}} \mathrm{O}+\mathrm{A}, \qquad \mathrm{O} \xrightarrow{k_\mathrm{prod}} \mathrm{O}+\mathrm{B} \tag{2.2c}$$

$$\mathrm{OA}_2 \xrightarrow{k_\mathrm{prod}} \mathrm{OA}_2+\mathrm{A}, \qquad \mathrm{OB}_2 \xrightarrow{k_\mathrm{prod}} \mathrm{OB}_2+\mathrm{B} \tag{2.2d}$$

$$\mathrm{A} \xrightarrow{\mu} \emptyset, \qquad \mathrm{B} \xrightarrow{\mu} \emptyset \tag{2.2e}$$

In this reaction scheme, O represents a stretch of DNA containing a regulation site (operator) and two adjacent genes sharing it. These genes code respectively for proteins A and B, transcribed in opposite directions, as shown in Figure 2.1. The promoter O can randomly produce both A and B with the same rate. Each protein can form a homodimer that can bind to the promoter; we note here that while mean-field analysis predicts that cooperative binding of the TFs to the DNA is essential for bistability [61], it has been demonstrated recently that cooperative binding is not critical for bistability when the discrete nature of the reactants is taken into account [56]. When an A(B) dimer is bound to O, the production of B(A) is blocked. Monomers can also decay, which models degradation and dilution in a cell. Clearly, when one species is abundant over the other one, many dimers of the majority species are formed, and the probability of finding one of them bound to O is therefore high. This effect in turn lowers the production of the minority species, leading to a stabilization of the state. If a rare fluctuation, however, is able to build up a substantial number of the minority species, these molecules will in turn dimerise and bind to O, leading to a stochastic flip of the switch. Here, we have assumed that transcription, translation and protein folding can be modelled as a single Poisson process, neglecting the molecular details and substeps that lead to the production of a protein. Ref. [49] discusses the effects of both shot noise and fluctuations in the number of proteins produced per mRNA transcript on the switch stability.

A previous work has demonstrated that the switch stability depends strongly on the

mean copy number of species A and B [55]. In the model described here, the mean copy number is given by the *ratio* of the production and decay rates, *i*.e. k_{prod}/μ. While this ratio is directly related to the switch stability, the individual values of these rate constants only set the time scale of the reactions. The choice of either k_{prod}^{-1} or μ^{-1} as the unit of time is therefore arbitrary. Following the earlier work, we use the faster reaction of the two: k_{prod}^{-1} is thus the unit of time [55]. Choosing the volume of the system, V, as the unit of volume, we define the following baseline set of parameters: $k_{\mathrm{f}} = 5k_{\mathrm{prod}}V$, $k_{\mathrm{b}} = 5k_{\mathrm{prod}}$ (so that $K_{\mathrm{D}}^{d} = k_{\mathrm{b}}/k_{\mathrm{f}} = 1/V$), $k_{\mathrm{on}} = 5k_{\mathrm{prod}}V$, $k_{\mathrm{off}} = k_{\mathrm{prod}}$ (so that $K_{\mathrm{D}}^{b} = k_{\mathrm{off}}/k_{\mathrm{on}} = 1/(5V)$), $\mu = 0.3k_{\mathrm{prod}}$. These numbers are biologically motivated, as we discuss in more detail in the Discussion Section. We assume that the rates for proteins A and B are the same, in order to have a completely symmetric switch.

The flipping of biochemical switches will be studied with the Forward Flux Sampling technique, described in Section 1.6.4: We have used the original FFS scheme to compute the rate constant and to generate members of the transition-path ensemble (TPE) [42, 43, 44]. In contrast to schemes such as Milestoning [62] and PPTIS [63], FFS does not have to make the Markovian assumption that the paths lose memory of where they come from. Moreover, in contrast to these approaches, FFS does not have to assume that the distribution of state points at the interfaces $\{\lambda_0, \cdots, \lambda_n\}$ equals the stationary distribution of states: each point at interface i lies on a path which originates in the initial state A. We show below that this is essential for this analysis of the genetic switch, even though this system is highly diffusive.

2.3 Results

The mean field analysis performed in [49] demonstrates analytically for this system the existence of three fixed points for the parameter values listed above: two symmetrical stable states, one rich in A and the other rich in B, separated by one unstable state where the total number of A equals the total number of B. This implies that the system can be considered as a truly bistable switch. However, while this mean-field analysis indicates the regions in parameter space where the system is bistable, it cannot predict the switch stability and elucidate the switching pathways.

To harvest switching pathways and to calculate the switching rate, we have performed Kinetic Monte Carlo simulations in conjunction with FFS. The order parameter to lay down the FFS interfaces (see Section 1.6.4) is the difference between the total numbers of the two kinds of molecules: $\lambda = n_{\mathrm{A}} + 2n_{\mathrm{A}_2} + 2n_{\mathrm{OA}_2} - (n_{\mathrm{B}} + 2n_{\mathrm{B}_2} + 2n_{\mathrm{OB}_2})$, where n_{X} is the number of copies of species X.

2.3.1 Switching rates

Figures 2.2A and 2.2B show the switching rate as a function of the dimerisation rate and the operator binding rate, respectively. In both cases, the forward and backward rate

constants are scaled such that the equilibrium constant is kept constant. It is seen that while the switching rate increases with the rate of dimerisation (panel A), it decreases with the rate of operator binding (panel B).

FFS not only makes it possible to compute rate constants, it also allows the calculation of stationary distributions, both for equilibrium and non-equilibrium systems [51]. Figure 2.3 shows the steady-state probability distribution $P(\lambda)$ of finding the system at a particular value of the order parameter λ, for different values of the dimerisation and operator binding rate. First of all, it should be noted that for the range of parameters considered here, the distribution is bimodal, as expected for a bistable system. Secondly, it is seen that the location of the basins of attraction are fairly insensitive to both the rate of dimerisation and the rate of operator binding. Also the shape of the stationary distribution does not depend much on these rate constants. However, as inset B of Figure 2.3 and Figure 2.2D show, the relative probability $P(\lambda=0)$ of being at the top of the barrier that separates the two stable states, does depend on the rate of operator binding. Interestingly, the probability of being at the top of the barrier is much less sensitive to the rate of dimerisation (inset B of Figure 2.3 and Figure 2.2C).

Eq. (2.1) makes it possible to derive from the computed rate constant k_{AB} and the calculated relative probability of being at the top of the barrier $P(q^*) = P(\lambda = 0)$, the kinetic prefactor R. Figures 2.2E and 2.2F show the kinetic prefactor as a function of the dimerisation rate and operator binding rate, respectively. It is seen that the kinetic prefactor strongly increases with the rate of dimerisation, while it decreases with the rate of operator binding.

2.3.2 Switching pathways

To elucidate the dependencies of the switching rate, "the barrier height" and the kinetic prefactor on the dimerisation and operator binding rates, we will examine the switching pathways as harvested by FFS. Here, and in what follows, we will focus on three sets of parameters: (1) the base-line set, with an operator binding rate of $k_{\mathrm{on}} = 5k_{\mathrm{prod}}V$ and a dimerisation rate of $k_{\mathrm{f}}=5k_{\mathrm{prod}}V$; (2) a set with slow dimerisation, $k_{\mathrm{f}}=0.1k_{\mathrm{prod}}V$, and $k_{\mathrm{on}}=5k_{\mathrm{prod}}V$; (3) a set with fast operator binding, $k_{\mathrm{on}}=500k_{\mathrm{prod}}V$, $k_{\mathrm{f}}=5k_{\mathrm{prod}}V$. As above, in all cases the backward rates are scaled such that the equilibrium constants remain constant: $K_{\mathrm{D}}^{d}=k_{\mathrm{b}}/k_{\mathrm{f}}=1/V$ and $K_{\mathrm{D}}^{b}=k_{\mathrm{off}}/k_{\mathrm{on}}=1/(5V)$.

To characterize the switching pathways, we need to define a parameter that measures the progress of the transition. One choice would be to take λ, the difference between the total copy numbers of A and B. However, as we will show below, λ is not sufficient to describe the progress of the reaction: the reaction coordinate involves at least one other order parameter. We have therefore also used this other parameter to characterize the progress of the reaction, namely the committor $P_B(x)$. As discussed in the introduction, the committor is the probability that a configuration reaches state B before state A when it is propagated in a random direction. After the transition paths have been harvested with FFS, the committor values can be computed for the configurations that these trajectories visit; for each configuration x, $P_B(x)$ was estimated by firing 100 test trajectories from

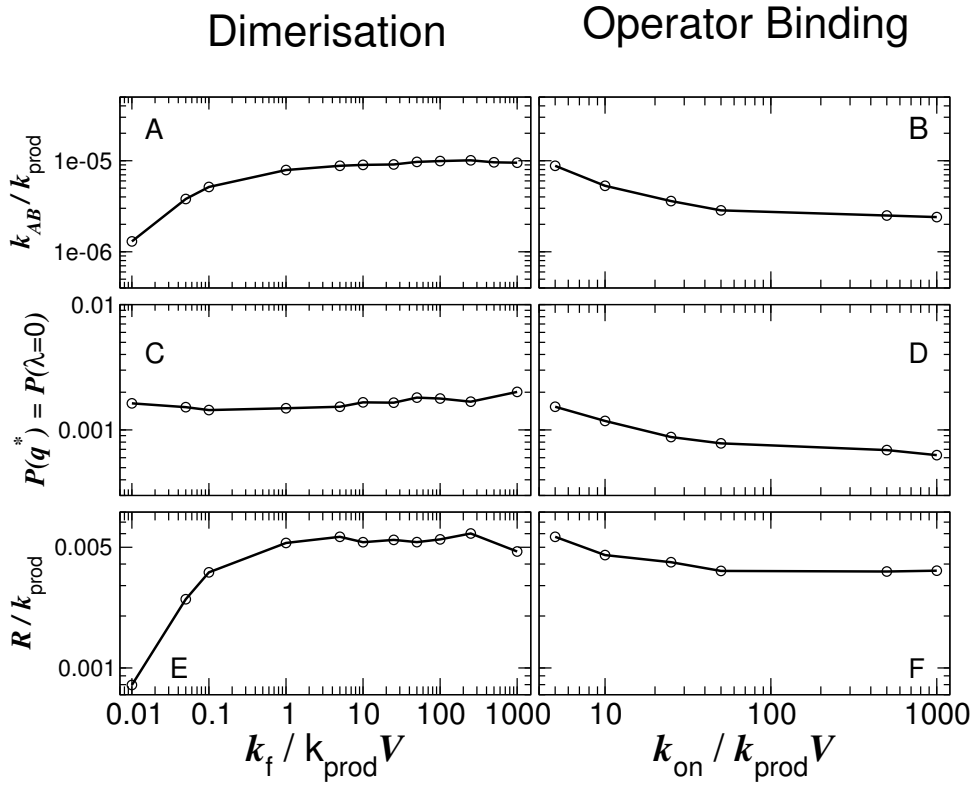


Figure 2.2: Dependence of the switching rate on the dimerisation speed (A) and on the rate of operator binding (B). Backward reaction rates are scaled such that equilibrium constants remain constant: $K_D^d = k_b/k_f = 1/V$ and $K_D^b = k_{off}/k_{on} = 1/(5V)$. The switching rate k_{AB} can be written as the product of two factors: $k_{AB} = RP(q^*) = RP(\lambda = 0)$, where R is a kinetic prefactor and $P(q^*) = P(\lambda = 0)$ is the relative probability of being at the dividing surface. The latter quantity, $P(q^*) = P(\lambda = 0)$, can be derived from the stationary distribution $P(\lambda)$, which can obtained from the FFS calculation to obtain k_{AB}, and which is plotted in Figure 2.3. The middle panels show $P(q^*)$ as a function of the dimerisation speed (C) and the rate of operator binding (D), while the lower panels show the kinetic prefactor R as a function of both rates (E and F, respectively). It is seen that both the kinetic prefactor R (F) and the probability of being at the diving surface $P(q^*)$ (D) decrease as the rate of operator binding increases. In contrast, the probability of being at the separatrix (C) is fairly insensitive to the rate of dimerisation, while the kinetic prefactor (E) increases with increasing dimerisation speed. The three pairs of panels have different scales on the y axis to improve visualization.

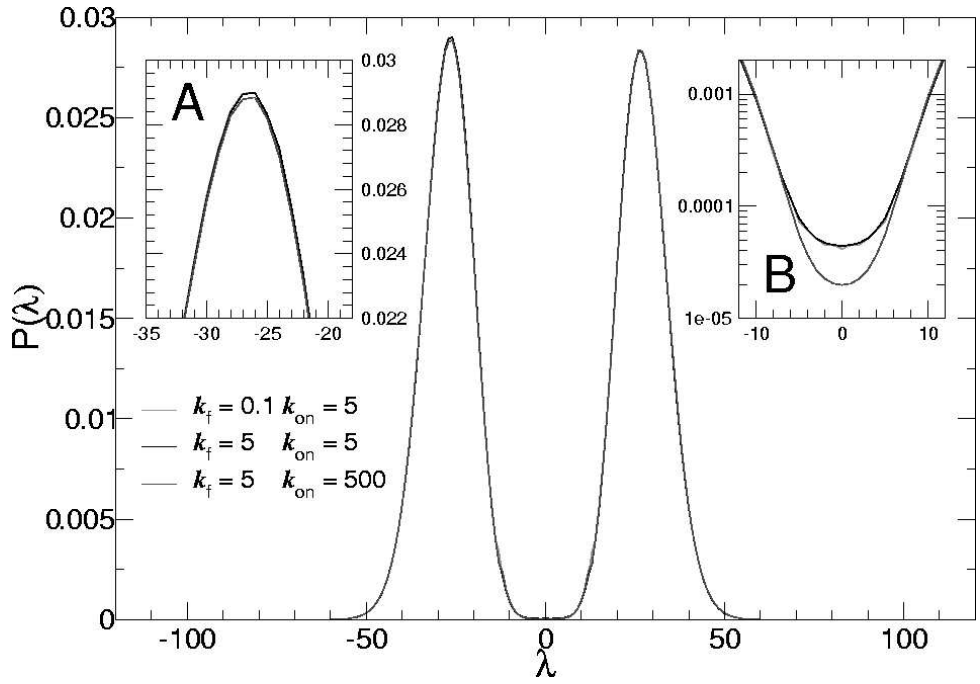


Figure 2.3: Probability distribution as a function of the order parameter $\lambda = n_A + 2n_{A_2} + 2n_{OA_2} - (n_B + 2n_{B_2} + 2n_{OB_2})$, as obtained from the FFS calculations [51], for three different sets of parameters. Inset A magnifies the left peak and demonstrates that in the basins of attraction, the stationary distribution does not depend upon the rate of dimerisation and operator binding. However, as shown in Inset B, fast operator binding does modify the distribution at the top of the "flipping barrier", *i*.e. around the unstable steady state $\lambda = 0$.

that point. Since P_B is the likelihood that a configuration will end up in B, it could be considered as the true reaction coordinate that measures the progress of the transition.

Figure 2.4 shows the average switching pathways in the n_A, n_B plane, where n_A and n_B are the total copy numbers of species A and B, respectively. The trajectories are averaged in the P_B ensemble: the values of n_A and n_B are averaged over those configurations with the same value of P_B. It is seen that the average paths in the P_B ensemble are rather "noisy": this is due to the fact that P_B is a stochastic quantity, which has to be estimated by a computationally demanding procedure. The inset therefore shows the average switching pathways in the λ ensemble; here, the values of n_A and n_B are averaged over those configurations with the same value of λ. Figure 2.4 shows that while the rate of dimerisation affects the location of the switching pathways as they leave the basin of attraction, it does not influence the location of the transition state ensemble (TSE). In contrast, the rate of operator binding affects the location of the TSE: it is seen that the switching pathways cross the dividing surface at lower values of n_A and n_B.

To characterize the switching pathways further, Figure 2.5 shows the probability that the operator is bound by a B_2 dimer, $\langle n_{OB_2} \rangle$ as a function of P_B, and, in the inset, as a func-

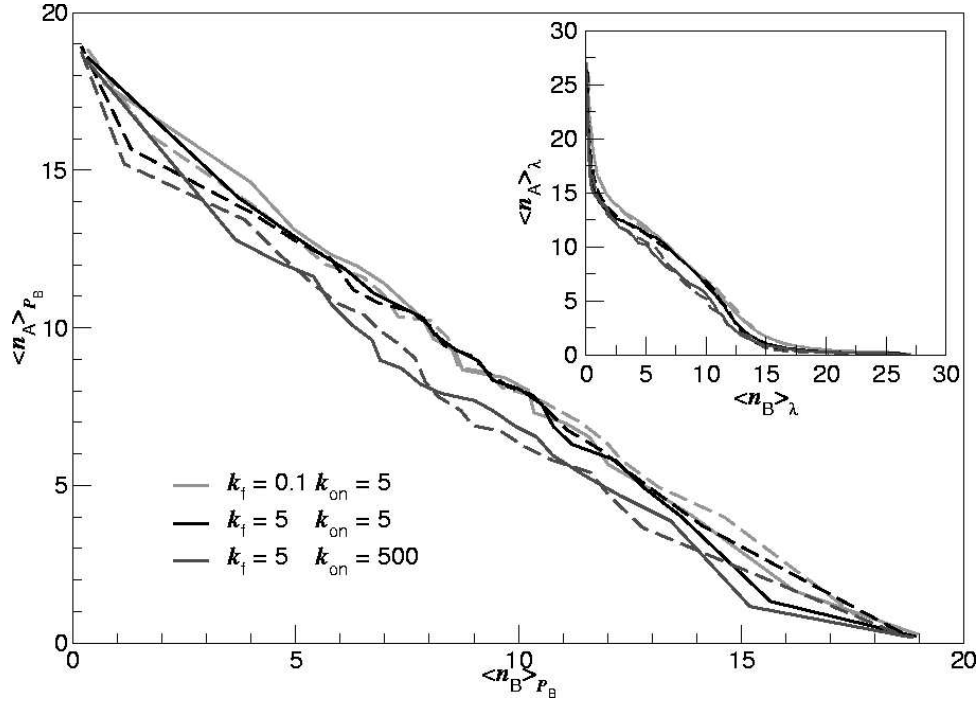


Figure 2.4: Switching paths projected onto the n_A, n_B surface, for three different sets of parameters. The main panel shows the paths averaged in the P_B ensemble, where n_A and n_B are averaged over configurations with the same value of the committor P_B. The inset shows the average paths in the λ ensemble, where the values of n_A and n_B are averaged over configurations with the same value of $\lambda = n_A + 2n_{A_2} + 2n_{OA_2} - (n_B + 2n_{B_2} + 2n_{OB_2})$, where n_X is the copy number of species X. The forward paths, corresponding to transitions from A to B are shown with solid lines, while the reverse transitions, from B to A are shown with dashed lines. Note that the location of the transition state ensemble depends on the rate of operator binding, but is fairly insensitive to the rate of dimerisation.

tion of λ. The solid lines correspond to the average switching paths of the transition from A to B, while the dashed lines corresponds to the paths of the reverse transition, from B to A. It is seen that when the rate of operator binding is fast, the forward and backward paths essentially coincide. This situation differs markedly for the system with the base-line parameter set and for the system with slow dimerisation: although the switch is symmetric on interchanging A and B, the transition path ensemble (TPE) for the transition from A to B does not coincide with that from B to A [42]. This is a manifestation of the fact that this switch is a non-equilibrium system: for equilibrium systems that obey detailed balance and microscopic reversibility, the forward and backward paths must necessarily coincide.

The fact that the forward and backward paths do not coincide also means that the switching paths do not follow the most probable steady-state path in phase space, which, for equilibrium systems, would correspond to the lowest free-energy path: Since this

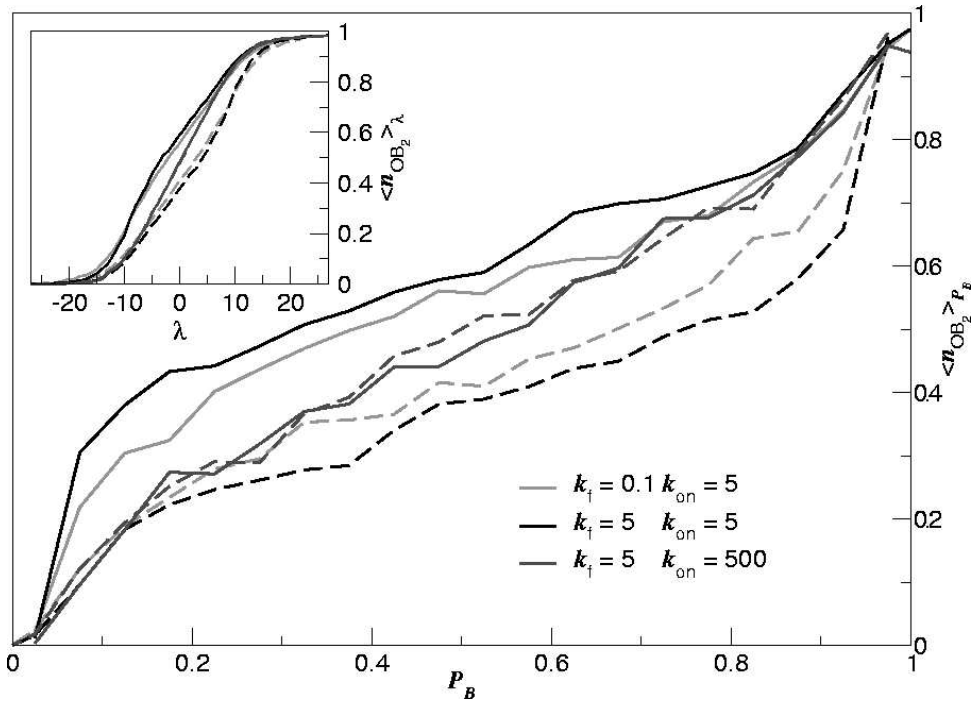


Figure 2.5: The probability that a B dimer is bound to the operator, $\langle n_{\mathrm{OB}_2} \rangle$, as a function of the committor P_B (main panel) and as a function of λ for three different sets of parameters. The solid lines correspond to the transition from A to B, while the dashed lines corresponds to the reverse transition from B to A. It is seen that when the rate of operator binding is fast ($k_{\mathrm{on}} = 500$) the forward and backward paths coincide, while for the other systems they do not.

system is symmetric, this "lowest-free energy path" is symmetric on interchanging species A and B, while Figure 2.5 shows that the dynamical switching trajectories are not (unless operator binding is fast). This also means that it is essential not to make the Markovian assumption of memory loss, which underlies path sampling schemes such as Milestoning [62] and PPTIS [63].

That the forward and reverse transition follow different routes in state space (Figure 2.5) shows that the system actively exploits operator state fluctuations to flip the switch [42, 50]. Indeed, the progress of the transition, and hence the reaction coordinate, is not only determined by the difference in the number of protein molecules, λ, but also by which type of protein happens to be bound to the operator. This is illustrated in Figure 2.6, which shows for the three sets of parameters, the probability distribution $P(\lambda)$ for the TSE for the forward transition from A to B, separated into the components due to the three operator states O, OA_2, and OB_2. First of all, it should be noted that in the TSE the state of the operator and λ are *correlated*: the histograms for OA_2 are shifted to higher values of λ with respect to those for OB_2. This means that if a B dimer is bound

to the operator, then, on average, the number of A molecules has to exceed the number of B molecules in order to have the same value of P_B, and vice versa. Secondly, for the system in which the operator binding and unbinding is fast, at the separatrix the probability $\langle n_{OA_2} \rangle$ that an A species is bound to the operator (which is given by the area under the histogram corresponding to OA_2), equals the probability $\langle n_{OB_2} \rangle$ that a B species is bound to the operator. In contrast, for the reference system and the system with slow dimerisation, $\langle n_{OB_2} \rangle > \langle n_{OA_2} \rangle$ in the transition state ensemble. This unambiguously demonstrates that the system uses the binding of B to the operator to flip the switch from state *A* to state *B*.

We can now understand the dependence of the flipping rate on the rate of operator binding (Figure 2.2). In the non-adiabatic limit of slow operator binding and unbinding [42, 50], the binding of the minority species to the operator strongly enhances the flipping of the switch: when the minority species happens to bind the operator, it will stay on the DNA for a relatively long time, thus blocking the synthesis of the majority species and allowing the production of the minority species. Indeed, in this limit, the system reaches the dividing surface with only a few operator binding and unbinding events. As the rate of operator binding and unbinding is increased, the state of the operator is increasingly being slaved to the difference in the total number of A and B molecules, λ. In the adiabatic limit of fast operator binding, the probability that a molecule of type *A* or *B* is bound to the operator is completely determined by λ [50]. In this limit, the dividing surface is located at $\lambda \approx 0$ and $\langle n_{OA_2} \rangle \approx \langle n_{OB_2} \rangle$; to reach the separatrix, the system has to wait for a series of fluctuations in the birth and decay of both species that lead to $n_A \approx n_B$. This implies that the total number of copies of A and B at the dividing surface decreases as the rate of operator binding increases (see inset Figure 2.4). It also explains why the prefactor *R* and the probability of being at the separatrix, $P(q^*)$, and hence the switching rate, decrease as the rate of operator binding increases (Figure 2.2).

Figures 2.4–2.6 suggest that the rate of dimerisation only has a marginal effect on the switching pathways. However, it should be realised that this could be a result of projecting the switching pathways onto the wrong coordinates. Indeed, Figure 2.2 shows that the rate of dimerisation does affect the switching rate if $k_f < k_{prod} V$. It is conceivable that dimerisation also affects the switching pathways in this regime, but that we have to find other order parameters to describe the switching pathway. We have investigated a number of order parameters, but with limited success. The most successful result is shown in Figure 2.7, which shows $\langle n_{B_2} \rangle$ as a function of $\langle n_B \rangle^2$. When the dimerisation reaction remains in equilibrium during the transition, this function should be a straight line with a slope given by the dissociation constant K_D. Figure 2.7 shows that this is the case for the system with the base-line parameters and for the system with the fast operator binding. For the system with slow dimerisation, however, deviations from this equilibrium scenario are visible: while at the transition state the dimerisation reaction is in equilibrium, at the "pre critical" and "post critical" side of the barrier the dimerisation reaction is out of equilibrium. Note that a similar behaviour can be seen for the projections of the switching pathways in the n_A, n_B plane (inset Figure 2.4). The reason for this behaviour is that the

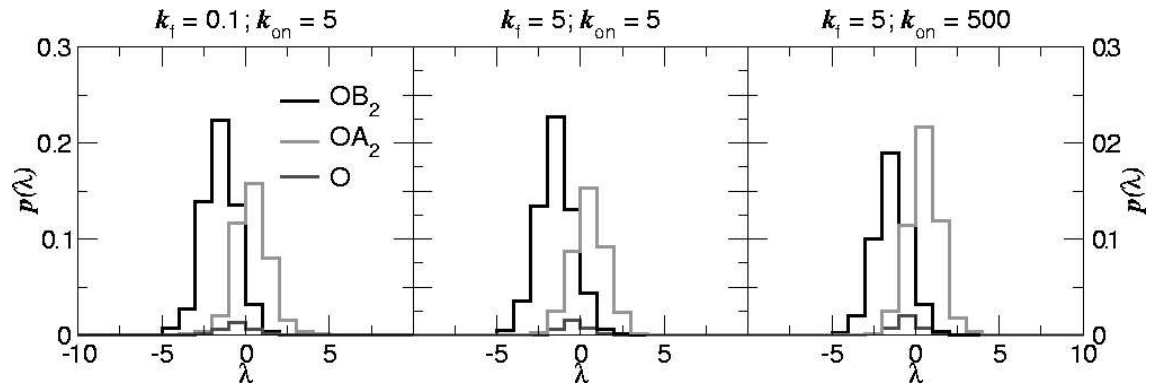


Figure 2.6: The probability $p(\lambda)$ for the transition state ensemble ($P_B = 0.5$) of the transition from A to B, split into color-coded contributions from the three operator states; the area under each histogram gives the probability $\langle n_{\mathrm{OX}} \rangle$ that the operator is bound to species X (the three areas thus sum to unity). The left panel corresponds to the system with slow dimerisation $k_{\mathrm{f}} = 0.1$; the middle panel corresponds to the system with the base-line parameters; the panel on the right corresponds to the system with fast operator binding $k_{\mathrm{on}} = 500$. Note that in all cases the state of the operator is correlated with λ. Note also that with fast operator binding (right panel), $\langle n_{\mathrm{OA}_2} \rangle \approx \langle n_{\mathrm{OB}_2} \rangle$, while in the other cases predominantly species B is bound to the operator in the TSE ensemble: $\langle n_{\mathrm{OA}_2} \rangle < \langle n_{\mathrm{OB}_2} \rangle$.

switching pathways spend relatively little time on the flanks of the barrier, while, because of the diffusive dynamics, they stay fairly long on the top of the barrier. On the top of the barrier, the dimerisation reactions thus have time to equilibrate, even when they are fairly slow (*i.e.* with $k_{\mathrm{f}} = 0.1 k_{\mathrm{prod}} V$).

The dependence of the switching pathways (Figures 2.4-2.7) on the rate of dimerisation helps to understand the importance of dimerisation for the switching rate. The dimerisation reactions mostly affect the dynamics of the trajectories, in particular as they leave their basins of attraction. In contrast, they hardly change the location of the transition state ensemble. Together, these observations explain why the dimerisation reaction affects the switching rate via the kinetic prefactor R, and not via the relative probability of being at the top of the barrier, $P(q^*)$. Our results thus show that the underlying dynamics of the system can have a large effect on the switching rate. In this case, it is caused by an interplay between the time scales of dimerisation and protein decay events: when the system is in a stable steady state, in order to start a switching event, two copies of the minority species must be produced. They must then dimerise and bind to the operator, to efficiently increase their production. If the dimerisation rate is comparable to the degradation rate, it becomes increasingly probable that copies of the minority species are removed from the system before they can for a dimer. This effect is thus truly dynamical in origin and fundamentally different from the enhanced switch stability via cooperativity due to nonlinear degradation [64].

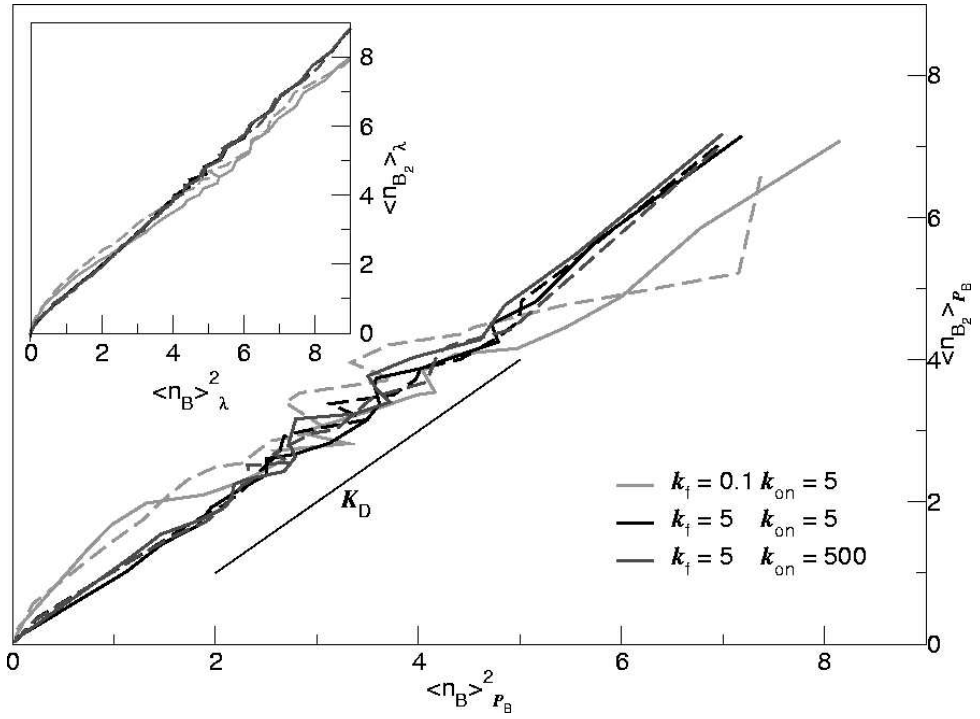


Figure 2.7: The copy number of the dimer B_2, $\langle n_{B_2} \rangle$, as a function of the square of the copy number of the monomer B, $\langle n_B \rangle^2$, for the forward paths (solid lines) and backward paths (dashed lines) for three different sets of parameters. If dimerisation would be in equilibrium during the transition, then $\langle n_{B_2} \rangle$ as a function $\langle n_B \rangle^2$ would be given by a straight line with a slope given by the dissociation constant K_D. It is seen that when the dimerisation is slow ($k_f = 0.1$), the dimerisation reaction is out of equilibrium during the transition.

2.4 Discussion

We have studied the flipping rate of a genetic toggle switch as a function of the rate of dimerisation and operator binding. To this end, we have varied the rate constants of these reactions over more than four orders of magnitude (see Figure 2.2). This large range is important, because the rate constants of biochemical reactions tend to vary over a wide range. For instance, in prokaryotic cells, the inverse rate of protein production, k_{prod}^{-1}, is in the range of seconds to minutes [65]. With the size of a typical *E. coli* cell being on the order of $1\mu\text{m}^3$, this corresponds to $k_{\text{prod}}V = 10^{-2} - 10\,\text{nM}^{-1}/\text{min}$. The rate of dimer association, k_f, is about $10^{-2} - 10^{-1}\text{nM}^{-1}/\text{min}$, while the dimer dissociation rate is on the order of $k_d = 10^{-2} - 10^3/\text{min}$, corresponding to dissociation constants in the range $K_D^d = 0 - 10^2\text{nM}$ [64]. This means that $k_f = 10^{-2} - 10 k_{\text{prod}}V$. Figure 2.2A shows that while for $k_f > k_{\text{prod}}V$ the switching rate is fairly insensitive to changes in the dimerisation rate, the flipping rate reduces strongly as the dimerisation rate is decreased when $k_f < k_{\text{prod}}V$.

Hence, under biologically relevant conditions, the rate of dimerisation can strongly affect the switching rate. The rate of operator (un)binding can vary over a similar broad range as the rate of dimerisation [65]. This means, as Figure 2.2B shows, that also operator state fluctuations can have a marked effect on the flipping rate of genetic switches in living cells.

Figure 2.2 shows that while dimerisation affects the switching rate predominantly via the kinetic prefactor, operator binding affects the switching rate via the kinetic prefactor and the probability of being at the separatrix: both decrease as the rate of operator binding increases. Interestingly, the steady-state phase-space density in the basins of attraction are much more robust to changes in the rate of operator binding and dimerisation (see Figure 2.3). Clearly, dimerisation and operator binding do not significantly affect the behaviour of the network in the stable state, but can strongly affect the switching of one stable state to the next.

Genetic switches have become a paradigm for rare events in non-equilibrium systems. In the analysis of rare events in equilibrium systems, it is often assumed that one coordinate, the reaction coordinate, is slow, while the other degrees of freedom are fast. If this is the case, then the transitions can be accurately modeled by assuming that the reaction coordinate evolves according to a Langevin equation, while the other degrees of freedom create friction and provide the activation energy to cross the free-energy barrier. The "barrier crossing" in the toggle switch differs fundamentally from this classical scenario. The reaction coordinate consists of at least two parameters, namely the difference in total copy number of species A and B and the state of the operator [42]. Moreover, these coordinates move on comparable time scales—the operator state fluctuates on time scales similar to those of protein production and decay; in addition, their dynamics mix in a non-equilibrium fashion [50]—the degradation and production of proteins are non-equilibrium processes. This hampers the application of standard theoretical tools to model barrier crossings [50]. Indeed, it appears that new theoretical approaches are required to accurately model such rare events in non-equilibrium systems.

Chapter 3

Eliminating fast reactions in stochastic simulations of a genetic switch

Our life is frittered away with detail.
Simplify, simplify, simplify!
Henry David Thoreau

In many stochastic simulations of biochemical reaction networks, it is desirable to "coarse-grain" the reaction set, removing fast reactions while retaining the correct system dynamics. For "fast" reactions we mean here reactions that have a high propensity to happen, because of fast reaction rates, or large number of reactants, or both. Various coarse-graining methods have been proposed, but it remains unclear which methods are reliable and which reactions can safely be eliminated. We address these issues for a model gene regulatory network that is particularly sensitive to dynamical fluctuations: a bistable genetic switch. We remove protein-DNA and/or protein-protein association-dissociation reactions from the reaction set, using various coarse-graining strategies. We determine their effects on the steady-state probability distribution function and on the rate of fluctuation-driven switch flipping transitions. We find that protein-protein interactions may be safely eliminated from the reaction set, but protein-DNA interactions may not. We also find that it is important to use the chemical master equation rather than macroscopic rate equations to compute effective propensity functions for the coarse-grained reactions.

3.1 Introduction

Biochemical reaction networks control how living cells function. Computer simulations provide a valuable tool for understanding how complex biochemical network architecture is connected to cellular function. A popular method for simulating biochemical networks is the "Stochastic Simulation Algorithm" (SSA) which was introduced in this field by Gillespie [66, 35]. For many reaction networks, however, SSA simulations are prohibitively expensive because of "time-scale separation": the reaction set contains some reactions which occur much more frequently than others. For every "slow" reaction event, many "fast" reaction events have to be simulated. This problem has led to the development of various methods for coarse-graining the reaction set [67, 68, 69, 70, 71, 72, 73] - that is, eliminating the fast reactions and simulating only the slow reactions. Key issues are which fast reactions can safely be eliminated, and how this should be done, so as not to disturb the original dynamics. In this Chapter, we address these issues for a biochemical network which is especially sensitive to dynamical fluctuations: a bistable genetic switch. Because of its sensitivity, this model provides a useful test system for assessing how to coarse-grain biochemical networks. We expect our conclusions to be valid for a wide range of biochemical networks where fluctuation-driven processes are important.

The SSA is a kinetic Monte Carlo method which generates trajectories for the number of molecules of each chemical species in the reaction system. The molecular discreteness of the reacting species is included, as well as stochastic fluctuations in the numbers of molecules, assuming that each reaction is a Poisson process. It is assumed that the system is well-stirred - *i.e.* possible inhomogeneities in the spatial distribution of the components are ignored (alternative methods that do include spatial effects have recently been developed [38, 39, 40, 41]). The SSA generates trajectories that are consistent with the chemical master equation. Its implementation is outlined in Section 1.6.1 - in brief, reaction propensities are computed, which are the probability per unit time for each reaction to occur. These propensities are used to determine the time and identity of the next reaction event. Modifications of the SSA for large numbers of reaction channels or for high copy numbers of the reacting species have been developed [37, 74].

The SSA becomes inefficient when some of the reaction channels ("the fast reactions") have much higher propensities than others ("slow reactions") - this is known as the time-scale separation problem, and several methods for dealing with it have been proposed. In all cases, the first step is to identify which reactions are "fast" and which are "slow". The criterion is generally that fast reactions should reach a steady state faster than the waiting time between slow reaction events. The slow reactions are generally simulated using the SSA, while the various methods differ in their treatment of the fast reactions. In one class of methods, the fast reactions are propagated using the deterministic or chemical Langevin equation [67, 75, 76]. Alternatively, one may assume that the propensity functions of the fast reactions do not change between firings of the slow reaction, as in the τ leap method of Gillespie [74, 70]. The accuracy of these techniques requires that the species in the fast reactions are present in large copy numbers. In another class of methods, which we consider here, the fast reactions are eliminated entirely and the slow reactions are

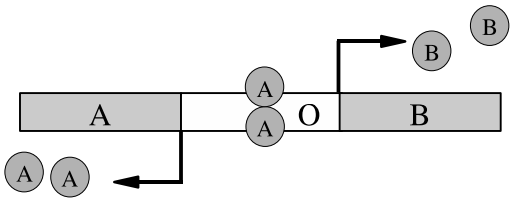


Figure 3.1: Pictorial representation of our model switch, corresponding to Eq. (2.2). Two divergently-transcribed genes are under the control of a shared regulating binding site on the DNA, called the operator. Both proteins can bind, in the homodimer form, to the operator and block the production of the other species.

propagated according to effective propensities that depend on steady-state averages over the chemical master equation for the fast reactions. Because these algorithms take account of molecular discreteness, they do not require the species in the fast reactions to be present in large copy numbers. These schemes are discussed in more detail in Section 3.3.

In this Chapter, we use a bistable genetic switch as a test system for assessing various coarse-graining strategies. Our model switch is a gene regulatory network with two genes, in which the protein product of gene A dimerises and represses gene B, and vice versa [61, 52, 55, 49, 42, 57, 56]. The network occasionally flips spontaneously between its two stable states due to random fluctuations. The switch flipping rate is highly sensitive to the fluctuations in the network and can be used as a measure for whether the coarse-graining schemes correctly reproduce the network dynamics. The model switch shows dynamics on a wide range of timescales - protein-protein and protein-DNA association and dissociation over seconds to minutes, synthesis and degradation of proteins over tens of minutes, and switch flipping over tens of hours. It is by no means obvious whether the protein-protein and protein-DNA interactions may safely be eliminated, since it is likely that fluctuations in these reactions are crucial in driving the longer timescale switch flipping events [42, 50]. We investigate the consequences of eliminating the protein-protein and protein-DNA interactions for this system, using various coarse-graining schemes for the SSA, in combination with the recently developed "Forward Flux Sampling" (FFS) method for rare event simulations [42, 43, 44]. We compute steady-state probability distributions, as well as the switch flipping rate. We compare coarse-graining strategies in which averages over the fast reactions are computed using the chemical master equation to those where macroscopic rate equations are used. Our results show that the chemical master equation rather than the macroscopic rate equations should be used for integrating out fast reactions. We find that the steady-state distribution of the system is quite insensitive to removing either protein-protein or protein-DNA association and dissociation reactions, but the switch flipping rate is strongly affected by coarse-graining over protein-DNA interactions and less affected by removing protein-protein interactions.

In the next Section, we describe the model genetic switch. In Section 3.3, we give background information on the various coarse-graining schemes, and in Section 3.4, we discuss these in the context of the model switch. In Section 3.5, we present results on the

Reaction	Propensity	Reaction	Propensity	
$A + A \rightleftharpoons A_2$	$k_f n_A(n_A-1), k_b n_{A_2}$	$B + B \rightleftharpoons B_2$	$k_f n_B(n_B-1), k_b n_{B_2}$	(3.1a)
$O + A_2 \rightleftharpoons OA_2$	$k_{on} n_O n_{A_2}$, $k_{off} n_{OA_2}$	$O + B_2 \rightleftharpoons OB_2$	$k_{on} n_O n_{B_2}$, $k_{off} n_{OB_2}$	(3.1b)
$O \to O + A$	$k_{prod} n_O$	$O \to O + B$	$k_{prod} n_O$	(3.1c)
$OA_2 \to OA_2 + A$	$k_{prod} n_{OA_2}$	$OB_2 \to OB_2 + B$	$k_{prod} n_{OB_2}$	(3.1d)
$A \to \emptyset$	μn_A	$B \to \emptyset$	μn_B	(3.1e)

Table 3.1: Reactions and propensity functions for the model genetic switch.

computational speed-up and the accuracy of the various coarse-graining procedures using the stationary distribution and the switching rates as read-outs. We end with a discussion on the implications of our findings for the simulation of complex biochemical networks.

3.2 The Model Genetic Switch

The model bistable genetic switch is shown schematically in Figure 3.1 and the set of reactions is listed in Table 3.1 [61, 77, 49]. As shown in Figure 3.1, two genes A and B are transcribed in divergent directions, under the control of a single operator region, O, which contains a single binding site. Coding region A encodes protein A, while coding region B encodes protein B. Both proteins A and B are transcription factors, which, upon homodimerisation, are able to bind to the operator sequence O. When A_2 is bound at O, the transcription of B is blocked [A_2 is a repressor for B]; while, conversely, when B_2 is bound at O, the transcription of A is blocked [B_2 is a repressor for A]. When neither A_2 or B_2 is bound at O, both A and B are transcribed at the same average rate k_{prod}. Protein monomers are also removed from the system with rate μ, modelling active degradation processes as well as dilution due to cell growth. For convenience in this model system, we use the rather high value $\mu = 0.3 k_{prod}$, corresponding to the case where removal from the cell is dominated by active degradation. In our model, we assume that all the steps leading to production of a protein molecule (transcription, translation and protein folding) can be modelled as a single Poisson process with rate constant k_{prod}. The system is symmetric on exchanging A and B.

For this model system, bistability has been demonstrated using a mean field analysis and with simulations [49]. In one stable state, a large number of A proteins are present; this ensures that the operator O is mostly bound by A_2, keeping B repressed. Conversely, in the other stable state, B proteins are abundant, so that O is mostly bound by B_2, and A remains repressed. Previous work has demonstrated that, when simulated stochastically with appropriate parameters, the system makes occasional random flipping transitions between these two stable states [49], as in Figure 3.2. In our simulations, we use the inverse of the production rate, k_{prod}^{-1} as the unit of time. We assume that the cell volume

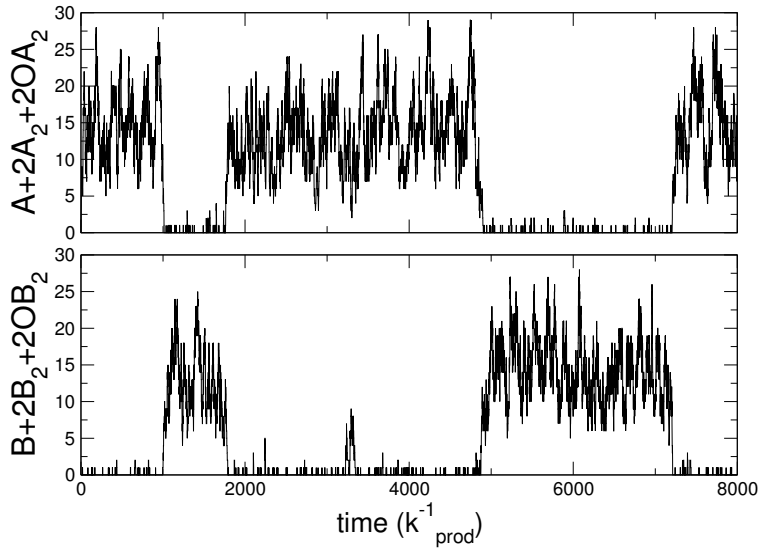


Figure 3.2: Typical simulation trajectory for the model switch, with baseline parameters except for μ which is replaced by $\mu = 0.45k_{\text{prod}}$. The total numbers of A and B molecules fluctuate around two stable states, one rich in A and the other rich in B. Transitions between these states are rapid, yet infrequent.

V remains constant. For simplicity, we use a value $V = 1$, and define our rate constants in appropriate units. We choose a "baseline" set of parameters, in the region of parameter space where the system has previously been found to be bistable: $k_{\text{f}} = 5k_{\text{prod}}V$, $k_{\text{b}} = 5k_{\text{prod}}$ (so that $K_{\text{D}}^{d} = k_{\text{d}}/k_{\text{f}} = 1/V$), $k_{\text{on}} = 5k_{\text{prod}}$, $k_{\text{off}} = k_{\text{prod}}$ (so that $K_{\text{D}}^{b} = k_{\text{off}}/k_{\text{on}} = 1/(5V)$), $\mu = 0.3k_{\text{prod}}$. This model system is loosely based on the bacteriophage λ genetic switch [78]. For the phage λ proteins cro and cI, assuming diffusion-limited protein-DNA association and using Refs. [79] and [80], $k_{\text{on}} \approx 5-10k_{\text{prod}}$. For cI and cro dimer formation, using Refs. [81] and [82], $k_{\text{f}} \approx 50-100k_{\text{prod}}$. Protein degradation rates are much lower for phage λ (of the order of $\mu \approx 0.1-0.01k_{\text{prod}}$) than for our model system - this contributes to the observed stability of the phage λ switch.

Throughout this Chapter, we represent the number of molecules of chemical species X which is present in the cell by n_{X}. Later in the Chapter, we will need to characterize the switching process by an "order parameter", which we denote λ. A natural choice is the difference between the total number of the two proteins in the cell: $\lambda \equiv n_{\text{A}} + 2n_{\text{A}_2} + 2n_{O\text{A}_2} - (n_{\text{B}} + 2n_{\text{B}_2} + 2n_{O\text{B}_2})$. Figure 3.2 shows λ plotted as a function of time for a simulation of this reaction set using the SSA. Bistable behaviour is indeed observed: the system spends most of its time in one of the two stable states with occasional transitions between states. The average duration of a flipping transition event is much shorter than the average "waiting time" between the flipping transitions.

3.3 Dynamical coarse-graining: background

Dynamical coarse-graining schemes begin by splitting the reaction set into fast and slow reactions, as described in Section 3.1. The slow reactions are generally simulated using the SSA. The fast reactions are approximated in ways that differ for different methods. In

this Chapter, we only consider approaches in which the fast reactions are removed entirely from the reaction set, by assuming that they relax to a steady state faster than the waiting time between slow reactions. Effective propensities for the slow reactions are computed as averages over the steady state distribution, obtained from the chemical master equation for the fast reactions.

The key step is the determination of the effective propensity functions $\overline{a}_j^s(n_s, n_f)$ for the slow reactions in the coarse-grained reaction scheme. These depend on the copy numbers n_s of the "slow species" (those that are only affected by the slow reactions), and n_f of the "fast species" (those that are affected by both the fast and the slow reactions). The effective propensities are given by

$$\overline{a}_j^s(n_f, n_s) = \sum_{n'_f} P_\infty(n'_f | n_f, n_s) a_j^s(n'_f, n_s), \tag{3.2}$$

where a_j^s denotes the propensity function for a given slow reaction j and $P_\infty(n'_f | n_f, n_s)$ is the probability of obtaining a given copy number n'_f for the fast species, at the end of a very long simulation of the fast reaction set only, starting from state space point (n_f, n_s). These effective propensities are designed to give the same flux along the slow reaction channel, on average, as in the full system.

The effective propensity functions in Eq. (3.2) can be obtained by performing short SSA simulations of the fast reactions at fixed copy numbers of the slow species [72, 73]. Alternatively, one may solve the chemical master equation for the fast reactions analytically or numerically [68, 69, 71].

It is important to discuss the definition of the "fast variables" (n_f in Eq. (3.2)) and "slow variables" (n_s) [68, 71, 34]. In the work of Cao *et al.* [71], the slow variables are the copy numbers of those species which are unaffected by the fast reactions, while the fast variables can be changed by both fast and slow reactions. During the coarse-grained simulation, both the fast and the slow variables are propagated in time, even though only the slow reactions are simulated. However, Bundschuh *et al.* [68] describe a way to eliminate not only the fast reactions but also the fast variables from the simulation scheme - so that the coarse-grained simulation includes only slow species and slow variables. Here, the fast variables are the copy numbers of all species which are affected by the fast reactions. The slow variables are made up of the copy numbers of species which are unchanged by the fast reactions, as well as *combinations* of the fast variables. These combinations are chosen so that they are unchanged by the fast reactions. For example, for a fast reaction set $2\mathrm{A} \rightleftharpoons \mathrm{A}_2$, an appropriate slow variable would be $n_{\breve{\mathrm{A}}} = n_{\mathrm{A}} + 2n_{\mathrm{A}_2}$. The original slow reaction set is then rewritten in terms of these new slow variables. This eliminates all the fast species from the slow reaction set and the simulation proceeds by simulating only the slow variables. This is the strategy which we adopt in this Chapter.

3.4 Coarse-graining for the model genetic switch

Stochastic simulations of the model genetic switch have characteristic features that depend on the timescale over which we observe the simulation. In a timeframe of about $0.2k_{\text{prod}}^{-1}$, we will observe mainly protein-protein association and dissociation events (typical timescale $[n_{\text{A}}(n_{\text{A}}-1)k_f]^{-1}$ and $[n_{\text{A}_2}k_b]^{-1}$ respectively), as well as protein-operator association and dissociation (typical timescale $[n_{\text{A}_2}k_{\text{on}}]^{-1}$ and $[k_{\text{off}}]^{-1}$). If we extend our observation "window" to a timeframe of about $4k_{\text{prod}}^{-1}$, we observe also protein production and degradation (typical timescale $[k_{\text{prod}}]^{-1}$) and $\mu^{-1} \approx 3k_{\text{prod}}^{-1}$). In a much longer timeframe, we observe flipping events between the two stable states. It is these flipping events that are the phenomenon of interest - yet for each "interesting" switch flipping event, very many "less interesting" association and dissociation events need to be simulated. This is an example of timescale separation, which we seek to overcome by coarse-graining - eliminating the protein-protein and/or protein-DNA association and dissociation reactions from the simulation scheme.

To coarse-grain the model genetic switch, we divide the full reaction set (3.1) into fast and slow reactions. We will consider three cases: (i) protein-protein association and dissociation reactions (3.1a) are fast, (ii): protein-DNA association and dissociation reactions (3.1b) are fast, and (iii): both reactions (3.1a) and (3.1b) are fast. For each of these cases, we define fast and slow variables. The fast variables are the copy numbers n_f of species which are affected by the fast reactions. The slow variables n_s are either the copy numbers of species unaffected by the fast reactions, or linear combinations of fast variables which are unchanged by any of the fast reactions - *e.g.* $n_{\breve{\text{A}}} + n_{\text{A}} + 2n_{\text{A}_2}$. To accompany these slow variables, we define "slow species". These may represent either single chemical species, or combinations of species. For example, the species $\breve{\text{A}}$ represents an A molecule which is in either monomer or dimer form. We then rewrite the slow reaction set in terms of the new slow species, and carry out a simulation using Gillespie's SSA, with effective propensities given by Eq. (3.2). For this system, only the first moment of $P_\infty(n'_f|n_f,n_s)$ is required. This can be obtained by solving the chemical master equation for the fast reactions at a fixed value of the slow variables. Alternatively, one may approximate the macroscopic rate equations for the fast reactions. A summary of the various coarse-graining methods used in this Chapter is given in Table 3.2. Table 3.2 also lists the coarse-grained reaction sets, and gives formulae for the effective propensity functions. The notation $\langle \text{X} \rangle^{\text{METHOD}}_{\text{SLOW VARIABLES}}$ denotes the first moment (average) of the steady state probability distribution function $P_\infty(n'_f|n_f,n_s)$, the superscript denoting whether the master equation or macroscopic rate equation is used to find the average, and the subscript indicating which slow variables the average depends upon.

3.4.1 Coarse-graining protein-DNA binding

We first remove the protein-DNA association and dissociation reactions

$$O + \text{A}_2 \rightleftharpoons O\text{A}_2 \qquad O + \text{B}_2 \rightleftharpoons O\text{B}_2 \tag{3.3}$$

Name	Reaction	Propensity	Method	C-G Var	Definition
EO	$\emptyset \to \mathrm{A}$	$k_{\mathrm{prod}}\langle n_O + n_{O\mathrm{A}_2}\rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2}$	Rate	$\hat{\mathrm{A}}_2$	$n_{\hat{\mathrm{A}}_2} = n_{\mathrm{A}_2} + n_{O\mathrm{A}_2}$
	$\emptyset \to \mathrm{B}$	$k_{\mathrm{prod}}\langle n_O + n_{O\mathrm{B}_2}\rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2}$	Equation	$\hat{\mathrm{B}}_2$	$n_{\hat{\mathrm{B}}_2} = n_{\mathrm{B}_2} + n_{O\mathrm{B}_2}$
	$\mathrm{A} + \mathrm{A} \rightleftharpoons \hat{\mathrm{A}}_2$	$k_{\mathrm{on}} n_{\mathrm{A}}(n_{\mathrm{A}}-1),\ k_{\mathrm{off}}\langle n_{\mathrm{A}_2}\rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2}$			
	$\mathrm{B} + \mathrm{B} \rightleftharpoons \hat{\mathrm{B}}_2$	$k_{\mathrm{on}} n_{\mathrm{B}}(n_{\mathrm{B}}-1),\ k_{\mathrm{off}}\langle n_{\mathrm{B}_2}\rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2}$			
	$\mathrm{A} \to \emptyset$	μn_{A}			
	$\mathrm{B} \to \emptyset$	μn_{B}			
ED1	$O + 2\check{\mathrm{A}} \rightleftharpoons O\mathrm{A}_2$	$k_{\mathrm{on}}\langle n_{\mathrm{A}_2}\rangle^{\mathrm{RE}}_{\check{\mathrm{A}}},\ k_{\mathrm{off}} n_{O\mathrm{A}_2}$	Rate	$\check{\mathrm{A}}$	$n_{\check{\mathrm{A}}} = n_{\mathrm{A}} + 2n_{\mathrm{A}_2}$
	$O + 2\check{\mathrm{B}} \rightleftharpoons O\mathrm{B}_2$	$k_{\mathrm{on}}\langle n_{\mathrm{B}_2}\rangle^{\mathrm{RE}}_{\check{\mathrm{B}}},\ k_{\mathrm{off}} n_{O\mathrm{B}_2}$	Equation	$\check{\mathrm{B}}$	$n_{\check{\mathrm{B}}} = n_{\mathrm{B}} + 2n_{\mathrm{B}_2}$
	$O \to O + \check{\mathrm{A}}$	$k_{\mathrm{prod}} n_O$			
	$O \to O + \check{\mathrm{B}}$	$k_{\mathrm{prod}} n_O$			
	$O\mathrm{A}_2 \to O\mathrm{A}_2 + \check{\mathrm{A}}$	$k_{\mathrm{prod}} n_{O\mathrm{A}_2}$			
	$O\mathrm{B}_2 \to O\mathrm{B}_2 + \check{\mathrm{B}}$	$k_{\mathrm{prod}} n_{O\mathrm{B}_2}$			
	$\check{\mathrm{A}} \to \emptyset$	$\mu\langle n_{\mathrm{A}}\rangle^{\mathrm{RE}}_{\check{\mathrm{A}}}$			
	$\check{\mathrm{B}} \to \emptyset$	$\mu\langle n_{\mathrm{B}}\rangle^{\mathrm{RE}}_{\check{\mathrm{B}}}$			
ED2	$O + 2\check{\mathrm{A}} \rightleftharpoons O\mathrm{A}_2$	$k_{\mathrm{on}}\langle n_{\mathrm{A}_2}\rangle^{\mathrm{ME}}_{\check{\mathrm{A}}},\ k_{\mathrm{off}} n_{O\mathrm{A}_2}$	Master	$\check{\mathrm{A}}$	$n_{\check{\mathrm{A}}} = n_{\mathrm{A}} + 2n_{\mathrm{A}_2}$
	$O + 2\check{\mathrm{B}} \rightleftharpoons O\mathrm{B}_2$	$k_{\mathrm{on}}\langle n_{\mathrm{B}_2}\rangle^{\mathrm{ME}}_{\check{\mathrm{B}}},\ k_{\mathrm{off}} n_{O\mathrm{B}_2}$	Equation	$\check{\mathrm{B}}$	$n_{\check{\mathrm{B}}} = n_{\mathrm{B}} + 2n_{\mathrm{B}_2}$
	$O \to O + \check{\mathrm{A}}$	$k_{\mathrm{prod}} n_O$			
	$O \to O + \check{\mathrm{B}}$	$k_{\mathrm{prod}} n_O$			
	$O\mathrm{A}_2 \to O\mathrm{A}_2 + \check{\mathrm{A}}$	$k_{\mathrm{prod}} n_{O\mathrm{A}_2}$			
	$O\mathrm{B}_2 \to O\mathrm{B}_2 + \check{\mathrm{B}}$	$k_{\mathrm{prod}} n_{O\mathrm{B}_2}$			
	$\check{\mathrm{A}} \to \emptyset$	$\mu\langle n_{\mathrm{A}}\rangle^{\mathrm{ME}}_{\check{\mathrm{A}}}$			
	$\check{\mathrm{B}} \to \emptyset$	$\mu\langle n_{\mathrm{B}}\rangle^{\mathrm{ME}}_{\check{\mathrm{B}}}$			
EO-ED1	$\emptyset \to \tilde{\mathrm{A}}$	$k_{\mathrm{prod}}\langle n_O + n_{O\mathrm{A}_2}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$	Rate	$\tilde{\mathrm{A}}$	$n_{\tilde{\mathrm{A}}} = n_{\mathrm{A}} + 2n_{\mathrm{A}_2} + 2n_{O\mathrm{A}_2}$
	$\emptyset \to \tilde{\mathrm{B}}$	$k_{\mathrm{prod}}\langle n_O + n_{O\mathrm{B}_2}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$	Equation	$\tilde{\mathrm{B}}$	$n_{\tilde{\mathrm{B}}} = n_{\mathrm{B}} + 2n_{\mathrm{B}_2} + 2n_{O\mathrm{B}_2}$
	$\tilde{\mathrm{A}} \to \emptyset$	$\mu\langle n_{\mathrm{A}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$			
	$\tilde{\mathrm{B}} \to \emptyset$	$\mu\langle n_{\mathrm{B}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$			
EO-ED2	$\emptyset \to \tilde{\mathrm{A}}$	$k_{\mathrm{prod}}\langle n_O + n_{O\mathrm{A}_2}\rangle^{\mathrm{ME}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$	Master	$\tilde{\mathrm{A}}$	$n_{\tilde{\mathrm{A}}} = n_{\mathrm{A}} + 2n_{\mathrm{A}_2} + 2n_{O\mathrm{A}_2}$
	$\emptyset \to \tilde{\mathrm{B}}$	$k_{\mathrm{prod}}\langle n_O + n_{O\mathrm{B}_2}\rangle^{\mathrm{ME}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$	Equation	$\tilde{\mathrm{B}}$	$n_{\tilde{\mathrm{B}}} = n_{\mathrm{B}} + 2n_{\mathrm{B}_2} + 2n_{O\mathrm{B}_2}$
	$\tilde{\mathrm{A}} \to \emptyset$	$\mu\langle n_{\mathrm{A}}\rangle^{\mathrm{ME}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$			
	$\tilde{\mathrm{B}} \to \emptyset$	$\mu\langle n_{\mathrm{B}}\rangle^{\mathrm{ME}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$			

Table 3.2: Summary of coarse-graining schemes for the original reaction set (3.1): eliminating operator binding (EO), eliminating dimerisation reactions using the Macroscopic Rate Equation (ED1) or the Master Equation (ED2), eliminating both dimerisation and operator binding using the Macroscopic Rate Equation (EO-ED1) or the Master Equation (EO-ED2). For each coarse-graining scheme, the coarse-grained reaction set is indicated together with the propensity function for each reaction. We also give definitions of the new slow variables for each scheme.

from the original reaction scheme (3.1). We denote this approach "Eliminating Operator state fluctuations (EO)". In our coarse-grained simulation, the system will still experience fluctuations due to protein-protein association and dissociation, protein production and protein decay, but not those due to the binding and unbinding of molecules to the DNA.

The "fast species", which are affected by reactions (3.3), are A_2, B_2, OA_2, OB_2 and O. The "slow species" are A, B, $\hat{A}_2$ and $\hat{B}_2$, where A and B are simply the protein monomers - these are unchanged by the fast reactions (3.3) - and $\hat{A}_2$ and $\hat{B}_2$ are new species, such that:

$$\begin{aligned} n_{\hat{A}_2} &= n_{A_2} + n_{OA_2} \\ n_{\hat{B}_2} &= n_{B_2} + n_{OB_2} \end{aligned} \tag{3.4}$$

$n_{\hat{A}_2}$ and $n_{\hat{B}_2}$ are simply the total numbers of dimers in the system - including both free and DNA-bound dimers. The new, coarse-grained, reaction set is given in Table 3.2, together with the effective propensities. As the operator O has been removed from the scheme, protein production has become a simple birth process, in which a monomer appears from "nowhere". The propensity for this birth of a monomer (say A) takes into account the "lost" reactions - it reflects the probability, in the full reaction scheme, of finding the promoter O in one of the states O and OA_2 that are able to produce A. The protein-protein interactions (3.1a) have been changed to reflect the fact that free dimers A_2 have been replaced by the new species $\hat{A}_2$. Two monomers can now reversibly associate to generate a molecule of $\hat{A}_2$, while the reaction representing dissociation of free dimers to monomers has an effective propensity that depends on the average number of free dimers that would be present in the full reaction scheme, for a given value $n_{\hat{A}_2}$ of total dimers. Protein degradations (3.1e) remain unchanged since these reactions affect only monomers.

To evaluate the effective propensities listed in Table 3.2, we require the averages $\langle n_O + n_{OA_2} \rangle_{\hat{A}_2,\hat{B}_2}$, $\langle n_O + n_{OB_2} \rangle_{\hat{A}_2,\hat{B}_2}$, $\langle n_{A_2} \rangle_{\hat{A}_2,\hat{B}_2}$ and $\langle n_{B_2} \rangle_{\hat{A}_2,\hat{B}_2}$. These depend on both slow species $\hat{A}_2$ and $\hat{B}_2$, because the two operator binding reactions are coupled. This arises from the competition between A_2 and B_2 for the same binding site. In this particular case, as it is shown in Appendix B, solving the master equation for the fast reactions (3.3) and approximating them by the corresponding macroscopic rate equations give the same result, so we will only consider the rate equation approach. Solving for the steady state of Eqs. (3.3), we obtain

$$\begin{aligned} K_D^b \langle n_{OA_2} \rangle^{RE}_{\hat{A}_2,\hat{B}_2} &= \langle n_O \rangle^{RE}_{\hat{A}_2,\hat{B}_2} \cdot \langle n_{A_2} \rangle^{RE}_{\hat{A}_2,\hat{B}_2} \\ K_D^b \langle n_{OB_2} \rangle^{RE}_{\hat{A}_2,\hat{B}_2} &= \langle n_O \rangle^{RE}_{\hat{A}_2,\hat{B}_2} \cdot \langle n_{B_2} \rangle^{RE}_{\hat{A}_2,\hat{B}_2}. \end{aligned} \tag{3.5}$$

Combining this with the fact that in our scheme there is only one DNA copy:

$$n_O + n_{OA_2} + n_{OB_2} = 1 \tag{3.6}$$

we obtain

$$\begin{aligned} a_{\mathrm{A,eff}} &= k_{\mathrm{prod}} \langle n_O + n_{O\mathrm{A}_2} \rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2} \\ &= k_{\mathrm{prod}} \frac{1 + (K_{\mathrm{D}}^b)^{-1} \langle n_{\mathrm{A}_2} \rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2}}{1 + (K_{\mathrm{D}}^b)^{-1} \left(\langle n_{\mathrm{A}_2} \rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2} + \langle n_{\mathrm{B}_2} \rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2} \right)}, \end{aligned} \tag{3.7}$$

and similarly for $a_{\mathrm{B,eff}}$. To find $\langle n_{\mathrm{A}_2} \rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2}$ and $\langle n_{\mathrm{B}_2} \rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2}$ in Eq. (3.7), we combine relations (3.5) and (3.6) with

$$\begin{aligned} n_{\hat{\mathrm{A}}_2} &= n_{\mathrm{A}_2} + n_{O\mathrm{A}_2} \\ n_{\hat{\mathrm{B}}_2} &= n_{\mathrm{B}_2} + n_{O\mathrm{B}_2}. \end{aligned} \tag{3.8}$$

Numerical solution techniques are required here, and we have used the Newton-Raphson method [83].

3.4.2 Coarse-graining protein-protein binding

We now remove instead the protein-protein association and dissociation reactions

$$\begin{aligned} \mathrm{A} + \mathrm{A} &\rightleftharpoons \mathrm{A}_2 \\ \mathrm{B} + \mathrm{B} &\rightleftharpoons \mathrm{B}_2 \end{aligned} \tag{3.9}$$

from our original reaction scheme (3.1). We denote this approach "Eliminating dimerisation (ED)". These interactions are particularly attractive candidates for coarse-graining, since they tend to consume a large fraction of the computational effort when there are significant numbers of free monomers and dimers in the system.

The "fast" species - those whose number is affected by reactions (3.9) - are A, A_2, B and B_2. The "slow species", which will remain in our coarse-grained reaction scheme, are O, $O\mathrm{A}_2$, $O\mathrm{B}_2$ - species from the original scheme which are not affected by reactions (3.9) - together with two new species, $\check{\mathrm{A}}$ and $\check{\mathrm{B}}$, defined by:

$$\begin{aligned} n_{\check{\mathrm{A}}} &\equiv n_{\mathrm{A}} + 2n_{\mathrm{A}_2} \\ n_{\check{\mathrm{B}}} &\equiv n_{\mathrm{B}} + 2n_{\mathrm{B}_2} \end{aligned} \tag{3.10}$$

These new species $\check{\mathrm{A}}$ and $\check{\mathrm{B}}$ are combinations of the fast species whose number remains unchanged by the fast reactions (3.1a). The new, coarse-grained, reaction set with the corresponding propensity functions, is given in Table 3.2. The protein production reactions (3.1c) and (3.1d) now produce the new species $\check{\mathrm{A}}$ and $\check{\mathrm{B}}$. In the original reaction set (3.1), the protein degradation reactions (3.1e) affected only monomers. The corresponding reaction in the new reaction set removes a molecule of the new species $\check{\mathrm{A}}$ and $\check{\mathrm{B}}$ from the system, but with an effective propensity that depends on the average number of monomers that would be obtained by a simulation of the fast reactions, at fixed $n_{\check{\mathrm{A}}}$ or $n_{\check{\mathrm{B}}}$. Similarly,

reactions (3.1b) in the original set, corresponding to the association and dissociation of dimers with the DNA, have been replaced by the association/dissociation of two units of $\check{\mathrm{A}}$ or $\check{\mathrm{B}}$ to O, with an effective propensity proportional to the average number of dimers given by the fast reaction set for fixed $n_{\check{\mathrm{A}}}$ or $n_{\check{\mathrm{B}}}$. Here, the averages required for the effective propensity functions depend on only one slow variable - either $\check{\mathrm{A}}$ or $\check{\mathrm{B}}$ but not both - in contrast to method EO, where the averages depend on both slow variables. This is because the two reactions (3.9) are not coupled to each other: dimerisation of A has no direct effect on the dimerisation propensity of B and vice versa.

We shall test two alternative approaches to the computation of the averages $\langle n_{\mathrm{A}}\rangle_{\check{\mathrm{A}}}$, $\langle n_{\mathrm{A}_2}\rangle_{\check{\mathrm{A}}}$, $\langle n_{\mathrm{B}}\rangle_{\check{\mathrm{B}}}$ and $\langle n_{\mathrm{B}_2}\rangle_{\check{\mathrm{B}}}$ in Table 3.2. In the first approach, which we denote ED1, we make the approximation that these averages correspond to the steady state solutions of the macroscopic rate equations corresponding to (3.9):

$$\begin{aligned} k_b\langle n_{\mathrm{A}_2}\rangle_{\check{\mathrm{A}}} - k_{\mathrm{f}}\langle n_{\mathrm{A}}\rangle^2_{\check{\mathrm{A}}} &= 0 \\ k_b\langle n_{\mathrm{B}_2}\rangle_{\check{\mathrm{B}}} - k_{\mathrm{f}}\langle n_{\mathrm{B}}\rangle^2_{\check{\mathrm{B}}} &= 0 \end{aligned} \tag{3.11}$$

so that

$$\begin{aligned} K_{\mathrm{D}}^d\langle n_{\mathrm{A}_2}\rangle_{\check{\mathrm{A}}} &= \langle n_{\mathrm{A}}\rangle^2_{\check{\mathrm{A}}} \\ K_{\mathrm{D}}^d\langle n_{\mathrm{B}_2}\rangle_{\check{\mathrm{B}}} &= \langle n_{\mathrm{B}}\rangle^2_{\check{\mathrm{B}}}. \end{aligned} \tag{3.12}$$

Relations (3.12) can be used together with the definitions (3.10) to give

$$\begin{aligned} \langle n_{\mathrm{A}}\rangle^{\mathrm{RE}}_{\check{\mathrm{A}}} &= K_{\mathrm{D}}^d\left(\sqrt{8n_{\check{\mathrm{A}}}/K_{\mathrm{D}}^d+1}-1\right)/4 \\ \langle n_{\mathrm{B}}\rangle^{\mathrm{RE}}_{\check{\mathrm{B}}} &= K_{\mathrm{D}}^d\left(\sqrt{8n_{\check{\mathrm{B}}}/K_{\mathrm{D}}^d+1}-1\right)/4. \end{aligned} \tag{3.13}$$

The average numbers of dimers $\langle n_{\mathrm{A}_2}\rangle_{\check{\mathrm{A}}}$ and $\langle n_{\mathrm{B}_2}\rangle_{\check{\mathrm{B}}}$ are given in this approximation by combining (3.13) with (3.12). Method ED1 is expected to give incorrect results when $n_{\check{\mathrm{A}}}$ or $n_{\check{\mathrm{B}}}$ is small, since the macroscopic rate equation approximation breaks down in this limit. This is expected to be a serious problem, because for our genetic switch model, both $n_{\check{\mathrm{A}}}$ and $n_{\check{\mathrm{B}}}$ will be small at the crucial moments when the switch is in the process of flipping between the two steady states [49]. Alternatively, one may solve the master equation corresponding to the eliminated reactions (3.9) to compute the averages. We denote this approach ED2. Numerical solution of this master equation, as described in Appendix A, results in the probability distribution functions $p(n_{\mathrm{A}}|n_{\check{\mathrm{A}}})$, $p(n_{\mathrm{A}_2}|n_{\check{\mathrm{A}}})$, $p(n_{\mathrm{B}}|n_{\check{\mathrm{B}}})$ and $p(n_{\mathrm{B}_2}|n_{\check{\mathrm{B}}})$ for the fast variables, for given values of $\check{\mathrm{A}}$ and $\check{\mathrm{B}}$. These can be used to find $\langle n_{\mathrm{A}}\rangle^{\mathrm{ME}}_{\check{\mathrm{A}}}$, $\langle n_{\mathrm{A}_2}\rangle^{\mathrm{ME}}_{\check{\mathrm{A}}}$, $\langle n_{\mathrm{B}}\rangle^{\mathrm{ME}}_{\check{\mathrm{B}}}$ and $\langle n_{\mathrm{B}_2}\rangle^{\mathrm{ME}}_{\check{\mathrm{B}}}$. These averages are then substituted into the expressions given in Table 3.2 to obtain effective propensities for the coarse-grained SSA simulation. We note that the effective propensity functions for methods ED1 and ED2 in Table 3.2 are identical. The only difference between the two schemes is the way in which the required averages are obtained: using a macroscopic rate equation approximation (ED1) or by numerical solution of the master equation (ED2).

3.4.3 Coarse-graining protein-DNA and protein-protein binding

We now eliminate both protein-DNA interactions [Eq. (3.3)] and protein-protein interactions [Eq. (3.9)]. We will be left with a coarse-grained scheme in which the only fluctuations are due to protein production and degradation. Our "fast reactions" are then (3.3) and (3.9), and our "fast species", whose number is changed by the fast reactions, are in fact *all* the species in the original scheme: O, $O\mathrm{A}_2$, $O\mathrm{B}_2$, A, A_2, B and B_2. Our only slow species, $\tilde{\mathrm{A}}$ and $\tilde{\mathrm{B}}$, are then combinations of the fast species whose number is unchanged in any of the fast reactions:

$$\begin{aligned} n_{\tilde{\mathrm{A}}} &\equiv n_{\mathrm{A}} + 2n_{\mathrm{A}_2} + 2n_{O\mathrm{A}_2} \\ n_{\tilde{\mathrm{B}}} &\equiv n_{\mathrm{B}} + 2n_{\mathrm{B}_2} + 2n_{O\mathrm{B}_2}. \end{aligned} \tag{3.14}$$

$n_{\tilde{\mathrm{A}}}$ and $n_{\tilde{\mathrm{B}}}$ correspond to the total number of A and B molecules in the system. On removal of reactions (3.3) and (3.9), our coarse-grained reaction scheme, given in Table 3.3 under the labels EO-ED1 and EO-ED2, consists simply of a pair of birth-death processes for species $\tilde{\mathrm{A}}$ and $\tilde{\mathrm{B}}$. The effects of the "lost" fast reactions are incorporated via effective rate constants that account for the average number of the relevant fast species expected in a simulation of the fast reaction set, for fixed numbers of the slow species. As in the EO coarse-graining scheme, but not in the ED schemes, the averages here depend upon both slow species, since the fast reactions for A and B are coupled by the shared DNA binding sites.

As for the ED schemes, we consider two alternative ways of obtaining the necessary averages $\langle n_O\rangle_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$, $\langle n_{\mathrm{A}}\rangle_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$, $\langle n_{\mathrm{A}_2}\rangle_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$, $\langle n_{O\mathrm{A}_2}\rangle_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$, $\langle n_{\mathrm{B}}\rangle_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$, $\langle n_{\mathrm{B}_2}\rangle_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$ and $\langle n_{O\mathrm{B}_2}\rangle_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$. Firstly, in approach EO-ED1, we approximate these averages by the steady state solutions of the macroscopic rate equations corresponding to the fast reactions (3.3) and (3.9). Following the same steps as in the previous two Sections (applying equations (3.5) and (3.11)), we arrive at

$$\begin{aligned} \langle n_{\tilde{\mathrm{A}}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}} &= \langle n_{\mathrm{A}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}} + 2(K_{\mathrm{D}}^{d})^{-1}\left(\langle n_{\mathrm{A}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}\right)^2 + \\ &\qquad \frac{2(K_{\mathrm{D}}^{d})^{-1}(K_{\mathrm{D}}^{b})^{-1}(\langle n_{\mathrm{A}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}})^2}{1+(K_{\mathrm{D}}^{d}K_{\mathrm{D}}^{b})^{-1}\left[(\langle n_{\mathrm{A}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}})^2+(\langle n_{\mathrm{B}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}})^2\right]} \\ \langle n_{\tilde{\mathrm{B}}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}} &= \langle n_{\mathrm{B}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}} + 2(K_{\mathrm{D}}^{d})^{-1}\left(\langle n_{\mathrm{B}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}\right)^2 + \\ &\qquad \frac{2(K_{\mathrm{D}}^{d})^{-1}(K_{\mathrm{D}}^{b})^{-1}(\langle n_{\mathrm{B}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}})^2}{1+(K_{\mathrm{D}}^{d}K_{\mathrm{D}}^{b})^{-1}\left[(\langle n_{\mathrm{A}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}})^2+(\langle n_{\mathrm{B}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}})^2\right]}, \end{aligned} \tag{3.15}$$

which can be combined with relations (3.14) and inverted numerically to give $\langle n_{\mathrm{A}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$ and $\langle n_{\mathrm{B}}\rangle^{\mathrm{RE}}_{\tilde{\mathrm{A}},\tilde{\mathrm{B}}}$ [83]. The other averages required in Table 3.2 are obtained using relations (3.5), (3.6) and (3.12). Approach EO-ED1 is approximate, since it assumes that the average numbers of molecules that would be produced by long stochastic simulations of the fast

Reaction	Propensity	Reaction	Propensity
$A + A \rightleftharpoons A_2$	$k_f n_A(n_A-1)$, $k_b n_{A_2}$	$B + B \rightleftharpoons B_2$	$k_f n_B(n_B-1)$, $k_b n_{B_2}$
$O + A_2 \rightleftharpoons OA_2$	$k_{on} n_O n_{A_2}$, $k_{off} n_{OA_2}$	$O + B_2 \rightleftharpoons OB_2$	$k_{on} n_O n_{B_2}$, $k_{off} n_{OB_2}$

Table 3.3: Reaction scheme for the preliminary simulations to compute the effective propensity functions given in Eqs. (3.16) and (3.17), for scheme EO-ED2.

reaction set are given by the steady-state solutions of the corresponding macroscopic rate equations. This approximation is avoided in approach EO-ED2, in which we calculate the averages of the fast variables A, A_2, OA_2, B, B_2, OB_2 and O from the master equation corresponding to the coupled reactions (3.3) and (3.9). This is difficult to do directly (as in scheme ED2), so we use simulations. We carry out a series of short preliminary simulations, using the SSA, of reactions (3.3) and (3.9), for fixed values of $n_{\tilde{A}}$ and $n_{\tilde{B}}$. The reaction scheme for these preliminary simulations is given in Table 3.3. From these, we compute the averages required for the effective propensities

$$a_{A,\mathrm{eff}} = k_{\mathrm{prod}} \left(\langle n_O + n_{OA_2} \rangle^{\mathrm{ME}}_{\tilde{A},\tilde{B}} \right) \tag{3.16}$$

and

$$\mu_{A,\mathrm{eff}} = \mu \langle n_A \rangle^{\mathrm{ME}}_{\tilde{A},\tilde{B}}. \tag{3.17}$$

These propensities are stored in a lookup table, which is referred to during the coarse-grained simulations of the slow variables.

3.5 Results

We now assess the performance of the various coarse-graining approaches, in terms of how much they speed up the simulations, and how accurately they reproduce the behaviour of the full system. Key features of the behaviour of this system are its bimodal steady state probability distribution function and its spontaneous flips between the two stable states. In this Section, we assess how well the coarse-grained simulations reproduce the bimodal probability distribution for the difference in total number between the A and B proteins, which we denote λ:

$$\lambda = n_A + 2n_{A_2} + 2n_{OA_2} - (n_B + 2n_{B_2} + 2n_{OB_2}) \tag{3.18}$$

We also test how well the various schemes reproduce the rate of fluctuation-induced switching between the A- and B-rich states, which we measure using the Forward Flux Sampling (FFS) rare event simulation method [42, 43, 44]. We compare our results to SSA simulations of the full reaction set (3.1) which we denote ORN ("Original Reaction Network"). The coarse-graining schemes are based on the assumption that the "fast"

reactions are indeed fast compared to the slow reactions. We therefore expect their accuracy to improve as the rate constants for the "fast" reactions increase. We will see that the accuracy also depends on which are the "fast" reactions, and on how we compute the averages needed for the propensity functions.

3.5.1 Computational Performance

We begin by running the system for a fixed simulation time with the various coarse-graining approximations, and measuring the computational speed-up. The coarse-graining procedure itself takes a negligible time in all schemes but EO-ED2; in this last case, making the lookup table is done with a separate code, which runs for 6-24 hours. The speed increase depends on the values that we choose for the rate constants. Table 3.4 shows the CPU time, in seconds (on an AMD Athlon 1600+ processor), required for a simulation run of length $10^5 k_{\rm prod}^{-1}$, for various values of the rate constants $k_{\rm on}$ for protein-DNA binding and $k_{\rm f}$ for dimerisation. In all cases, $k_{\rm off}$ and $k_{\rm b}$ are adjusted so as to keep the equilibrium constants $K_{\rm D}^d$ and $K_{\rm D}^b$ fixed. Considering the full reaction network (ORN) - top row in the table - we observe that the CPU time is much more sensitive to the dimerisation rate than to the protein-DNA binding rate. This shows that the SSA mostly executes monomer-monomer association and dimer dissociation reactions (reactions that are likely to have the largest propensities), even when $k_{\rm on}$ is much greater than $k_{\rm f}$. This is because the propensity for dimerisation depends on (roughly) the square of the number of free monomers, which is generally quite large. Protein-protein association/dissociation is therefore the performance bottleneck for this system. Bearing this in mind, it is not surprising that when we consider the next row in Table 3.4, we see that removing protein-DNA association and dissociation reactions (EO), is only useful when the rate constants for these reactions are exceedingly large. Eliminating the protein-protein association and dissociation reactions (ED1 and ED2) results in a dramatic speed-up compared to the ORN case (rows 3 and 4). This speed-up is most impressive when the dimerisation rate is high. There is no significant difference in the CPU time required between the ED1 and ED2 methods. When we eliminate both protein-DNA and protein-protein association and dissociation (bottom two rows in Table 3.4), we obtain a further factor 2-25 decrease in CPU requirement. Again, there is no significant difference in CPU time between methods EO-ED1 and EO-ED2. We therefore conclude that, in the physiological parameter range, some computational speed-up can be obtained by removing protein-DNA binding reactions; however, much more computer time can be saved by coarse-graining protein-protein association and dissociation reactions.

3.5.2 Steady-state probability distribution

We now compute the steady-state probability distribution function $P(\lambda)$ for the difference λ between the total number of A and B molecules, as given by Eq. (3.18). We expect $P(\lambda)$ to have two peaks around the known stable steady states $\lambda = \pm 27$ [49], and a "valley" around the unstable steady state $\lambda = 0$. To compute $P(\lambda)$, we carry out a long SSA

	$k_f = 5$ $k_{on} = 5$	$k_f = 100$ $k_{on} = 5$	$k_f = 5$ $k_{on} = 100$	$k_f = 100$ $k_{on} = 100$
ORN	5.81	113	5.55	118
EO	5.25	103	5.08	70.7
ED1	0.18	0.19	1.94	1.94
ED2	0.18	0.19	1.92	1.91
EO-ED1	0.085	0.082	0.081	0.081
EO-ED2	0.083	0.082	0.081	0.084

Table 3.4: CPU time (in seconds) required to simulate the system for $t_{sim} = 10^5 k_{prod}^{-1}$, for different parameter sets. The coarse-graining procedure takes a negligible time in all schemes but EO-ED2; in this last case, the lookup table is made by a separate code that runs for 6-24 hours. Simulations were performed on an AMD Athlon 1600+ processor. The dissociation rates were scaled such that the equilibrium constants for dimerisation and operator binding were kept constant at $K_D^d = 1/5$ and $K_D^b = 1$.

simulation, during which we compile a histogram for the probability of finding the system at each λ value. This procedure is repeated for all the coarse-grained simulation schemes in Table 3.2. However, it is hard to achieve good sampling of $P(\lambda)$ in the "valley" region close to $\lambda = 0$, where the system is very unlikely to be found. In this region we use the FFS method to compute $P(\lambda)$ more accurately [51]. This method is described briefly in the following Section and in Section 1.6.4. Results are shown in Figure 3.3, for the "baseline" parameter set given in Section 3.1. As expected, $P(\lambda)$ is clearly bimodal and shows symmetric peaks flanking a "valley" at $\lambda = 0$. The location of the peaks and valley correspond to the stable and unstable solutions of a mean field analysis [49] of the switch. Comparing the results for the different coarse-graining schemes in Figure 3.3, we see that they all appear to reproduce $P(\lambda)$ quite well, giving the correct position, height and width of the peaks. Inset A magnifies the left probability peak, showing that the only methods displaying a small systematic error are ED1 and EO-ED1, *i.e.* the coarse-graining schemes relying on the solutions of the Macroscopic Rate Equation. In general, we can conclude that all the methods reproduce $P(\lambda)$ rather well in the peak regions. However, when we investigate in more detail the results for the "valley" region around $\lambda = 0$, clear differences are observed between the coarse-graining methods. Inset B of Figure 3.3 shows on a logarithmic scale the results for $P(\lambda)$ in this region, generated using the FFS method. All the coarse-graining methods deviate from the results of the full reaction network (ORN). The apparent effect of removing dimerisation (ED1/ED2) is to shift the minimum up in probability, with the macroscopic rate equation approach (ED1) having a stronger effect. Removing operator state fluctuations (EO) has the opposite effect, shifting the minimum down in probability. Methods EO-ED1 and EO-ED2 appear to show a combination of these two effects. Although these deviations from the ORN results are small, they will turn out to be rather important for the dynamical switching behaviour to be discussed in

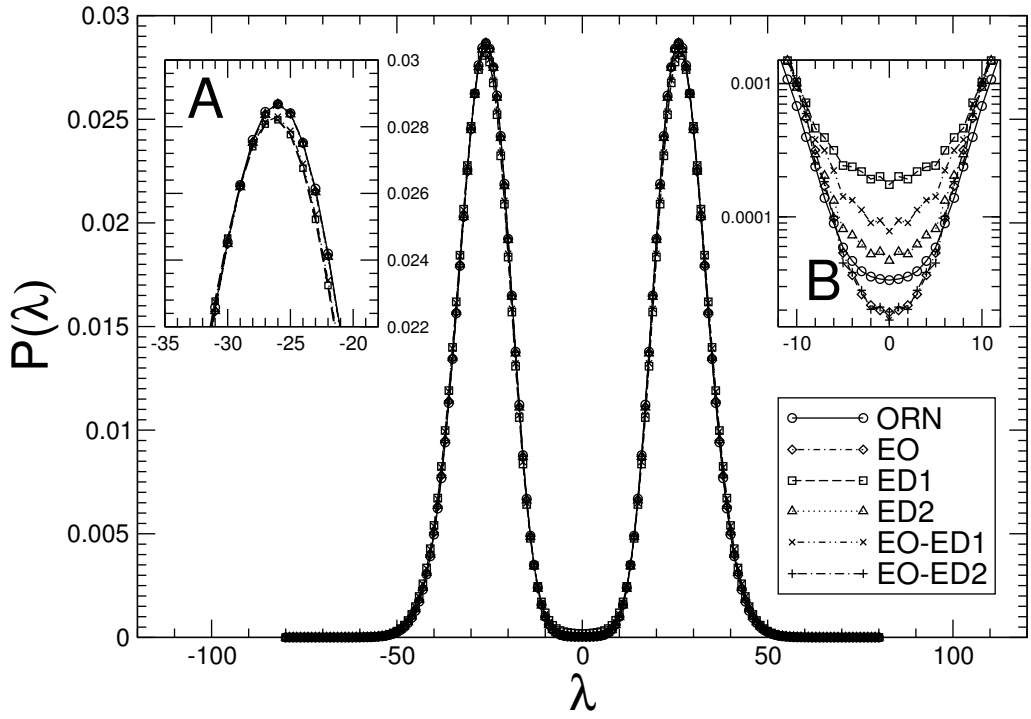


Figure 3.3: Probability distribution $P(\lambda)$ as a function of the order parameter λ. Inset A zooms in on the left peak and shows how all the methods are able to reproduce the positions and heights of the peaks. Inset B shows, on a logarithmic scale, deviations between the different coarse-graining schemes and the original network for the region around $\lambda = 0$. FFS was used to sample $P(\lambda)$ in this region.

the next Section. However, if one is only interested in the steady state distribution, the choice of the particular coarse-graining method does not appear to be crucial.

3.5.3 Rate of stochastic switch flipping

In many cases, fluctuation-driven dynamical properties are an important output of a simulation of a biochemical network. This is especially true of genetic switches, where a key characteristic is the rate of flipping between stable states (as shown in Figure 3.2 for the model genetic switch). When simulating these systems, one requires not only an accurate representation of the steady-state distribution, but also of the dynamical behaviour of the system. We now test whether the various coarse-graining methods are able to reproduce the correct rate of stochastic flipping of the model switch. This is a particularly stringent test, since this fluctuation-driven process is likely to be highly sensitive to the accuracy with which dynamical fluctuations are reproduced in the different schemes.

To measure the rate k_{AB} of stochastic switch flipping, we use the FFS method [42, 43, 44], which allows the calculation of rate constants and the sampling of transition paths for

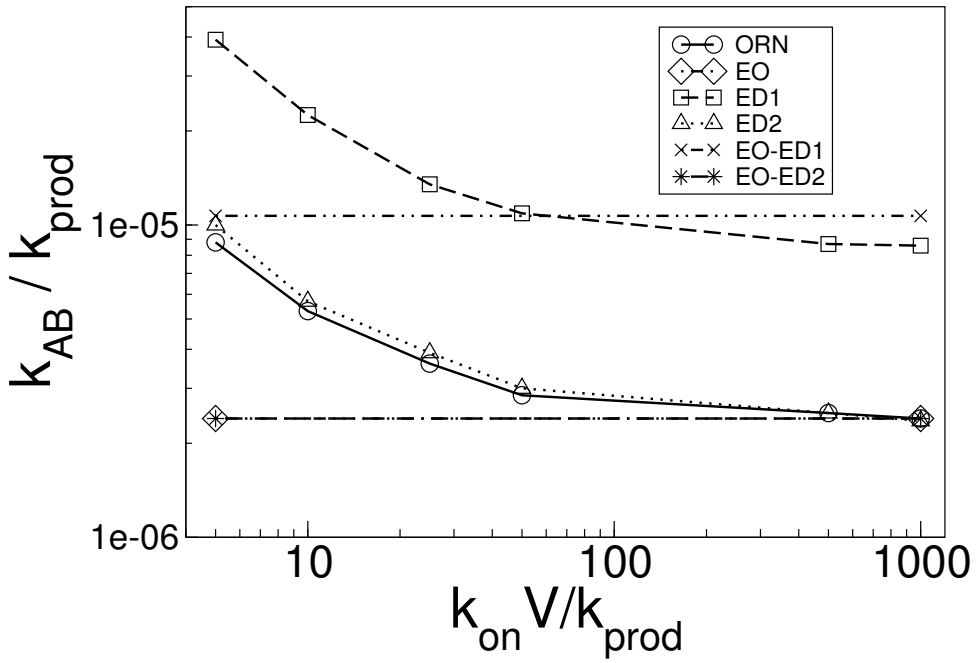


Figure 3.4: Switch flipping rate k_{AB} as a function of the dimer-DNA association rate k_{on}, adjusting k_{off} so that the equilibrium constant for DNA binding remains unchanged. The switch is more stable when operator binding/unbinding is rapid, suggesting that fluctuations in these reactions play an important role in the switch flipping. Methods that remove these reactions yield a straight line in the Figure; among those, only methods EO and EO-ED2 are able to capture the asymptotic behaviour of the curve for $k_{\mathrm{on}} \to \infty$. Method ED2 always gives a good approximation of the rate, while ED1 consistently overestimates it by about one order of magnitude.

rare events in stochastic dynamical systems. FFS is described in more detail in Section 1.6.4, and in Refs. [42, 43, 44]. The method can also be used to obtain the steady state probability distribution as a function of the λ parameter, as in inset B of Figure 3.3 [51].

Figure 3.4 shows the switch flipping rate k_{AB}, as a function of the dimer-DNA association rate k_{on}. The dimer-DNA dissociation rate is adjusted to keep $K_{\mathrm{D}}^{d} = 1$. The other parameters are fixed at $k_{\mathrm{f}} = 5k_{\mathrm{prod}}$, $\mu = 0.3k_{\mathrm{prod}}$ and $K_{\mathrm{D}}^{b} = 1/5$. For the full reaction network (ORN; solid line), the flipping rate decreases as DNA binding becomes faster, flattening for very fast ($k_{\mathrm{on}} > 500k_{\mathrm{prod}}$) operator association/dissociation. The switch is more stable (*i.e.* its spontaneous switching rate k_{AB} is lower) when operator binding/unbinding is rapid, suggesting that fluctuations in these reactions play an important role in switch flipping. For the EO method, in which protein-DNA association/dissociation reactions are "lost", the flipping rate does not depend on k_{on} (since only the equilibrium constant K_{D}^{b} features in this method and this is kept constant). As expected, the flipping rate for the EO method corresponds to the ORN result in the limit of large k_{on}. When we coarse-

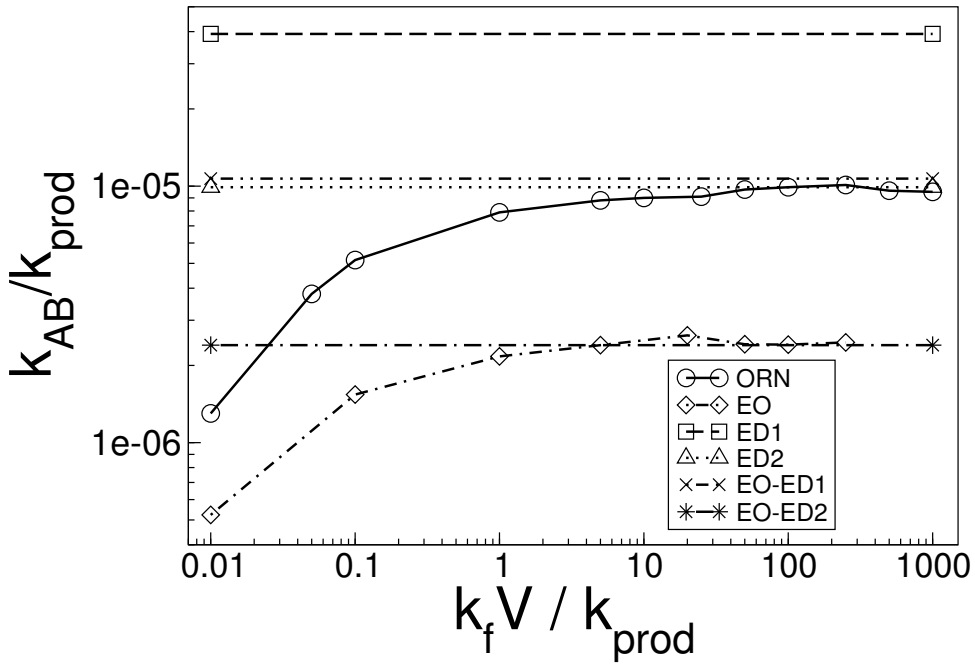


Figure 3.5: Switch flipping rate k_{AB} as a function of the protein-protein association rate k_{f}, adjusting k_{b} so that the equilibrium constant for dimerisation remains unchanged. For the full reaction set, the switching rate increases with the rate of dimerisation. The EO curve consistently underestimates the rate by approximately an order of magnitude. The methods that remove the dimerisation reactions yield a constant line in Figure. Among them, only methods EO-ED1 and EO give good results for high k_{f}, although for the latter we suspect this comes from a lucky cancellation of errors.

grain over protein-protein interactions (ED1 and ED2), our results are very dependent on whether the macroscopic rate equations or the master equation is used to compute propensity functions. For the rate equation approach (ED1), the decrease in k_{AB} with k_{on} is reproduced, but the switch is almost an order of magnitude less stable than for the full reaction set. However, when the chemical master equation is used to compute the propensities, the results are remarkably accurate - the ED2 approach gives switch flipping rates in good agreement with the full reaction set. The two methods EO-ED1 and EO-ED2, which coarse-grain over both DNA binding and dimerisation reactions, also show this behaviour: for EO-ED1, where the rate equation approximation is used, the switch flipping rate is similarly almost an order of magnitude too high, whereas for EO-ED2, where the master equation is used, k_{AB} is indistinguishable from that given by method EO (where dimerisation is simulated explicitly). These results show that dimer-DNA binding plays an important role in switch flipping for association/dissociation rates in the physiological range, and that, for reliable coarse-graining, effective propensities need to be computed with the master equation rather than the macroscopic rate equation approximation.

Figure 3.5 shows the equivalent result when the monomer-monomer association rate k_f is varied, adjusting k_b so that $K_D^d = 1$. The other parameters are $k_{on} = 5k_{prod}$, $\mu = 0.3k_{prod}$ and $K_D^b = 1/5$. For the full reaction set (ORN), k_{AB} increases with k_f: as the dimerisation reactions become faster, the switch becomes less stable. This is in contrast to the behaviour observed in Figure 3.4. It appears that switch flipping is hindered by fluctuations in the monomer-dimer reactions. This apparently somewhat counter-intuitive result can perhaps be explained as follows: protein is produced in the monomer form. To flip the switch, it needs to dimerise and bind to the operator. If dimerisation is slow, the monomer may be degraded before it has a chance to dimerise, and in this case it does not contribute to flipping the switch. On the other hand, if dimerisation is fast, then every monomer that is produced makes a contribution to the dimer pool and can potentially bind to the operator, leading to switch flipping. The EO approach (eliminating dimer-DNA binding) shows the same increase in k_{AB} with k_f, but underestimates the value of k_{AB} by about an order of magnitude. This supports our view that fluctuations in operator binding are important for switch flipping. On eliminating dimerisation fluctuations (ED1 and ED2), we observe the same problem with the macroscopic rate equation approximation as in Figure 3.4 - ED1 produces a flipping rate that is too high, while ED2, where the master equation is used, gives good agreement with the full reaction set (ORN) in the high k_f limit. When both protein-protein and protein-DNA association/dissociation reactions are eliminated, method EO-ED2 gives results in agreement with EO in the high k_f limit. Method EO-ED1 gives unexpectedly good results, in fairly close agreement with ED2 and ORN. However, given that we expect removing DNA binding to reduce the rate constant, while using the macroscopic rate equation approximation increases it, this is likely to be just a lucky cancellation of errors for this particular parameter set. Figure 3.5 therefore demonstrates that fluctuations in the monomer-monomer association/dissociation reactions actually disfavour switch flipping. Moreover, as for Figure 3.4, we see that the macroscopic rate equation approximation is not reliable for predicting switch flipping rates, while coarse-graining over the dimerisation reactions using the master equation approach (ED2) becomes reliable when $k_f > 5k_{prod}$ (for this parameter set).

We have also tested the various coarse-graining approaches for the case where both protein-protein and protein-DNA association/dissociation reactions are fast ($k_f = 100k_{prod}, k_{on} = 100k_{prod}$). In this case, as expected, methods EO, ED2 and EO-ED2 all give flipping rates in agreement with the full reaction set (ORN), while ED1 and EO-ED1 do not. This indicates once again that in general the macroscopic rate equation approach is not suitable for computing switching rates.

3.6 Discussion

Understanding cellular control systems is likely to require the study of very complex biochemical reaction networks. Computer simulations clearly have an important contribution to make in this area, since they can provide quantitative understanding of how biochemical networks work. It is clear that in many cases (including the understanding

on gene regulation), stochastic simulations are required. However, the more reactions a biochemical network has, the more computationally expensive it is to simulate. Eliminating fast reactions will be essential for simulating biochemical networks of the scale and complexity that is relevant for biology. It is therefore very important to understand how this can be done reliably, while preserving the correct dynamical features of the full reaction network. In this Chapter, we have made a systematic study of the computational speedup and accuracy of a range of coarse-graining schemes, for a model gene regulatory network. All gene regulatory networks involve protein-protein and protein-DNA interactions. These tend to be rapid in comparison to protein production (transcription, translation and folding) and removal from the cell (active degradation and dilution due to growth and division). We try to address the general question of what the consequences are of eliminating protein-protein or protein-DNA association and dissociation reactions from stochastic simulations of gene regulatory networks. We use as our case study a bistable genetic switch, since this gives us a very sensitive readout, in the form of the switch flipping rate, of the accuracy with which dynamical fluctuations are reproduced by the various coarse-graining schemes. We hope that our results will prove relevant to simulations of real genetic switches and gene regulatory networks in general.

To coarse-grain the reaction scheme for the model genetic switch, within the context of Gillespie's Stochastic Simulation Algorithm (SSA), we have used the approach described by Bundschuh *et al.* [68]. Here, the reaction set is divided into "fast" and "slow" reactions. Chemical species whose number is changed by the fast reactions are designated "fast". A set of "slow" chemical species is constructed, which consists of the original species that were unaffected by the fast reactions, together with new species, formed from linear combinations of the fast species, such that their number is unaffected by the fast reactions. The slow reactions are then rewritten in terms of the set of slow species, with effective propensity functions that depend on averages (and in some cases variances) of the fast reaction set, for fixed numbers of molecules of the slow species. These averages may be obtained by explicit or numerical solution of the chemical master equation for the fast reactions. Alternatively, one may make the approximation that the averages are well represented by the steady-state solutions of the corresponding macroscopic rate equations for the fast reactions. In either case, having computed the effective propensity functions, one simply implements the SSA for the slow reaction set, propagating the set of slow variables, using these effective propensities.

For the model genetic switch, we investigated the effects of eliminating protein-protein association/dissociation reactions, and/or protein-DNA association/dissociation reactions, from the full reaction set. We also compared the macroscopic rate equation approximation to the master equation approach for computing the effective propensities. Using all the coarse-graining schemes, we computed the steady-state probability distribution as well as spontaneous switch flipping rates. We found that all the coarse-graining methods gave good agreement with the full reaction network for the steady-state probability distribution, although small deviations were observed around the unstable steady state. However, dramatic differences were observed in the switch flipping rates computed using the different

coarse-graining schemes. Elimination of protein-DNA association/dissociation increased the stability of the switch (but agreed with the full reaction set in the fast reaction limit). In contrast, elimination of protein-protein association/dissociation decreased the stability of the switch (again, agreeing with the full reaction set in the fast reaction limit). However, over most of the range of parameters tested, protein-protein association/dissociation reactions can be eliminated with a minimal effect on switching rates, and with the advantage of an impressive computational speed-up. This result is likely to prove very useful when simulating complex and computationally expensive networks. The implications of these observations for the physics of the switching mechanism for this model switch are be investigated in Chapter 2.

We also observed that the macroscopic rate equation approximation does not produce reliable switching rates, even though the steady-state probability distribution is reasonably well reproduced. Typically, switching rates computed using the macroscopic rate equation approximation are an order of magnitude too high, even in the limit of fast reactions. In contrast, when the chemical master equation is used to compute the effective propensities, results are in excellent agreement with the full reaction set for fast reaction rates. This result serves as a warning that care must be taken in how coarse-graining is applied. As an example, the lysogeny-lysis switch of bacteriophage λ is extremely stable to fluctuations [84, 53], a fact that computational modelling (using macroscopic approximations) has thus far been unable to satisfactorily explain [85, 53]. In such a case, careful coarse-graining is crucially important.

Our results show that under certain biologically relevant conditions fast reactions can be eliminated while preserving the correct dynamical characteristics of the system, even when highly sensitive fluctuation-driven quantities such as switch flipping rates are considered. This is very encouraging for the simulation of more complex reaction networks, and it would be interesting to apply these approaches to more complicated genetic switches, and also to other gene regulatory networks where dynamical fluctuations are important. We hope that this work will be of use as a "tutorial" in designing and implementing coarse-graining schemes, and that it may aid in pointing the way to accurate and efficient coarse-grained simulations of a wide variety of interesting and important biochemical networks.

Chapter 4

The bacteriophage λ genetic switch

Stability is not immobility.
Klemens von Metternich

Genetic networks allow a cell to respond to different environmental conditions, and to take decisions accordingly. Some networks are able to establish a stable epigenetic state and to maintain it in the cell for many generations. The cell then can "choose" among a small number of alternative states. The case of two states is most typical. Systems showing two alternative stable states are called bistable. One of the best-characterised examples of a bistable genetic network is the bacteriophage λ genetic switch: a regulatory circuit which allows the phage to maintain either of two alternative stable states, and to reliably switch from one to the other upon a change in the environmental conditions. One state, where the phage stays dormant in the host DNA (lysogenic) exhibits extreme stability. The reasons for this stability are not yet completely understood, given that the system is continuously subjected to stochastic fluctuations which might flip the switch. Several models, based on equilibrium assumptions [53, 86], have failed to explain the measured spontaneous switching rate, which is less than 10^{-9} per generation per cell. In this work, we design a fully stochastic model aiming to explain the stability of the lysogenic state of bacteriophage λ. The model, based on a set of chemical reactions, describes the dynamics of transcription factors binding to the O_R operator and exploits the Forward Flux Sampling technique to measure the spontaneous switching rate between the two stable states of the system. In order to speed up the simulations, we apply a recently-developed approximation scheme that dynamically integrates out fast reactions, and we obtain results compatible with the available literature data. Furthermore, we investigate the effects of macromolecular crowding and of DNA looping on the system, and we find that both mechanisms could increase the stability of the lysogenic state.

4.1 Introduction

Genetic networks consist of genes whose protein products regulate the expression of other genes via activation or repression mechanisms. These proteins, called transcription factors (TFs) typically bind to the DNA upstream of the gene they regulate and either block the access of RNA polymerase (RNAp) to the gene promoter or help the RNAp to unwind the DNA and initiate transcription. Some networks exist stably in either of two different states, which are inherited by the cell progeny and propagated into the cell population. These states are called epigenetic, because they are not directly encoded in the genome itself, but rather in the patterns of regulation to which the genes are subjected. Even very simple organisms can display very stable epigenetic states in which one state of the switch is extremely stable, despite the cell being prone to strong stochastic fluctuations, which might lead to an accidental flip. These fluctuations come from the intrinsic stochasticity of biochemical reactions and can give rise to effects that cannot be accounted for at a mean-field level, but require solution of the chemical master equation of the system. In this Chapter, we focus on one of the best characterised bistable systems in biology, the bacteriophage λ genetic switch.

Bacteriophage λ is a virus which infects the bacterium *E. coli*. After inserting its DNA into a host cell, it takes the decision to enter either of two alternative pathways, called lytic and lysogenic, which lead to different behaviour. In the lytic case, the phage uses the biochemical machinery of the host to replicate as much as possible, thereby killing the bacterium and releasing its progeny. In the lysogenic state, the virus integrates its DNA into the bacterial chromosome, and stays dormant for a large number of host generations. The phage reliably switches from the lysogenic to the lytic state upon UV damage of the bacterial genetic material (an event called prophage induction). The two states of the phage have been demonstrated to be stable [87], *i.e.* to be robust against variations in the environmental conditions. The system can then be considered a bistable genetic switch. Spontaneous flipping events are extremely rare, with a rate that has been estimated to be not higher than 10^{-9} per generation per cell [84], which is extremely low (lower than the rate of spontaneous DNA mutations). Two transcription factors (TF), called cI and cro, regulate the early stages of the switching mechanism, and are ultimately responsible for its extreme stability. As depicted in Figure 4.1, the genes encoding these two proteins are divergently transcribed, and share a regulatory region of the DNA, called the O_{R} *operator*. Both cI and cro can bind, in homodimer form, to the operator and repress the production of either protein. Protein cI is present in high concentrations when the phage is in the lysogenic state, whereas cro is abundant in the lytic state. On the DNA, in the region between the two genes, 3 binding sites, each spanning 17 base-pairs, are found (see Figure 4.1). cI binds preferably to O_{R1}, thereby blocking the production of cro. At higher concentrations, cI can also bind cooperatively to O_{R2}, thus recruiting an RNA polymerase molecule and enhance its own production. Finally, at very high concentrations, cI binds to O_{R3}, repressing its own production, thus preventing the build up of an excessive concentration in the cell. Cro shows similar behaviour (with the notable absence of cooperative interactions), with reversed affinity to the operator sites. The two

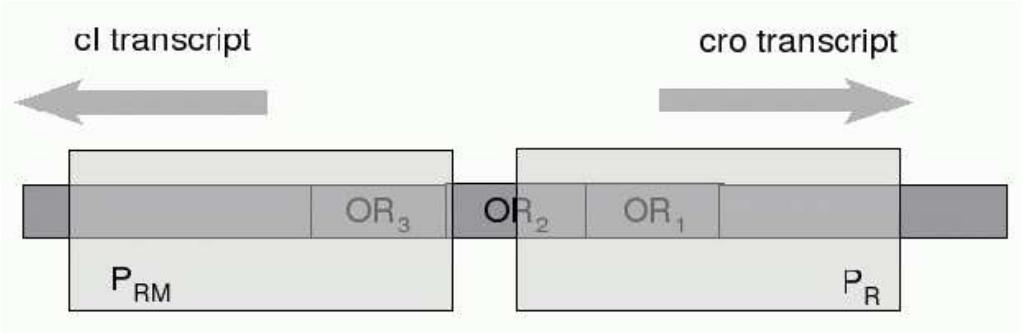


Figure 4.1: Pictorial representation of the phage λ binding sites on the O_R operator. The *cI* and *cro* genes are transcribed divergently and share the O_R regulatory region. cI and cro proteins can bind, in homodimer form, to three distinct binding sites on the operator. cI has higher affinity to O_{R1}, and binds less tightly to O_{R2} and O_{R3}; conversely, the affinity of cro is higher for O_{R3}. The promoters for the two genes asymmetrically overlap with the operator binding sites; their asymmetry has been exaggerated in the cartoon, to reflect the interference between the binding of RNA polymerase and the transcription factors.

genes mutually repress each other, which enables the existence of two stable states rich in either species; only a rare fluctuation in the number of the minority species can trigger the flipping of the switch. A combination of positive and negative regulation is exploited to prevent this undesirable event.

The λ system has been studied thoroughly over the last 60 years, and a large amount of data is available. For our purposes, the binding energies of cI and cro to O_R have been accurately measured, together with their oligomerization energies [79, 81, 82, 88]. The O_R operator is commonly accepted to be the core of the genetic network determining the bistability of the system. Our model focuses then on this region, neglecting the role of other λ genes that intervene later in the lytic pathway, or are activated when the virus first infects the host. We consider the phage genome to be already integrated in the host chromosome and we predict the behaviour of the system in the two stable epigenetic states, and how often the switch will spontaneously flip in either direction. Several other modeling attempts have taken the same approach [53, 86, 89]. However, all these models rely on the key assumption of rapid equilibrium for the operator binding and dimerisation reactions. Notably, none of these models is able to correctly reproduce the measured spontaneous switching rate from the lysogenic to the lytic state, all overestimating it by at least 1-2 orders of magnitude. We argue that the stochastic fluctuations neglected in the above-mentioned models need to be considered in order to achieve a faithful representation of the system and could play a crucial role in the stability of the switch.

When pondering the faithfulness of the model in representing reality, we should realise that the cellular environment is packed with macromolecules and cellular structures. Reaction equilibria "feel" the consequent "lack of space" and achieve steady states with different concentrations of reactants and products than would be expected for *in vitro*

dilute conditions. This effect is ubiquitous in cells, yet few studies pay attention to it. We argue that macromolecular crowding could further stabilise the lysogenic state, by favouring compact structures such as TF-DNA complexes.

Recent experiments have highlighted the role of another, distant operator called O_L in maintaining the stability of the lysogenic state. cI dimers bound to O_L could interact with the dimers bound to O_R, locking the DNA in a looped state and reinforcing the repression of the P_R promoter [90, 91, 92]. We include the effects of crowding and of the O_L operator in our model, and we find that both can help in increasing the stability of the lysogenic state.

The rest of the Chapter is organized as follows: in Section 4.2, we describe our stochastic model, and we discuss the choice of parameters: many of these are available (sometimes already for decades!) in the literature, yet we have nevertheless made some assumptions, which will be justified. In the following Section, we briefly describe the numerical methods used to solve the model at the level of the chemical master equation, and to compute the switching rate, while in Section 4.4 we explain the results obtained. As simulations of large reaction networks are computationally demanding, we use the coarse-graining procedure described in Chapter 3 to integrate out some of the "fast" reactions, preserving the fluctuations exploited by the system in the switching process. Furthermore, the model is modified to probe its validity on some relevant mutant phages. In Section 4.5, we investigate the effects of macromolecular crowding on the bistability of the system, and we find that excluded volume interactions could be exploited by the phage to increase the stability of the lysogenic state. Finally, in Section 4.6, we extend our model to include the O_L auxiliary operator, which is thought to interact with O_R via a long DNA loop.

4.2 Model

In Chapter 2, we considered the dynamics of a model switch, loosely based on bacteriophage λ, and we pointed out how dynamical properties of the system, such as the switching rate, are fluctuation-driven and could depend crucially on the individual rate constants of the reactions involved. Therefore, in order to properly study the dynamics of a genetic switch and predict a correct switching rate, a fully stochastic model is required. In the case of bacteriophage λ, the large body of data available in the literature made it possible to obtain measured values for most of the necessary parameters for the detailed reaction set we describe below, leaving only few quantities to be estimated.

We model the system as a set of chemical reactions between transcription factors and DNA, of which the former can form homodimers or bind to specific sites on the DNA, here represented as distinct chemical species. The model is designed to describe the core of the bacteriophage λ genetic switch, *i.e.* the dynamics of the transcription factors cI and cro. In the following, a brief description of the model and of how the reaction rates are obtained will be given.

Both transcription factors can bind to the operator region O_R. As we have discussed

in Section 4.1, in the short piece of DNA (82 base-pairs) stretching between the two divergently-transcribed genes *cI* and *cro*, there are multiple binding sites: three for the cI and cro dimers, and two for RNA polymerase. Binding sites for RNA polymerase are called *promoters* and asymmetrically overlap with the binding sites for TFs: the promoter for *cI*, P_{RM}, overlaps with O_{R3}, and marginally with O_{R2}, while the promoter for *cro*, P_{R}, overlaps with both O_{R1} and O_{R2}. An RNAp molecule bound to P_{RM} therefore blocks only the binding of a TF to O_{R3}, whereas when RNAp is bound to P_{R}, TFs can bind to neither O_{R2} nor O_{R1}. Therefore, there are in total 40 possible occupation states for the operator. Our scheme contains a reaction for every possible transition among these states. The most recent measurements of the free energies of binding, to our knowledge, were obtained by Darling *et al.* [79]. These data correct the early measurement of Shea and Ackers [89], adding a weak cooperativity between sites occupied by cro dimers. However, these binding free energies do not suffice to define our model, because they do not provide any information about the time scales of the binding and unbinding events—only their ratio. We thus have to make another assumption, namely that the binding of a TF molecule to the DNA is diffusion-limited [1]. The forward reaction rate of is then set to the Smoluchowski rate $k_{\mathrm{ass}} = 4\pi DL$, where D is the diffusion constant of the protein and L is the linear protein size. cI and cro molecules have about average size and weight for small proteins in bacteria, and we therefore assume for them the typical values of $D = 5\mu\mathrm{m}^2/\mathrm{s}$ [93], L=5nm [94], which gives k_{f}=$0.314\mu\mathrm{m}^3/\mathrm{s} = 0.188(\mathrm{nM})^{-1}/\mathrm{s}$. Combining this association rate with the free energies of binding, we can extract dissociation rates for each binding site:

$$k_{\mathrm{diss}} = k_{\mathrm{f}}[V_{\mathrm{cell}}(l)/V_{\mathrm{cell}}(\mu m^3)]\exp(\Delta G/RT)\,, \tag{4.1}$$

where the factor $V_{\mathrm{cell}}(\mathrm{l})/V_{\mathrm{cell}}(\mu\mathrm{m}^3) = 6.023\cdot 10^8$ converts the standard volume of 1 liter into $\mu\mathrm{m}^3$ units, and T is assumed to be 310K, corresponding to the physiological temperature of 37 °C. The volume conversion factor is needed because the standard free energy ΔG is measured with reference to a concentration of 1 mol/liter [95].

The dimerisation reactions for cI and cro are explicitly simulated. The most recent data we are aware of measure the free energy of dimerisation to be about -11 kcal/mol for cI [81, 88, 82] and -8.7 kcal/mol for cro [81]. At a temperature of 37 °C, the equilibrium dissociation constants are, respectively, $K_{\mathrm{D}} = 15\,\mathrm{nM}$ for cI and $K_{\mathrm{D}} = 740\,\mathrm{nM}$ for cro. Cro dimers appear then to be much more unstable than cI dimers, in contrast with what was assumed in early papers [96, 97, 89, 27]. Recently, Jia *et al.* [98] have measured the dissociation rate of cro dimers *in vitro* to be about $0.02 - 0.04\,\mathrm{s}^{-1}$ using FRET. However, as will be explained more clearly in the next Sections, we believe that the conditions in the cell can have a major effect on these time scales, and therefore, analogous to the binding reactions, we use a diffusion-limited association rate for the dimerisation reactions (this time using the relative diffusion constant between two diffusing molecules $D = D_1 + D_2$), and we obtain dissociation rates from Eq. (4.1).

The production of transcription factors is modelled as a two-step process leading first to the synthesis of a messenger RNA (mRNA) transcript of the corresponding gene, and later to its translation into a protein. In order to start the whole process, an RNA poly-

merase molecule must first bind to a promoter, which must not be blocked by a transcription factor bound to one of the overlapping operators. The RNAp then forms a DNA-bound state, called *closed complex*, from which it can either dissociate or unwind the DNA to start transcription of the gene, forming an *open complex*. Once the open complex is formed, transcription proceeds irreversibly [99] and eventually leads to the production of an mRNA molecule. We combine the many stages of the transcription process into an effective mRNA production reaction [53], assuming that the rate of mRNA synthesis is proportional to the frequency of transcription initiation [89] and neglecting details such as gene length, efficiency of the transcriptional stop signal and promoter-specific number of abortive inititations. The P_{RM} promoter is much weaker than P_{R}, *i.e.* its open complex formation rate is lower. However, when a cI dimer is bound to O_{R2}, the rate for open complex formation of the *cI* gene is significantly increased. Ref. [89] provides the ratios between the various transcription rates: $k_{\mathrm{RM2}}/k_{\mathrm{RM1}} = 11$ and $k_{\mathrm{R}}/k_{\mathrm{RM1}} = 14$, where k_{RM1} is the basal transcription rate of P_{RM}, k_{RM2} is the enhanced transcription rate of the same promoter, and k_{R} is the transcription rate of P_{R}. Averaging the values measured in [100, 101, 102] for k_{RM1} to $0.001\mathrm{s}^{-1}$, we obtain $k_{\mathrm{RM2}} = 0.011\mathrm{s}^{-1}$ and $k_{\mathrm{R}} = 0.014\mathrm{s}^{-1}$, which will be used in our model. Translation of mRNA molecules is modelled as a reaction that produces proteins from transcripts; mRNAs can undergo repeated translations and are stochastically degraded with a rate $\mu_{\mathrm{mRNA}} = 0.0058\mathrm{s}^{-1}$, corresponding to a typical half life of 2 minutes [103]. The translation rate of the messenger RNA is set to $k_{\mathrm{prod}} = S\mu_{\mathrm{mRNA}}$, where S indicates the average number of proteins produced from one mRNA. S strongly depends on the gene transcript: it is reported that a cI mRNA transcript produces a factor 20-70 less proteins than a *lacZ* mRNA transcript [104], while cro transcripts produce about half the number of proteins as *lacZ* transcripts [105]. This data allows us to estimate that $S_{\mathrm{cI}} = 6$ and $S_{\mathrm{cro}} = 20$. Having determined S and μ_{mRNA}, we can obtain k_{prod}, bearing in mind that this rate is a coarse-grained description of a complex process, depending on numerous factors affecting translation, such as ribosomes binding sites, limiting amino acids and rare codons. Moreover, this reaction also accounts for post-translational protein modifications and folding.

Proteins in general have a long half-life, which can greatly exceed the cell generation time. They are thus mainly removed from the cell by dilution due to cell growth and division. We model the depletion of proteins due to dilution by introducing a degradation reaction for all the proteins in the scheme (monomers and dimers), with rate $\mu_{\mathrm{dilution}} = \ln 2/t_{\mathrm{cell\,cycle}}$, We assume here $t_{\mathrm{cell\,cycle}} = 34$ min as in the experiments of Little *et al.* [84]. In addition to dilution effects, [106] reports that the protein cro has a half-life of 30-60 minutes due to active degradation processes. As in Ref. [53], we choose a mean half-life of 42 minutes, and add an active degradation term for cro monomers $\mu_{\mathrm{cro}} = \ln 2/(42\mathrm{min}) + \mu_{\mathrm{dilution}}$. cI monomers and both cI and cro dimers TFs are removed from by dilution.

Rapidly growing bacteria contain multiple replication forks in their chromosome. Therefore, it seems reasonable to assume that more than one copy of the operator is present in the system. We assume here that the number of operators, averaged over the

cell cycle, is $N_O = 3$ [107]. We neglect the possibility of multiple copies of the lambda genome being integrated in tandem into the host chromosome [108].

The concentration of free RNA polymerase in the cell is set to 30 nM [109] and it is kept constant during the simulations. The average bacterial volume is assumed to be $2\mu\text{m}^3$ [53].

4.3 Methods

The system we are investigating involves 190 reactions, listed in Appendix E, and an analytical solution is not a feasible strategy. Moreover, approximating the dynamics using macroscopic rate equations would have strong effects on the dynamics of spontaneous switching events (as shown in Chapter 3, which are typically driven by fluctuations in systems out of equilibrium. The reaction set will therefore be simulated using the Stochastic Simulation Algorithm (SSA) which propagates the system according to the chemical master equation [35], described in Section 1.6.1.

As the system is subjected to fluctuations in this simulation scheme, we expect spontaneous transitions to happen between the two stable states. However, especially for the lysogenic to lytic switching, these transitions are extremely rare, and therefore they can not be efficiently sampled by our SSA. To measure the switching rate, the Forward Flux Sampling method, described in Section 1.6.4, will be used.

In the field of soft condensed matter physics a number of simulation schemes have been developed in recent years, which make it possible to zoom in on the rare events themselves [110, 111, 112, 113, 114, 115, 116, 117, 118, 119]. However, these schemes require knowledge of the phase space density. For systems that are in equilibrium—systems that obey detailed balance and microscopic reversibility—the phase space density is known: it is given by the Boltzmann distribution. In contrast, for systems that are out of equilibrium, the phase space density is usually not known. Hence, the application of most numerical techniques for simulating rare events is limited to equilibrium systems, and thus is not applicable to the case we are investigating. However, the Forward Flux Sampling (FFS) technique [42, 43, 44], which has recently been developed, makes it possible to compute rate constants in both equilibrium and non-equilibrium systems with stochastic dynamics, Furthermore, FFS allows for sampling of the transition path ensemble. Even with FFS, measuring the switching rate for our model remains a challenging task that can be achieved only with very long serial simulations. The original FFS code was parallelised with MPI and the load was distributed over several processors. A parallel version of the kinetic Monte Carlo code was also written.

The SSA is an event-driven scheme, and quickly becomes inefficient when some of the reactions occur with very high frequency, either due to a large number of reactant molecules or to a fast reaction rate. In our reaction set, dimerisation of transcription factors meets these conditions: most of the computational effort in our simulations is devoted to these "fast" reactions. "Slow" events, like binding of a protein to DNA, are selected

only rarely. Our previous study of a model genetic switch [49], developed in Chapter 2 shows that dimerisation does not affect the switching pathways in phase space significantly, and has a limited impact on the rate of the transitions between the two stable states when $k_f > k_{prod} V_{cell}$, as this is the case for our model. It would therefore be useful to integrate the dimerisation reactions out, using some of the approximations available in the literature [120, 68, 34, 31] and tested in Chapter 3, and simulate a coarse-grained version of the model. We must be careful, however, that the approximation we choose does not effect the dynamical behaviour of the system, *i.e.* that the new coarse-grained scheme correctly reproduces the dynamical properties of the original network. In Chapter 3, we have studied how the equilibrium and dynamic properties of a model genetic switch change when it is coarse-grained according to several alternative approximation methods. We found that fast dimerisation equilibria can be safely integrated out, provided that the effective rate constants of the slow reactions are obtained by solving the Master Equation for dimerisation, thus properly incorporating fluctuations in the monomer-dimer equilibrium. In analogy with Chapter 3, we substitute the original molecules cI, cI_2, cro and cro_2, with new, fictitious, "slow" species cI_{tot}=cI+2cI_2 and cro_{tot}=cro+2cro_2, whose number does not change with the fast dimerisation reactions. We then write a new reaction set, where the fast reactions are eliminated and the slow reactions are rewritten in terms of the slow variables. We solve the dimerisation Master Equation for fixed values of the slow species, to obtain the probability distributions $p(cI|cI_{tot})$ and $p(cro|cro_{tot})$, which can be used to compute averages of the fast variables, *e.g.* $\langle cI \rangle_{cI_{tot}}$ and $\langle cI_2 \rangle_{cI_{tot}}$, as described in Appendix A. Finally, the effective rates for the new slow reactions are obtained in terms of these averages.

4.4 Results

We simulate first the full reaction set using the SSA. In these simulations, we can follow the dynamical evolution of the system from a given initial condition to the closest stable steady state in phase space. We expect the system to be bistable, namely to show one stable state rich in cI and another one rich in cro. We first check the existence of these hypothesised stable states, starting from suitable initial conditions. The system should display fluctuations around the fixed points and spontaneous transitions to the other basin should be extremely rare, especially when the system resides in the lysogenic state. In order to visualise the steady states, we choose a simple, one-dimensional, order parameter λ for the system, defined as the difference between the total number of cI molecules and the total number of cro molecules present at a given time [42]. The same order parameter will be used for our FFS simulations. We have seen in Chapter 2 that this order parameter in fact corresponds quite closely to the committor function for the transition.

We initially prepare the system with a large number of cI molecules and no cro, as if the host were in the lysogenic state. The situation is maintained throughout the whole simulation (Figure 4.2, left panel), corresponding to a simulated time of several hours: the cI concentration fluctuates around 250 nM, while cro is tightly repressed. When the initial

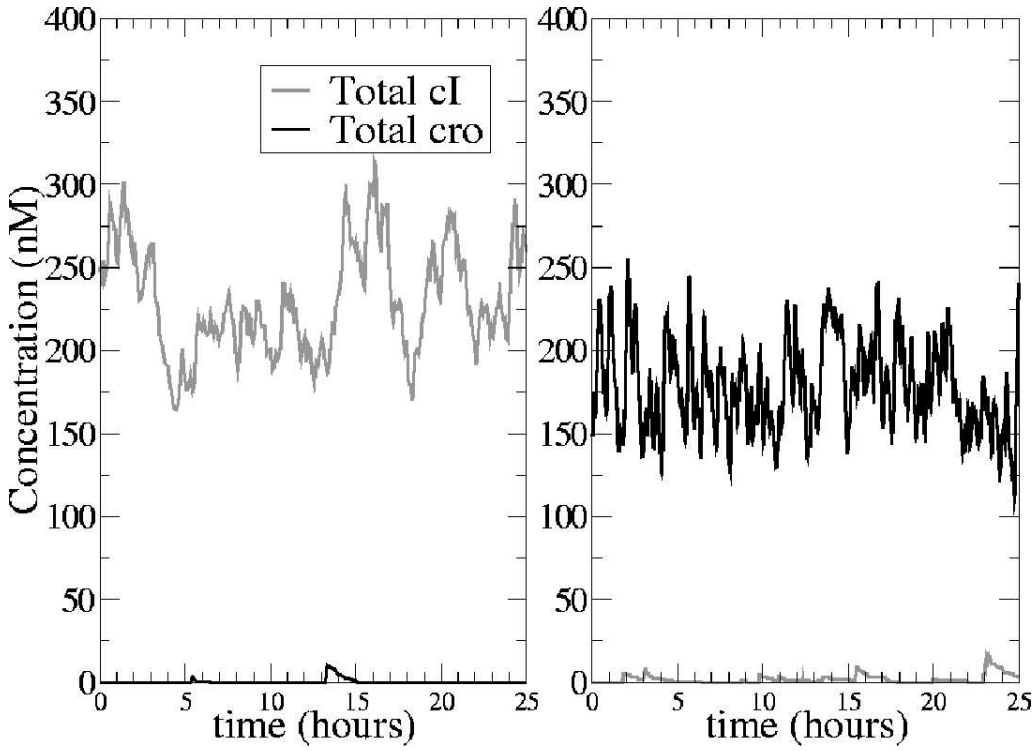


Figure 4.2: Time evolution of the total number of cI and cro molecules in the system, showing the bistability of the system. Left panel: the system is prepared in the lysogenic state with 250nM of cI. The concentration of the majority species settles around the same value, while the concentration of the minority species remains close to zero. Right panel: a similar situation is observed for the lytic state, where cro is found in an average concentration of 180nM. Occasional production events of the minority species are immediately repressed by the reaction network.

condition is set to 150 nM cro and no cI (lytic state), we see the system fluctuating around a stable state rich in cro, with an average concentration of about 180 nM, as shown in the right panel of Figure 4.2. This time, however, the spikes in the concentration of the minority species are more frequent and have a higher intensity, which could indicate a lower stability of this steady state. In both cases, the concentration of the minority species starts to increase from time to time, but it almost immediately decays back to zero: stochastic fluctuations give rise to sporadic production events, but the genetic network exerts tight control and forces the system back to the original steady state. We ran the system in the lysogenic state over a simulated period of 10 days and did not record any spontaneous switching events. This situation is compatible with the experimental observations of [87]. Unlike [53, 86], the average concentrations of cI and cro in the two stable states were not imposed in the model, but arose from the stochastic simulation itself. The compatibility with the data measured in [91, 53, 121] provides validation of the model. Ref. [87],

however, shows a strongly varying cI concentration in the lysogen, in the range of 80-300 nM.

In the simulations of Figure 4.2, 99.97% of the computational time is spent on simulating the dimerisation reactions. We proceed then to the elimination of these reactions, as indicated in Section 4.3. This resulted in a gain in speed of about a factor 100 for this system, and the trajectories of the resulting coarse-grained system are statistically indistinguishable from those shown in Figure 4.2. This shows that not only the steady state behaviour, but also the fluctuations in the total number of molecules are preserved when the dimerisation reactions are integrated out.

Having found the two stable states, we can now measure the spontaneous switching rate with the Forward Flux Sampling method. Data reported in [87] suggest that the lytic state is much less stable than the lysogenic. We therefore expect to measure a lysogenic-to-lytic switching rate that is much lower than the rate of the reverse process. The FFS simulation is performed on the original full reaction set, using the order parameter λ defined above. In the steady states, the minority species is practically absent, so we defined the lysogenic steady state to be the region of phase space where $\lambda \geq 480$, corresponding to $[\mathrm{cI_{tot}}] > 240\mathrm{nM}$ (in the absence of cro), and the lytic state state the region where $\lambda \leq -300$, corresponding to $[\mathrm{cro_{tot}}] > 150\mathrm{nM}$ (in absence of cI).

The simulations of the original reaction set are extremely slow, and only a small number of points (30-70) per FFS interface can be afforded on a standard computer. Under these conditions, the switching rate we measure has a large error bar, about the size of the mean. This means that we cannot determine whether or not our rates are lower than the average. Hence, we are only able to provide a reliable upper bound for the switching rates: $k_{\mathrm{lys}\to\mathrm{lyt}} \leq 10^{-14}s^{-1} \approx 10^{-11}$ per generation per cell, and $k_{\mathrm{lyt}\to\mathrm{lys}} \leq 10^{-5}s^{-1} \approx 10^{-2}$ per generation per cell. No previous model of the bacteriophage λ switch has been able to predict a switching rate compatible with the experimentally measured value, which is less than $2 \cdot 10^{-9}$ per generation per cell [84]. As we reported in Section 4.1, the rate of spontaneous switching is lower than the rate of spontaneous mutation in the genome. Therefore, the experiments of Ref. [87] observe a number of spontaneous flipping events which yield a rate higher than what has been reported. However, sequencing the DNA of the cells that have undergone a spontaneous flip shows that about 99% of these events were triggered by a mutation in the DNA which decreased the stability of the lysogen state, and were not classified as "genuine". Spontaneous switching events due to a rare burst in production of RecA (the protein that induces the switching to the lytic state upon DNA damage by cleaving the cI dimers) are avoided in Ref. [84] by working with *recA*-lysogens.

Previous models assumed equilibrated dimerisation and TF binding to the operator, and overestimated the switching rate by several orders of magnitude. Conversely, our model fully accounts for stochastic fluctuations and predicts a rate which is compatible with the experimental measurements. The lytic to lysogenic rate is hard to measure, because once the virus enters the lytic pathway, it quickly destroys the host. However, in

im- mutants (also known as "anti-immune" cells) the *cro* gene is disconnected from the lytic machinery, and cro protein can accumulate without interfering with growth and division of the host. Experiments performed by Calef *et al.* [108] show that 0.1-1% of these mutant cells switch to the lysogenic state over 5-10 generations, providing evidence of a reverse transition with a much higher rate. To our knowledge, these, together with [122] and [123], are the only experimental measurements of the rate of the lytic to lysogenic transition, and they are in qualitative agreement with our findings.

Encouraged by these preliminary results, we ran FFS simulations on the coarse-grained reaction set, substantially increasing the number of points collected at each interface. This led to more accurate results, within the previous upper bounds: $k_{\mathrm{lys}\to\mathrm{lyt}} = 2.2 \pm 0.2 \cdot 10^{-15} s^{-1} \approx 5 \cdot 10^{-12}$ per generation per cell, and $k_{\mathrm{lyt}\to\mathrm{lys}} = 2.3 \pm 0.310^{-7} s^{-1} \approx 5 \cdot 10^{-4}$ per generation per cell. The agreement between these two FFS results suggests that the fluctuations introduced by the dimerisation of transcription factors are not essential in the switching process, confirming the legitimacy of the coarse-grained approach. Our model of the simple O_{R} operator then produces two stable states, one of which has only a tiny probability of spontaneously switching to the other state.

In Ref. [84], some mutant phages were constructed experimentally, in which the binding site motif on the operator was modified. In particular, the usual sequence of binding sites in the operator, $O_{\mathrm{R}123}$, was mutated to either $O_{\mathrm{R}121}$ or $O_{\mathrm{R}323}$. These variants maintain the same qualitative behaviour (stable lysogens, prophage induction upon DNA damage by UV light) as the wild-type phage. However, the spontaneous switching rate from the lysogenic to the lytic state is about one order of magnitude higher for the mutant $O_{\mathrm{R}121}$ and more than two orders of magnitudes higher for $O_{\mathrm{R}323}$. Notably, in the simulation model [53], the stability of $O_{\mathrm{R}323}$ cannot be reconciled with the stability of the wild type phage.

We have implemented these mutant patterns of binding sites in our chemical model. Again, the simulation of the coarse-grained reaction set without dimerisation reactions gives results which are statistically indistinguishable from those obtained with the original reaction set. We find that the $O_{\mathrm{R}121}$ mutant shows a bistable behaviour qualitatively similar to that of the wild type, with several quantitative differences: in a lysogen, the concentration of cI is only 60nM; the concentration of cro is almost zero, but displays frequent and vigorous bursts, indicating stronger fluctuations, as it can be seen in the left panel of Figure 4.3. Moreover, the stability of the lytic state appears to be reduced: after a few hours of simulation time, the system always switches to the lysogenic state (Figure 4.3, right panel). The switch appears then to be less stable than for the wild-type. This qualitative observation is confirmed by computation of the spontaneous switching rate with FFS: $k_{\mathrm{lys}\to\mathrm{lyt}} = (4.4 \pm 0.4) \cdot 10^{-11} s^{-1} \approx 9 \cdot 10^{-8}$ per generation per cell, several orders of magnitude higher than in the wild type. The switching rate for the reverse transition is $(4.8 \pm 0.1) \cdot 10^{-5} s^{-1} \approx 10^{-2}$ per generation per cell.

In general, our analysis of the $O_{\mathrm{R}121}$ mutant is in qualitative agreement with the experimental results of Ref. [84]. However, we cannot obtain similarly good agreement

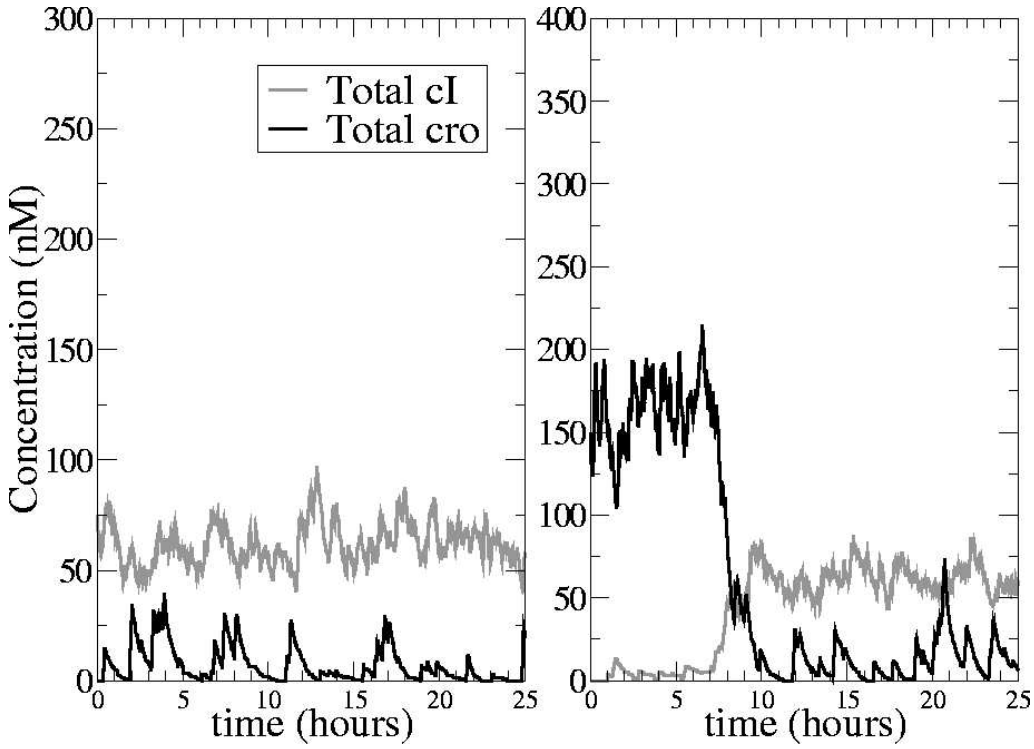


Figure 4.3: Time evolution of the total number of cI and cro molecules in the system, for the mutant phage $O_{\mathrm{R}121}$, whose behaviour is experimentally analysed in Ref. [84]. Compared with the wild type (Figure 4.2), the concentration of cI in the lysogenic state is lower (left panel). This, together with the looser repression of cro, indicates a decreased stability of the lysogenic state. In the right panel the system was prepared in the lytic state, and after few hours a transition to the lysogenic state occurred. We never observed similar event when simulating the wild type phage over a comparable time scale. Therefore, we conclude that the stability of the lytic state is decreased in this mutant. These findings are compatible with those in Ref. [84].

for the $O_{\mathrm{R}323}$ mutant: our model predicts a completely unstable lysogenic state. As it is depicted in Figure 4.4, when the system is prepared with many cI and no cro, it immediately switches to a state rich in cro (about 100 nM) which appears to be very stable. The monostability of the system is caused by the properties of the binding site $O_{\mathrm{R}3}$: it displays both the highest and the lowest affinities found in the O_{R} operator, binding cro very strongly ($\Delta G = -13.4$kcal/mol) and cI very weakly ($\Delta G = -9.5$kcal/mol [81]). The weak promoter P_{RM} is then tightly repressed by a the presence of cro dimers, and a single large fluctuation is enough to exit the lysogenic state.

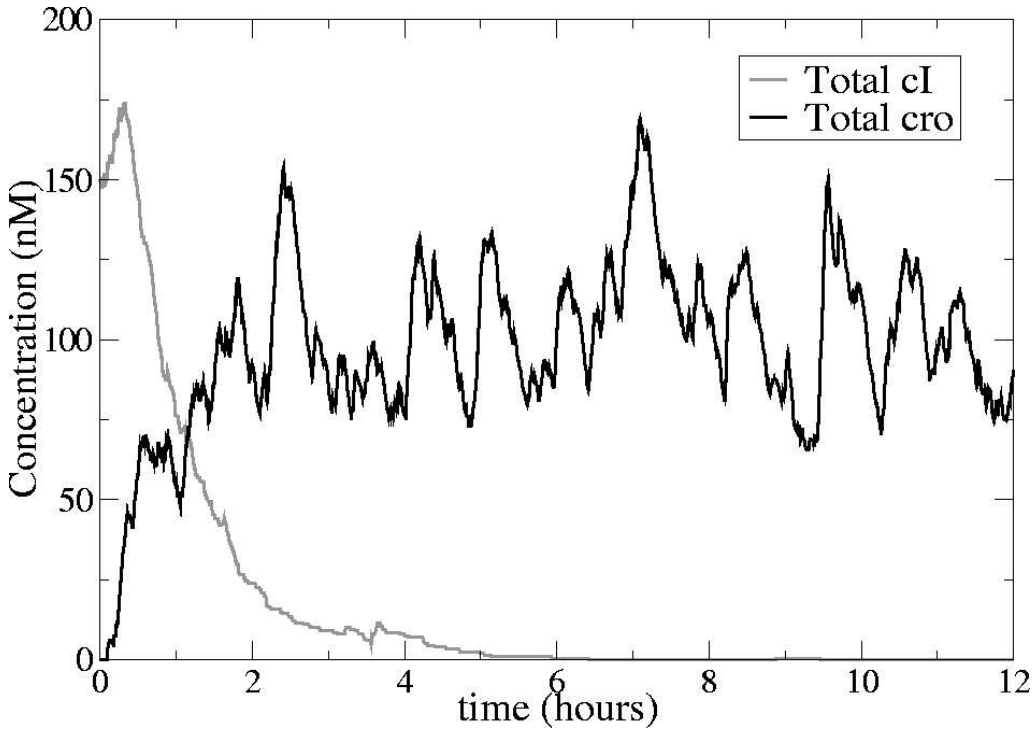


Figure 4.4: Time evolution of the total number of cI and cro molecules in the system, for the mutant phage O_{R323}, whose behaviour is experimentally analysed in Ref. [84]. The lysogenic state is not stable in our model: when the system is prepared with a high concentration of cI, it immediately switches to a state rich in cro, which is sustained for long time. The experiments of Ref. [84] indicate that this mutant shows the same bistable behaviour as the wild type (although the switching rate is even higher than in the O_{R121} mutant), a property that we can not reproduce with our model.

4.5 Macromolecular Crowding

Most of the reaction constants available in the literature have been obtained *in vitro*. The highly dilute conditions in a test tube prevent interactions between the solute molecules and approach "ideal" conditions. Conversely, the *in vivo* environment is notably different– in particular, many large molecules are present which may interact with the reactants [124]. Even though these interactions are mostly nonspecific, *e.g.* due to excluded volume effects, the high macromolecular concentration of a living cell (up to 400 g/l [124]) is likely to make the conditions very different from ideal. The average interaction of a species i with the other molecules in the cell can be described as a non-ideal contribution to the chemical potential $\mu_i^{\mathrm{non-id}} = kT \ln \gamma_i$. This extra term arises from interactions between solutes molecules, and must be added to the usual expression for the chemical potential in ideal conditions: $\mu_i^{\mathrm{id}} = \mu_i^0 + kT \ln c_i$, where c_i is the (normalized) concentration of species i. Denoting γ_i the *activity coefficient* of species i, its *thermodynamic activity*

can be defined as $a_i = \gamma_i c_i$. If the conditions are ideal, $\gamma_i = 1$ for all species; when solute molecules interact repulsively, $\gamma_i > 1$ (it costs more work than in ideal conditions to insert a molecule of species i), whereas if the interactions are attractive, $\gamma_i < 1$. Equilibrium constants are strictly speaking defined in terms of activities [95], which can be approximated to concentrations *only* when conditions are close to ideality. For instance, for the simple reaction $A + B \rightleftharpoons C$, the relation between the equilibrium constant of a reaction and the free energy change of the system becomes:

$$K_{\text{eq}} = \frac{a_C}{a_A a_B} = \frac{\gamma_C}{\gamma_A \gamma_B} \frac{c_C}{c_A c_B} = \gamma \exp\left(-\frac{\Delta G}{RT}\right) = \gamma K_{\text{eq}}^{\text{vitro}}, \tag{4.2}$$

where γ is the ratio between the activity coefficients of the products and the reactants. In this case, $\gamma > 1$ means that the non-ideal environment shifts the reaction equilibrium towards the products (with respect to the ideal case), whereas $\gamma < 1$ indicates a shift towards the reactants.

According to [125], the interactions between macromolecules in the cell are mostly non-specific, that is, they do not depend on structural details, but only on general macromolecular properties, like charge or volume. We assume that the interactions between molecules are due solely to excluded volume. In this case, the system minimises its free energy by favouring reactions which reduce the total excluded volume [126]. This effect results in a depletion interaction between reactants, due to the crowding agents. The reactants tend to collapse into more compact structures. In particular, for association-dissociation equilibrium, the bound state experiences a "caging" effect which keeps the reactants together, lowering the dissociation rate and shifting the equilibrium of the dimerisation reaction [127]. Supposing that *in vitro* experiments only measure $K_{\text{eq}}^{\text{vitro}}$, we model this effect by setting $\gamma > 1$ in Eq. (4.2). We keep the diffusion-limited association rate, and use Eq. (4.2) to compute the dissociation rate.

Furthermore, the crowded environment affects some dynamical properties of molecules in the cell, namely their diffusive motion. Intuitively, diffusing in a crowded environment is harder than in a dilute solution: the motion of several tracers has been measured *in vitro* for different concentrations of crowding agents, showing that the particles continue to diffuse, but with a diffusion coefficient which decreases exponentially with the concentration of crowding agents [128]. The estimated macromolecular density inside a living cell is 200-400g/l: assuming an analogous behavior to what measured in Ref. [128], diffusion coefficients are reduced by a factor 3-30. In order to account for this effect in our model, we modify the diffusion-rate association rate $k_{\text{ass}} = 4\pi DL$ of reacting species accordingly to the assumed macromolecular density. Diffusion constants of proteins have been measured *in vivo* in *E. coli* [93, 129, 130] by means of FRAP (Fluorescence Recovery After Photobleaching) techniques [131] to be between 0.4 and 10 μ^2/s. We note however that most of these works report widely varying data: it is not clear if this is to be attributed to limitations in the experimental techniques or to real variations in the diffusion coefficients across different cells. We decided then to scale the diffusion coefficients in our model as a function of the cellular macromolecular density, according to the results of Ref. [128].

The values we obtain are always comprised within the error bars of the experimental studies [93, 129, 130].

In summary, the crowded environment of the cell influences chemical reactions in two distinct ways: by reducing the diffusion-limited association rate, and by shifting the equilibria of association-dissociation reactions towards the bound products, which have lower excluded volume.

We have investigated the effects on our model of bacteriophage λ when we include macromolecular crowding using the procedure described above. It is not easy, however, to quantify a precise γ coefficient for all the bimolecular reactions considered in the model. In principle, γ can differ between reactions. For instance, while cI and cro molecules are both average sized proteins for *E. coli*, with a weight of a few tens of kDa, RNA polymerase is a much bigger molecule (about 400kDa). However, crowding effects depend mostly on the linear dimension of a molecule, and the difference in radius between these two species approximately scales with the cubic root of the mass, yielding comparable dimensions. For simplicity, we assume that RNAp and transcription factors are equally affected by macromolecular crowding. Moreover, while it is obvious that crowding favours the formation of oligomers, for protein-DNA associations the situation is less intuitive. The difference in excluded volume between the bound complex, and the bare DNA and an unbound molecule is not easy to compute, due to the details of the interactions (the case of RNAp is instructive: the molecule "grabs" the DNA and the excluded volume of the bound complex is only marginally different from that of the reactants). Moreover, these reactions happen within the bacterial nucleoid, where the high concentration of nucleic acids forms a dense polymer mesh, which is thought to exacerbate crowding effects. Furthermore, experimental data reveals that, in highly crowded conditions, the binding equilibrium constant for DNA replication proteins to phage T4 DNA is strongly increased [127]. We make then a drastic assumption, in supposing that the effect of crowding on DNA-protein complexes is the same as for protein-protein complexes. This rough approximation can be justified given the uncertainty already present in some of the parameters in this Section, and bearing in mind that our main aim is to investigate the qualitative effects of crowding on the model system, rather than to try to determine exact quantities. Because of these approximations, the switching rates measured for $\gamma > 1$ must be taken *cum grano salis*.

We proceed then by repeating the approach of Section 4.4 for several values of γ, and try to detect trends in the switching rate; all the simulations are performed on the coarse-grained reaction set, where dimerisation reactions are integrated out. As a first attempt, we apply the SSA to the system at $\gamma = 10$ (corresponding to a macromolecular concentration of about 200 g/l [124]). The simulation reveals decreased stability of the lytic state (Figure 4.5): after about 10 hours, we observe a spontaneous transition to the lysogenic state (we never observed such an event for the wild type system at $\gamma = 1$). We systematically investigate the effect of increasing γ on the switching rates, and the results are collected in Figure 4.6. For each case, we first run an SSA simulation to define the

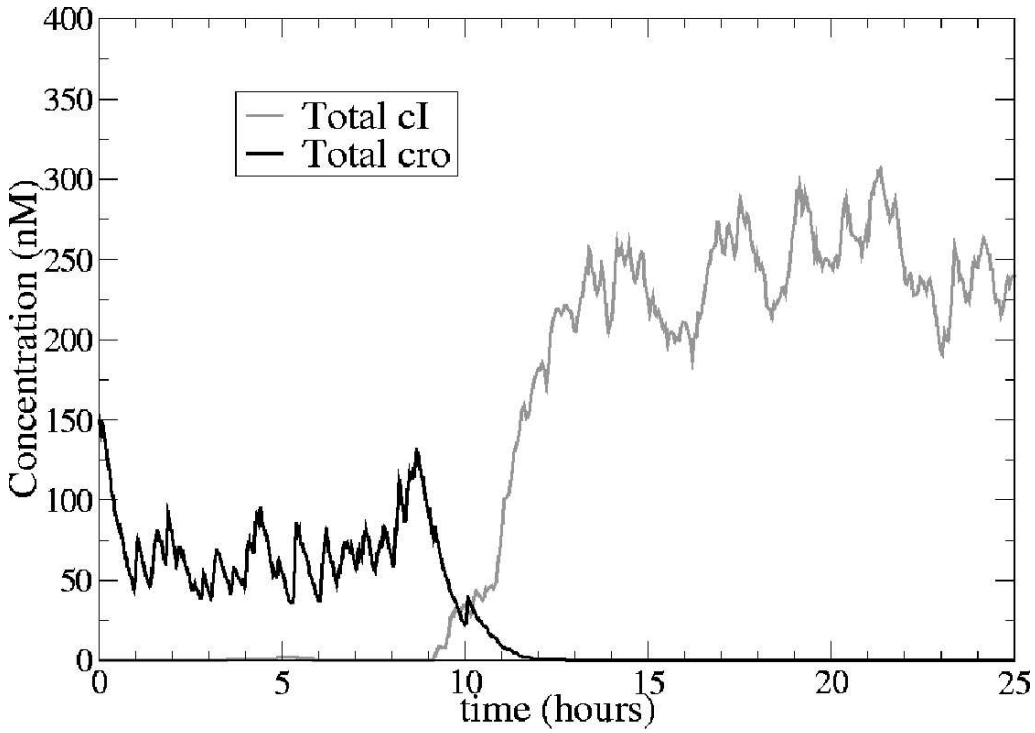


Figure 4.5: Time evolution of the total number of cI and cro molecules in the system for $\gamma = 10$ (corresponding to a macromolecular concentration of about 200 g/l [124]). We observe a spontaneous transition from the lytic to the lysogenic state, suggesting decreased stability of the former.

boundaries of the lytic and lysogenic states; we then design a suitable set of FFS interfaces and compute the spontaneous switching rates. For increasing γ, the lysogenic-to-lytic rate shows a minimum before increasing again around $\gamma = 10$ (Figure 4.6, panel A), while the reverse rate tends to increase with γ (Figure 4.6, panel B). We conclude that a crowded environment has the overall effect of increasing the stability of the lysogenic state and decreasing the stability of the lytic state. This effect is more pronounced for the lytic state, but it never change the rates by more than two of orders of magnitude.

It is interesting to investigate the effect of crowding on the Little mutants discussed in Section 4.4: for the O_{R121} mutant, crowding leads again to an increased stability of the lysogenic state, together with a loss of stability of the lytic state. For $\gamma = 5$, switching rates are comparable with the ideal case ($k_{\mathrm{lys}\to\mathrm{lyt}} = (7 \pm 1) \cdot 10^{-11}\mathrm{s}^{-1}$ and $k_{\mathrm{lyt}\to\mathrm{lys}} = (4.6 \pm 0.1) \cdot 10^{-5}\mathrm{s}^{-1}$), but already at $\gamma = 10$, the system is able to reside in the lytic state for a few hours at most, before switching to the lysogenic state. We are not aware of any measurement of the stability of the lytic state or any experiment conducted in presence of crowding agents for an *im-* O_{R121} mutant, so we are not able to check this prediction of the model. On the other hand, the increased stability of the lysogenic state due to crowding has a more interesting effect on the O_{R323} mutant: on increasing γ, the lysogenic state

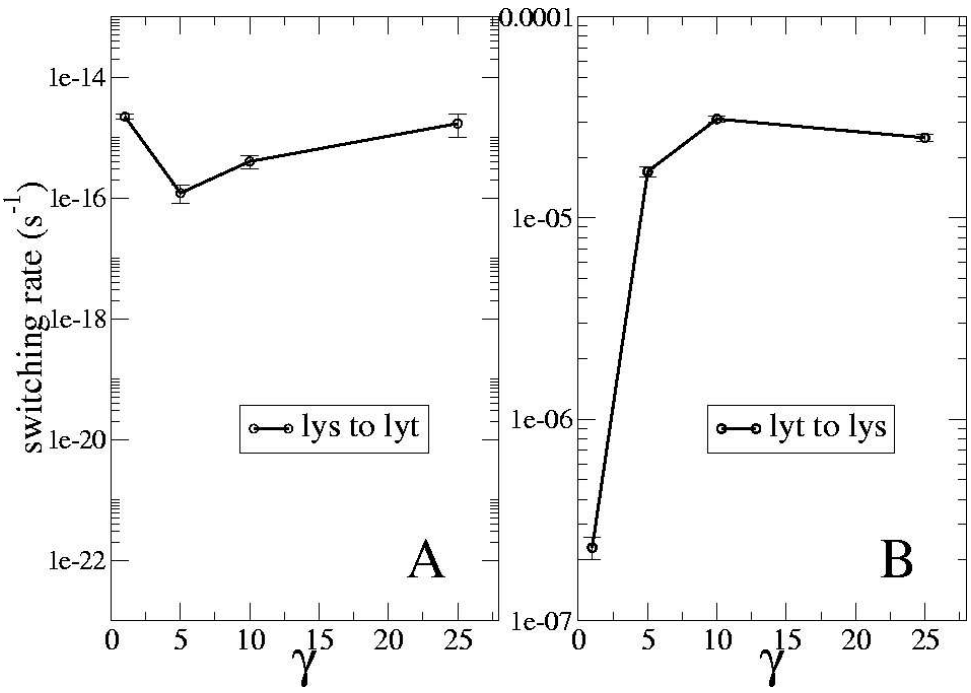


Figure 4.6: Switching rates computed with FFS, for several activity coefficients, reflecting increasing crowding in the cell ($\gamma=1$: ideal conditions, $\gamma=25$: macromolecular concentration ~ 300g/l). The lysogenic-to-lytic switching rate (panel A) shows a minimum for $\lambda=5$, while the reverse rate (panel B) increases steeply for low γ, and then and reaches a plateau. Macromolecular crowding can change the switching rates by 1-2 orders of magnitude.

gradually becomes stable (for $\gamma=10$, $k_{\mathrm{lys}\to\mathrm{lyt}}=(2.2\pm0.1)\cdot10^{-6}\mathrm{s}^{-1}$ and $k_{\mathrm{lyt}\to\mathrm{lys}}=(6.2\pm0.2)\cdot10^{-6}\mathrm{s}^{-1}$), while the lytic state loses its stability around $\gamma=25$. This observation suggests that the data in [84] can be explained within a single model, if we suppose that the effects of the cellular environment are compatible with a global activity coefficient of $\gamma>10$.

We account for crowding here in a very crude way, yet, nevertheless, we are able to observe a trend common to the wild type and two mutant phages, namely that the crowded *in vivo* conditions tend to stabilise the lysogenic state and could be partly responsible for its exceptional stability.

4.6 DNA looping

Recently, the role of another operator on the phage λ genome has been found to be very important for the stability of the switch [90, 132]. This operator, known as O_{L}, is located 2400 bp away on the λ genome, and appears to be very similar to O_{R}, with an analogous

pattern of binding sites for cI and cro dimers. Only one promoter, P_{L}, is found in the O_{L} region. This overlaps with the binding site $O_{\mathrm{L}1}$, and is thus prone to blocking by a bound cI dimer. P_{L} lies ahead of the *n* gene, which codes for a late lytic protein. The long distance between the two operators has led to the long-lived assumption that it plays a role as an unimportant auxiliary operator in the stability of the switch. However, it has recently been reported that mutants defective in O_{L} have considerably more cI in the lysogenic state [90, 91]. As a consequence, these mutants cannot reliably switch to lysis when the bacterial DNA is damaged. A recent paper [92] points out the possibility of a DNA loop being formed between the two operators, stabilised by the formation of a cI octamer, as depicted in Figure 4.7. This octamer would involve pairs of cI dimers already bound to the sites $O_{\mathrm{R}1}, O_{\mathrm{R}2}$ and $O_{\mathrm{L}1}, O_{\mathrm{L}2}$, interacting through their C-termini (usually called "heads"), which could "lock" the DNA, increasing the stability of the lysogenic state. Once the loop has formed, the weak association of a cI dimer to $O_{\mathrm{R}1}$ and the consequent autorepression of the *cI* gene would be helped by the close presence of $O_{\mathrm{L}3}$. This last site has a much higher affinity for cI than $O_{\mathrm{R}3}$ and thus a higher probability of binding a cI dimer. A cooperative interaction can arise between the heads of an $O_{\mathrm{L}3}$-bound dimer and a free dimer, to make a tetramer and increase the local concentration of cI, thus facilitating the binding of the second dimer to $O_{\mathrm{R}3}$. It has to be noted that the formation of cI octamers has been observed *in vitro*, but only at cI concentrations much higher than those found in a cell [133], suggesting that the DNA could play an important role by locally increasing the cI concentration.

In order to include the effects of looping in our stochastic model, we extended the reaction set to include binding of cI and cro dimers to O_{L}. The free energies for these reactions, as reported in [92], show that O_{L} binds cI tighter than O_{R}. Furthermore, we considered all the possible ways of forming the loop starting from different operator states (*i.e.* we added all the reactions coupling 2 operators whose binding sites 1 and 2 are occupied by cI dimers). Ref. [92] estimates free energies for the cooperative interactions that lead to the formation of cI tetramers and octamers. However, we prefer to model the loop formation in a slightly different manner: we consider the loop to be formed by an equilibrium fluctuation of the DNA, and stabilised by the octameric bond, whose energy has been measured *in vitro*.

We are then left with estimating the typical association and dissociation times of the loop. We begin by realising that the DNA reaches its looped state due to a conformational change and that acts as a scaffold for the cI complex formation. We therefore treat the loop formation as a first order process. On the length scale of the loop, the DNA is completely floppy (2400bp correspond to about 15 persistence lengths) and the elastic bending energy is thus completely negligible. Furthermore, we assume that the two ends of the loop come into contact due to an equilibrium fluctuation of the DNA polymer. Subsequently, the interaction between the heads of the cI tetramers bound to O_{R} and O_{L} can quickly lock the loop. In analogy with our treatment of TF-DNA association, we neglect the time it takes for the bond to be formed: in this scheme, the mean association time of the loop

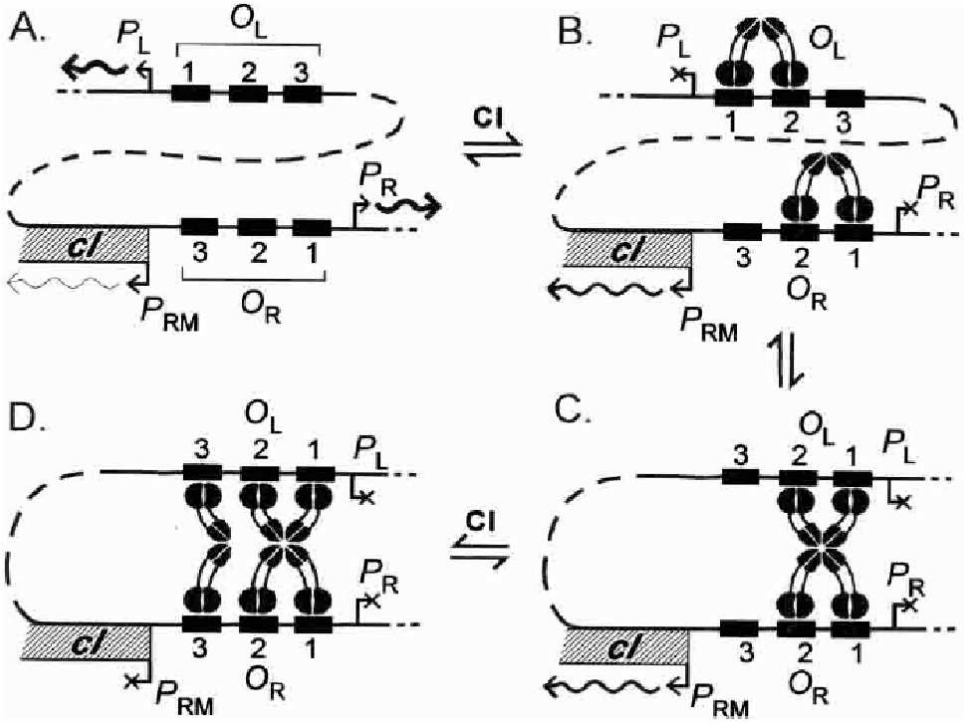


Figure 4.7: Figure taken from [92], describing the DNA looping interactions between O_R and O_L. The cartoon depicts the cI-DNA complexes at O_L and O_R on the λ genome and their effect on gene transcription as the cI concentration increases. cI dimers first bind cooperatively to O_{R1}, O_{R2}, and O_{L1}, O_{L2}, switching off the *cro* and *n* genes, and forming two tetramers connected by the heads of the proteins (B). A fluctuation of the DNA configuration can bring the heads of the tetramers together, where they can bind and lock the DNA in the looped state (C). At high cI concentrations, a dimer bound to O_{L3} can guide a second dimer to O_{R3}, thus repressing the *cI* gene (D).

is equal to the mean relaxation time of a polymer whose ends have been in brought as close as the loop width, which can be computed analytically [134]. For a loop length $L_0 = 2400\text{bp} \approx 800\text{nm}$, persistence length $l_p = 150\text{bp} = 50\text{nm}$, DNA thickness $d = 2.5\text{nm}$, and cellular friction coefficient $\eta = 0.005$ Pa·s, the typical association time is of the order of tens of seconds. Finally, we model the dissociation process as an escape from a potential well whose depth is set by the binding free energy of the octamer: as the DNA is not "pulling" on the loop, the dissociation time depends only on the strength of the interaction between the two tetramers bound to the two operators: $t_{\text{diss}} = t^0_{\text{diss}} \exp(-\Delta G_{\text{oct}}/K_B T)$, where t^0_{diss} can be estimated as the mean escape time for two diffusing particles initially in contact: $t^0_{\text{diss}} = L^2/D$. For $\Delta G_{\text{oct}} = -9.1\text{kcal/mol}$, $L = 10\text{nm}$, $D = 5\mu\text{m}^2/\text{s}$, $t^0_{\text{diss}} = 3.3\mu\text{s}$, and $t_{\text{diss}} \approx 10\text{s}$. These last values come from an estimate, and a number of factors could significantly change the parameter values. Nevertheless, we believe that these rates are

not too far from the right order of magnitude, and that they provide us with a starting point for our simulations.

When we include looping in our model, the number of reactions in our scheme increases to 401 and the simulations are consequently much slower. Brute-force simulations show that the stability of both the lysogenic and lytic states is maintained, as depicted in Figure 4.8. The formation of the loop enhances the negative autoregulation of cI, and lowers its concentration in the lysogenic state to about 80nM. Clearly, the more cI or cro molecules a cell accumulates in a steady state, the more difficult it will be to get rid of them (keeping all the other parameters constant) and spontaneously switch to the other steady state. However, maintaining a high number of proteins in a cell costs energy: although producing few more hundreds of them is not a heavy burden for an *E. Coli*, Ref. [135] reports that, in the case of the *lac* system, bacterial cells can evolve towards optimal protein levels, differing only by few percents to initial levels. It is then likely that even the levels of phage λ protein are the result of an optimization process, and the looping could then potentially be exploited to increase the stability of the lysogenic steady states without increasing the concentrations of the regulating proteins. We expect the lysogenic-to-lytic switching rate to decrease as an consequence of the stabilization induced by the DNA loop. However, the increased complexity of the reaction set significantly slows down the simulations, and, even with coarse-graining of dimerisation reactions, we could only measure an upper bound for the lysogenic-to-lytic switching rate: $k_{\mathrm{lys}\to\mathrm{lyt}} < 10^{-21}\mathrm{s}^{-1} \approx 10^{-18}$ per generation per cell. The reverse switching process is facilitated because the cI concentration in stable state is lower, and it takes a smaller effort to be reached. We measure a reverse switching rate of $(1.8 \pm 0.1) \cdot 10^{-5}\mathrm{s}^{-1}$, a couple of orders of magnitude lower than the result obtained without looping.

Finally, we include macromolecular crowding effects in the model with looping by varying the coefficient γ. These simulations are again extremely computationally demanding, and only upper bounds can be computed for the most important switching rate. The results are collected in Figure 4.9. Unfortunately, our data allow only limited quantitative conclusions about the influence of looping on the switching rate. In comparison with Figure 4.6, the stability of the lysogenic state is certainly not decreased, and there are indications that it could on the contrary be increased, as we expect. Especially for $\gamma = 1$, the lysogenic state shows a much higher stability. The reverse rate is 2 orders of magnitude higher for $\gamma = 1$, and slightly increased for higher concentrations of crowding agents (we should not forget that, for high γ, the stability of the lytic state is always marginal).

4.7 Discussion

In this Chapter, we have designed a chemical model of the core genetic network governing the bacteriophage λ genetic switch, and we have solved it numerically at the level of the chemical master equation. Including stochasticity in the model is crucial, as we have

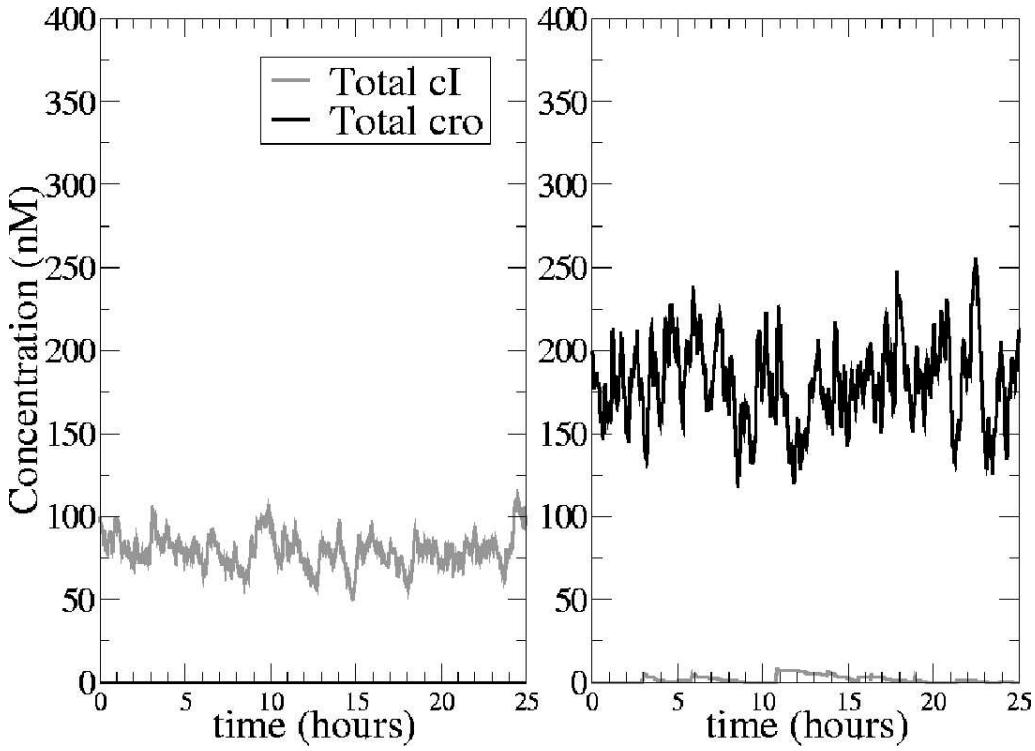


Figure 4.8: Time evolution of the total number of cI and cro molecules in the system accounting for DNA looping. Compared to Figure 4.2, the system is still bistable, with occasional production events of the minority species immediately repressed by the reaction network. The concentration of cI in the lysogenic steady state is lower, as the loop brings together binding sites O_{R1} and O_{L1}, favouring the negative autoregulation of the *cI* gene.

seen in Chapter 2 for a model switch (loosely inspired by this biological system). Bacteriophage λ has been studied extensively over the last 60 years, and contributed greatly to the birth of a new branch of science, today known as molecular biology. The amount of data collected during this large stretch of time probably makes bacteriophage λ one of the best-characterized systems in biology. This wealth of data make it possible to construct a quantitative, detailed model of the system without being forced to guess a large number of parameters. The reaction set which we have written only describes the dynamics of one operator (two in the case of looping) within the λ genome. Despite O_{R} being universally recognized as the core of the biochemical network underlying the switch, many other proteins come into play in other phases of the phage life cycle, and interact with the basic switch network. We did not consider the initial infection of the host by the phage, nor the process of prophage induction (the induced switching from the lysogenic to the lytic state), nor the insertion of the phage genome into the *E. coli* chromosome, nor any other behaviour of the phage, many of which have received great attention in the literature. Our

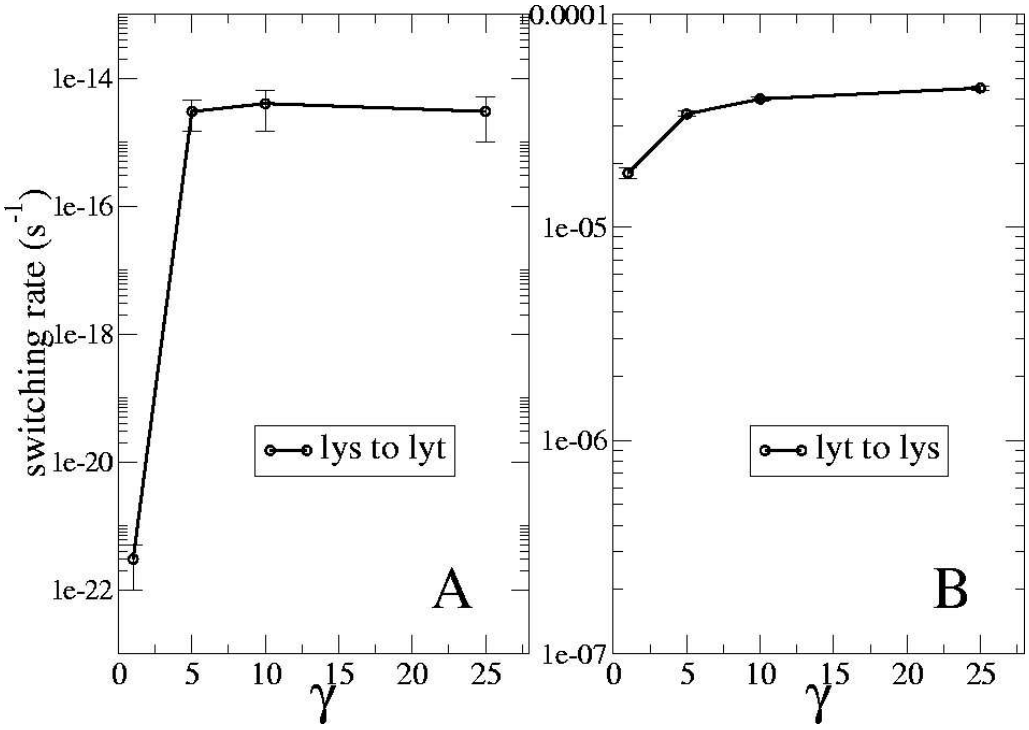


Figure 4.9: Switching rates computed with FFS, for several activity coefficients, for the model with DNA looping. The simulations are computationally demanding, and the results for the rate of spontaneous switching allow only for limited comparison with Figure 4.6. We observe the larger differences for $\gamma = 1$, where the DNA loop strongly increases the stability of the lysogenic state, and increases the stability of the lytic state by about two orders of magnitude.

choice of the reaction network was then based on the "classical" interpretation of the genetic switch. Recent studies challenge this interpretation and suggest that the role of cro is less relevant in the switching process. In [121] it is argued that cro is needed to repress P_{RM} only when the phage has already committed to the lysogenic pathway. In [90, 136] this analysis is completed focussing on the role of cro during prophage induction. Moreover, a new wave of papers, aiming at a more quantitative understanding of the system is becoming available, together with a number of recent reviews, confirming a renewed interest in this system [137, 138, 139].

We have emphasized throughout this Chapter that the differences between *in vivo* and *in vitro* conditions are notable, and, typically, underappreciated. We acknowledge that it is not an easy task to account for the cellular environment: it is virtually impossible to model all the essential features that differentiate a cell from a test tube. We focussed our attention here on the fact that such an environment is crowded, namely that it is tightly packed with macromolecules and other cellular structures. These crowding agents mainly exert a nonspecific repulsive interaction (due to excluded volume effects) on the proteins, "push-

ing" them together and favouring associated protein-protein and protein-DNA states. We build the consequences of crowding into our model by assuming a slower diffusion of the proteins, and by shifting reaction equilibria towards the aggregated state. This last result is obtained by modifying a simple parameter γ, defined as a combination of the activity coefficients of the participating species in a reaction. For γ values compatible with concentration of 250-300g/l we find that all the experimental data on the spontaneous switching rate can be reconciled to our model. We consider this a success, and a confirmation of the importance of the role that is played by the cellular environment.

Clearly, this oversimplified approach to modelling the *in vivo* conditions does not address other features that could play an important role. In particular, the issue of proteins binding nonspecifically to DNA has recently been raised in the context of bacteriophage λ: according to the model in [140, 141], about the 90% of cI and 50% of the cro in the cell is nonspecifically bound to DNA at physiological concentrations. This would imply that the number of available molecules in the cytoplasm would be much lower than that which we have considered. However, these studies assume that the whole genome is covered with potential binding sites, and do not account for other proteins which could possibly be nonspecifically bound. Moreover, there is evidence that histone-like proteins that bind nonspecifically to DNA could further reduce the regions of bacterial DNA where proteins can bind [142, 143].

The shape and dynamics of bacterial DNA is another much debated subject: the biopolymer is confined in the bacterial cell, whose typical linear dimensions are much smaller than the radius of gyration of the DNA. Histone-like proteins can help the bacterium to pack the genetic material, compacting it into the nucleoid. Another viewpoint is, however, present in the literature: in [144], an experiment in which crowding agents are added to an expanded *E. coli* genome suggests that depletion forces alone are strong enough to cause the formation of the compact nucleoid. The formation of the nucleoid can then be interpreted as a phase separation of superhelical DNA in a suspension of proteins, as computed theoretically in [145]. General agreement in this area has not yet been achieved, and therefore all models of the properties of DNA must rely on some hypothesis. Finally, supercoiling of DNA could also influence its physical properties. DNA loops span generally only a small fraction of the genome length, yet they can nevertheless be influenced by the issues we briefly discussed above. DNA looping is nowadays thought to occur quite frequently: the DNA acts as a scaffold and, by enhancing the local concentrations of specific proteins, can drive the association of complexes which would form only at much higher concentrations *in vitro* [132, 146, 147]. We have modelled the DNA in our loop as a short stretch of DNA as a worm-like chain, without considering effects due to the presence of histone-like proteins, specific kinks or DNA-bending proteins. Crowding was crudely accounted for through an increased cytoplasmic viscosity. The results we obtained seem to indicate that the O_R—O_L loop could lead to increased stability of the lysogenic state. Although this is only a preliminary result, it points in the right direction and could act as a basis for a more detailed model, including detailed modelling of DNA dynamics in the regulatory properties of a genetic network.

Chapter 5

A Brownian Dynamics Algorithm for Reaction-Diffusion systems

Fiet uti nusquam possit consistere finis
effugiumque fugae prolatet copia semper
Titus Lucretius Carus

Brownian Dynamics algorithms are widely used for simulating soft-matter and biochemical systems. Recently, they have been applied to the simulation of coarse-grained models of cellular networks in simple organisms like bacteria. In these systems, components move by diffusion, and can react with one another upon contact. However, when reactions are incorporated into a Brownian Dynamics algorithm, attention must be paid to avoid violation of the detailed-balance rule, which could include systematic errors in the simulation. In this Chapter, we present a Brownian Dynamics algorithm for reaction-diffusion systems which rigorously obeys the detailed-balance rule. After subjecting our algorithm to stringent tests on elementary systems, we apply it to the simulation of a "push-pull" network in which two antagonistic enzymes covalently modify a substrate. Our results highlight that the diffusive behaviour of the reacting species can strongly reduce the gain of the response curve of this network.

5.1 Introduction

Computer simulations are essential tools for understanding the behavior of biological systems and biochemical reactions. Among the numerous techniques used in these fields, Molecular Dynamics and Brownian Dynamics (BD) have played a role of fundamental importance. While the former is more rigorous and detailed, it is not, however, able to reach long time scales, because of the smallness of the time steps required for a correct simulation. Conversely, the latter accounts for the faster solvent dynamics at a mean-field level: the solvent is treated implicitly, and the solute molecules experience a random displacement, as a net result of the huge amount of fast collisions with the solvent molecules. BD is thus able to use longer time steps and achieve longer simulation times. The first BD algorithm was introduced by Ermak and McCammon [148] and has been widely used to study the behavior of protein-protein association in atomic-detailed simulations [149, 150, 151, 152, 153, 154]. In particular, in the case of diffusion-limited chemical reactions, the method of Northrup and Erickson [155] provided a way to describe the encounter complex first formed by the two reactants, before the subsequent, short-distance rearrangements bring the molecules in the final associated state. Effects of crowding [156] and competition of other charged molecules with substrate molecules binding an enzyme [157] were also considered.

More recently, the evidence that biomolecules move in prokaryotic cells primarily by diffusion has suggested that biological systems can be modelled as reaction-diffusion systems, and that Brownian Dynamics could be a useful tool in this field. In these simulations, molecules are often coarse-grained to the level of simple geometrical objects (typically spheres), that can react with other chemical species in a confined geometry. The reaction partners can either propagate diffusively or be immobilised in special locations (as for instance, interfaces representing external and intracellular membranes). Within this simplified version of the cell, simulation times can be brought up to physiological time scales. Many stochastic simulators for biochemical reactions have recently been developed [158, 159, 160, 161, 162, 39, 163]. Some of them account for individual particles in space and make use of BD or BD-like algorithms for their propagation: Ref. [160] uses a BD version specialised in simulating reactions between free-diffusing ligands and stationary surface receptors, while Ref. [159] numerically solves the Smoluchowski model of diffusion-limited reactions.

These stochastic simulators have been applied also to biochemical networks. These networks are composed by proteins and DNA that chemically and physically interact with each other, and allow a cell to detect and respond to changes in its environment. Stochastic computation methods are usually required to study biochemical networks, as they are generally prone to fluctuations. The origin of fluctuations can be *temporal*, *i.e.* rooted in the intrinsic stochasticity of chemical reactions and in the stochastic distribution of the reaction times. However, when reactions are diffusion-limited, the position of the reactants can introduce a further source of fluctuations, having a *spatial* origin. When the diffusion constants of particles are low, spatial fluctuations can dominate the behaviour of the network. While temporal fluctuations can be accounted for by describing the system

with its Master Equation (see Section 1.5.3) and solving it with the Stochastic Simulation Algorithm described in Section 1.6.1, spatial fluctuations are ignored in this approach, which assumes that the system is well-stirred. A Brownian Dynamics algorithm is able to bridge the gap and correctly account for both sources of fluctuations, thus unravelling effects that have a purely spatial origin.

In a reaction-diffusion Brownian Dynamics scheme, associations between molecules, in a coarse-grained description, must be introduced carefully: the detailed-balance rule must strictly be obeyed in order to avoid introducing systematic errors in the simulation. A critical analysis of second-order reaction mechanisms in BD must thus be addressed. Usually, in simulation packages based on BD, a reaction is supposed to happen when a move brings two reactants to an overlap; in [160] the reaction rate is used to compute a probability of reaction, but the determination of this probability is not reported. In [159] reaction rates are translated into binding and unbinding radii, setting the distance of association and dissociation events to the center of mass of the two particles, which is rather unphysical.

In this Chapter, we present a Brownian Dynamics algorithm which rigorously obeys detailed balance and is thus able to reproduce equilibrium properties of a reaction-diffusion system. In Section 5.2, we derive our algorithm on the basis of the statistical mechanics of chemical reactions. The algorithm will be subjected to stringent tests in Section 5.3: besides equilibrium properties, we test also how well the algorithm reproduces the dynamical behavior of the system, for different values of the time step. A comparison with a stochastic algorithm that does not account for spatial fluctuations of particles is also presented. Finally, in Section 5.4 we show an illustrative application of our algorithm to a simple coarse-grained model of a chemical species subjected to the action of two enzymes, operating in opposite directions (the so-called "push-pull" model system) [164]. The BD algorithm allows us to assess the effect of both spatial and temporal fluctuations in reducing the gain of the response of the system.

5.2 Methods

5.2.1 System

Despite its wide use, few Brownian Dynamics algorithms take proper care of the detailed-balance rule when treating second-order reactions. When two reaction partners come into close physical proximity, they can react with a probability related to the reaction rate k. It is common practice to evaluate this probability only when a diffusive move has led the two particles to (a partial) spatial overlap. However, in the case of a reversible reaction, when the two reactants dissociate, the products are usually positioned at contact, or in close proximity [159]. The dissociation move must be chosen such that the whole algorithm does obey detailed balance, otherwise the equilibrium properties of the system will not be correctly reproduced. An intuitive explanation is as follows: suppose that a particle jumps from a position x to a new position overlapping with a reaction partner, and a reac-

tive event is accepted. According to the standard procedure, the reverse move will never bring the reactive particle back to its initial position x, thus violating the detailed-balance rule. Alternatively, one can consider positioning the dissociated particle *exactly* at the initial position it has reacted from. This is not completely correct either, as such a procedure introduces an angular bias which violates the hypothesis of angular isotropy of the two dissociating species. In this Chapter, we introduce a new method of treating second-order reactions in a Brownian Dynamics simulation, which rigorously obeys detailed balance and reproduces satisfactorily the dynamical properties of the system, when the time steps are small enough.

We begin by considering the elementary reaction:

$$A + B \rightleftharpoons C \qquad (k_\mathrm{f}, k_\mathrm{b}), \tag{5.1}$$

where k_f is the forward rate for the association of molecules A and B, and k_b the backward rate for dissociation. This is the basic building block of our simulation scheme, and we will therefore study it in detail. First, we evaluate the partition function for the system, while later we extend our derivation to account for the particles' positions in space.

Let N_A, N_B, N_C be the number of A, B and C molecules and V the volume of the system. The partition function of the system can be written as the following sum of terms in the canonical ensemble:

$$\mathcal{Z} = \sum_{\{N\}} Z(N_A, N_B, N_C), \tag{5.2}$$

where $\{N\}$ denotes all possible combinations of $\{N_A, N_B, N_C\}$. The choice of the canonical ensemble is motivated by the assumption that the cell is a closed system that does not exchange particles with the environment.

Let us consider the case where $\{A, B, C\}$ are ideal particles in a volume V, except for the fact, of course, that A and B can form C.
The partition function Z for $\{N_A, N_B, N_C\}$ particles is then:

$$\begin{aligned} Z(N_A, N_B, N_C) &= \frac{q_A^{N_A} q_B^{N_B} q_C^{N_C}}{N_A! N_B! N_C!} \\ &= \frac{q_{A,\mathrm{cm}}^{N_A} q_{B,\mathrm{cm}}^{N_B} q_{C,\mathrm{cm}}^{N_C} (V/\Lambda^3)^{N_A+N_B+N_C}}{N_A! N_B! N_C!}, \end{aligned} \tag{5.3}$$

where $q_{A,\mathrm{cm}}$ accounts for the internal degrees of freedom of the particle, relative to its center of mass (so that $q_A = q_A^\mathrm{id} q_{A,\mathrm{cm}}$), $q_A^\mathrm{id} = V/\Lambda^3$ is the partition function for a molecule in an ideal gas, $\Lambda = h/(2\pi m k_\mathrm{B} T)^{1/2}$ is the thermal wavelength, and the factor $1/(N_A!)$ comes from indistinguishability of particles. The probability that the system has $\{N_A, N_B, N_C\}$ molecules, $P(N_A, N_B, N_C)$, can be written as:

$$P(N_A, N_B, N_C) = Z(N_A, N_B, N_C)/\mathcal{Z}. \tag{5.4}$$

Let us now consider the transition from $\{N_A, N_B, N_C\}$ to $\{N_A - 1, N_B - 1, N_C + 1\}$ molecules. The ratio between the probability of being in the state after and before the transition, is easily computed:

$$\begin{aligned} \frac{P(N_A-1,N_B-1,N_C+1)}{P(N_A,N_B,N_C)} &= \frac{N_A N_B}{N_C+1}\frac{\Lambda^3}{V}\frac{q_{C,\mathrm{cm}}}{q_{A,\mathrm{cm}}q_{B,\mathrm{cm}}} \\ &= \frac{N_A N_B}{N_C+1}\frac{K_{\mathrm{eq}}}{V} \\ &= \frac{N_A N_B}{N_C+1}\frac{1}{V}\frac{k_{\mathrm{f}}}{k_{\mathrm{b}}}. \end{aligned} \tag{5.5}$$

Using the detailed balance rule [165], we can now determine the transition probability to be used in a Monte Carlo scheme, in which the system is considered to be well-stirred, and space is not present:

$$P_{\mathrm{unbound}} P_{\mathrm{u}\to\mathrm{b}} = P_{\mathrm{bound}} P_{\mathrm{b}\to\mathrm{u}}, \tag{5.6}$$

where P_{unbound} is the probability of being in the state $\{N_A, N_B, N_C\}$, $P_{\mathrm{u}\to\mathrm{b}}$ is the probability of a transition from $\{N_A, N_B, N_C\}$ to $\{N_A - 1, N_B - 1, N_C + 1\}$, $P_{\mathrm{b}\to\mathrm{u}}$ is the probability of the reverse move, and P_{bound} is the probability of being in the state $\{N_A - 1, N_B - 1, N_C + 1\}$. Eq (5.5) and the detailed balance rule lead to:

$$P_{\mathrm{u}\to\mathrm{b}} = \frac{k_{\mathrm{f}}}{V} N_A N_B \quad \text{and} \quad P_{\mathrm{b}\to\mathrm{u}} = k_{\mathrm{b}}(N_C + 1). \tag{5.7}$$

In the previous derivation, we did not specify the positions of the particles in the volume V. However, Brownian Dynamics algorithms propagate particles in space and time: reactions happen between molecules at specific coordinates in space. We need therefore to compute a quantity analogous to (5.5), where the locations of particles are made explicit.

Let us now consider the probability $P(\mathbf{r}_A^{N_A}, \mathbf{r}_B^{N_B}, \mathbf{r}_C^{N_C}; \{N_A, N_B, N_C\})$ that the system has (N_A, N_B, N_C) molecules *and* that these molecules are located at positions $\{\mathbf{r}_A^1, \cdots, \mathbf{r}_A^{N_A}\}$, $\{\mathbf{r}_B^1, \cdots, \mathbf{r}_B^{N_B}\}$, $\{\mathbf{r}_C^1, \cdots, \mathbf{r}_C^{N_C}\}$. Applying the Bayes' rule, this probability is given by

$$P(\mathbf{r}_A^{N_A}, \mathbf{r}_B^{N_B}, \mathbf{r}_C^{N_C}; \{N_A, N_B, N_C\}) = \tag{5.8}$$

$$P_N(N_A, N_B, N_C)\mathcal{P}(\mathbf{r}_A^{N_A}, \mathbf{r}_B^{N_B}, \mathbf{r}_C^{N_C} | \{N_A, N_B, N_C\}), \tag{5.9}$$

where $\mathcal{P}$ is the *conditional* probability that the given number $\{N_A, N_B, N_C\}$ of molecules occupy those particular positions. The conditional probability is nothing but the probability of finding $\{N_A, N_B, N_C\}$ indistinguishable ideal particles in a volume V:

$$\mathcal{P}(\mathbf{r}_A^{N_A}, \mathbf{r}_B^{N_B}, \mathbf{r}_C^{N_C} | \{N_A, N_B, N_C\}) = \frac{N_A! N_B! N_C!}{V^{N_A+N_B+N_C}}. \tag{5.10}$$

Hence, we have that

$$P(\mathbf{r}_A^{N_A},\mathbf{r}_B^{N_B},\mathbf{r}_C^{N_C};\{N_A,N_B,N_C\}) = \frac{q_{A,\mathrm{cm}}^{N_A} q_{B,\mathrm{cm}}^{N_B} q_{C,\mathrm{cm}}^{N_C}}{\mathcal{Z}} \frac{1}{\Lambda^{3(N_A+N_B+N_C)}} \tag{5.11}$$

and

$$\frac{P(\mathbf{r}_A^{N_A-1},\mathbf{r}_B^{N_B-1},\mathbf{r}_C^{N_C+1};\{N_A-1,N_B-1,N_C+1\})}{P(\mathbf{r}_A^{N_A},\mathbf{r}_B^{N_B},\mathbf{r}_C^{N_C};\{N_A,N_B,N_C\})} = \frac{q_{C,\mathrm{cm}}}{q_{A,\mathrm{cm}} q_{B,\mathrm{cm}}}\Lambda^3. \tag{5.12}$$

We focus on the following setup: a particle A and a particle B, freely diffusing in a box of volume V with diffusion coefficients D_A, D_B respectively. Particles have finite radii, R_A, R_B, respectively and can not interpenetrate. The analysis above still holds, with the condition of replacing everywhere the volume of the box V with the accessible volume for the A and the B particle, $V^* = V - \frac{4}{3}\pi(R_A+R_B)^3$.
We consider an elementary reaction: setting $N_A = 1, N_B = 1, N_C = 0$, equation (5.11) becomes:

$$\frac{P(\mathbf{r}_C;0,0,1)}{P(\mathbf{r}_A,\mathbf{r}_B;1,1,0)} = \frac{q_{C,\mathrm{cm}}}{q_{A,\mathrm{cm}} q_{B,\mathrm{cm}}}\Lambda^3 = K_{\mathrm{eq}} = \frac{k_{\mathrm{f}}}{k_{\mathrm{b}}}. \tag{5.13}$$

In Eq. (5.13) we have used the chemical definition of equilibrium constant and its macroscopic relation to the forward and backward reaction rates. Relations (5.7) or (5.13) must be obeyed by a correct, rigorous simulation scheme.

5.2.2 Simulation scheme

We describe now a Brownian Dynamics scheme, in continuous space, for the setup described at the end of the previous Section.

We can assume without loss of generality, that $D_A = 0$, *i.e.* that the A particle does not diffuse in the simulation box. It is then convenient to position it at the center of the box. The single B particle moves by free diffusion with coefficient $D_B \equiv D$. At every simulation step, the system is propagated by a fixed time Δt.

In the absence of the A particle, the motion of the B particle is simply described by the Einstein equation:

$$\frac{\partial}{\partial t}\, p(\mathbf{r}',t+\Delta t|\mathbf{r},t) = D\nabla^2 p(\mathbf{r}',t+\Delta t|\mathbf{r},t), \tag{5.14}$$

where $p(\mathbf{r}',t+\Delta t|\mathbf{r},t)$ is the probability of finding the particle at position $\mathbf{r}'$ at time $t+\Delta t$, given that it was at $\mathbf{r}$ at time t.
We know with certainty the position of the particle at the initial time. We also know that at time $t+\Delta t$ the probability of finding the particle in space vanishes as we move far away from the initial position $\mathbf{r}$. We can then formulate the following boundary conditions for Eq. (5.14):

$$p(\mathbf{r}',t+\Delta t|\mathbf{r},t) = \delta(\mathbf{r}'-\mathbf{r}), \tag{5.15}$$

$$p(|\mathbf{r}'|\to\infty,t+\Delta t|\mathbf{r},t) = 0. \tag{5.16}$$

The solution of (5.14) with conditions (5.15) and (5.16) is Gaussian, whose variance is proportional to Δt:

$$p(\mathbf{r}',t+\Delta t|\mathbf{r},t) = \frac{1}{(2\cdot 2D\Delta t)^{3/2}} \exp\left\{\frac{(\mathbf{r}'-\mathbf{r})^2}{2\cdot 2D\Delta t}\right\}. \tag{5.17}$$

This time-dependent probability distribution can be used to generate new positions for the B particle at every time step Δt [83].

When we consider the presence of the particle A in the box, a reaction can occur with the B particle. In our scheme, in order for a reaction to happen, the two particles have to come into contact, *i.e.* they must overlap. However, this is not a sufficient condition: many effects, as for instance an unfavorable contact angle between the two molecules or a high reaction barrier, could impede the progress of the reactive event. From a coarse-grained point of view, we consider then a probability $P_{\mathrm{acc,f}}$ of accepting a reaction, *given* that the two particles overlap. Clearly, this quantity must be related to the intrinsic reaction rate at contact k_f. In analogy with a Monte Carlo scheme, we can similarly define the probability of *generating* a move that leads to a possible reaction, that is to an overlap: $P_{\mathrm{gen,f}}(\mathbf{r})$. This quantity can be computed analytically: let us consider the single particle A held fixed in a center of a large box, whose edges lie far enough to be neglected in the following derivation. Using a polar reference frame whose origin coincides with the center of the A sphere, we can compute the probability that a B particle initially at position $\mathbf{r}$ is displaced to a position $\mathbf{r}' \in \Sigma$, where Σ is the excluded volume for B (a sphere, centered in the origin, with radius $R = R_A + R_B$):

$$p(\mathbf{r}\to\Sigma) = \int_0^R r'^2 dr' \int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\varphi\, p(\mathbf{r}',t+\Delta t|\mathbf{r},t) \equiv g(r,\Delta t). \tag{5.18}$$

The function g can be computed analytically, is radially symmetric and depends on the Brownian Dynamics time step Δt. Details are given in Appendix D. We will not indicate anymore the dependence of various quantities on Δt, since this parameter is kept constant during the whole simulation. We set then $P_{\mathrm{gen,f}}(\mathbf{r}) = g(r)\Omega(\theta,\varphi)$, where $\Omega(\theta,\varphi)$ is the uniform angular distribution on the sphere.

The detailed balance rule between the bound and unbound states imposes

$$P_{\mathrm{unbound}}(\mathbf{r})P_{\mathrm{gen,f}}(\mathbf{r})P_{\mathrm{acc,f}} = P_{\mathrm{bound}}P_{\mathrm{gen,b}}P_{\mathrm{acc,b}}, \tag{5.19}$$

where P_{unbound} is the probability of being in the unbound (that is $N_C = 0$) state, $P_{\mathrm{gen,b}}$ is the probability of generating a move leading to a backward reaction (association) and $P_{\mathrm{acc,b}}$ the probability of accepting such a move. Combining this last relation with (5.13), we obtain a relation between the generation and acceptance of a move and its reverse counterpart:

$$\frac{P_{\mathrm{bound}}}{P_{\mathrm{unbound}}(\mathbf{r})} = \frac{k_\mathrm{f}}{k_\mathrm{b}} = \frac{P_{\mathrm{gen,f}}(\mathbf{r})\,P_{\mathrm{acc,f}}}{P_{\mathrm{gen,b}}\,P_{\mathrm{acc,b}}}. \tag{5.20}$$

Using Eqs. (5.18) and (5.20) we can now fix all the remaining probabilities. Dissociation events are first-order reactions. Assuming that events may happen with a constant probability per unit time yields a Poissonian distribution of waiting times between reactions: $P(t) = k_\mathrm{b} \exp\{-k_\mathrm{b} t\}$. The probability that the reaction has not happened at time t is then $S(t) = -\int_0^t P(t')dt' = \exp\{-k_\mathrm{b} t\}$. If we choose time steps Δt such that $\Delta t \ll 1/k_\mathrm{b}$, the probability that an event happens within Δt is just $k_\mathrm{b}\Delta t$. This can be used to determine the acceptance probability of a dissociation reaction: $P_{\mathrm{acc,b}} = k_\mathrm{b}\Delta t$.

Once we have determined that a dissociation event has happened, we must set a new position for the B particle in the reaction box. To generate the reverse move, we choose the same distribution $g(r)$ we have obtained before, with the proper normalisation. In fact, the probability of starting at position $\mathbf{r}$ in space and landing on the A particle integrates to a volume much smaller than the volume V of the system, for small values of Δt:

$$\int_R^\infty dr \int d\Omega\, \Omega(\theta,\varphi) g(r,\Delta t) r^2 = 4\pi I(\Delta t) \ll V, \tag{5.21}$$

where $\Omega(\theta,\varphi)$ is the uniform angular distribution on the sphere and $I = \int_R^\infty g(r) r^2 dr$. When evaluating the position of the dissociating particle, we already know that a dissociation has happened; the distribution $g(r)$ must now be normalised to one: $P_{\mathrm{gen,b}} = \frac{1}{4\pi I} g(r)\Omega(\theta,\varphi)$. If a trial dissociation move leads to an overlap, then this move should be rejected.

Using detailed balance (5.20), we can now obtain the desired acceptance probability for the forward move:

$$\begin{aligned} P_{\mathrm{acc,f}} &= \frac{P_{\mathrm{bound}}}{P_{\mathrm{unbound}}} \frac{P_{\mathrm{gen,b}}}{P_{\mathrm{gen,f}}} P_{\mathrm{acc,b}} \\ &= \frac{k_\mathrm{f}}{k_\mathrm{b}} \frac{g(r)\Omega(\theta,\varphi)d\mathbf{r}}{g(r)\Omega(\theta,\varphi)d\mathbf{r}\, 4\pi I} k_\mathrm{b}\Delta t \\ &= \frac{k_\mathrm{f}\Delta t}{4\pi I}. \end{aligned} \tag{5.22}$$

We have determined all the quantities we need, and we can proceed to show our BD algorithm.

5.2.3 Algorithm outline

Let us consider a system with M particles of type B and one particle of type A, held fixed at the center of box of volume V. For convenience, we choose as initial state the situation in which there is no bound state C.

1. Generate an initial position for the B particles in the available volume.

2. Select randomly one of the particles among species B and C.

3. (a) If the particle is type B, generate a new position according to a gaussian distribution with zero mean and standard deviation $\sqrt{2D\Delta t}$: $x_{\text{new}} = x_{\text{old}} + N(0, \sqrt{2D\Delta t})$, where Δt the Brownian Dynamics time step.
   (b) If the displacement move leads to an overlap of the B particle with A, that is if $|\mathbf{r}_A - \mathbf{r}_B| < R_A + R_B$, attempt a reaction according to a probability $P_{\text{acc,f}} = k_{\text{f}}\Delta t/(4\pi I)$.
   (c) If the trial reaction move is accepted, remove the B particle from the box, and substitute the A particle with a C. This new particle is not diffusing in the box, as A.
   (d) If the trial reaction move is rejected, put the B particle back to its original position.
4. (a) If the particle is type C, try a backward reaction with probability $P_{\text{acc,b}} = k_{\text{b}}\Delta t$.
   (b) If the trial reaction move is accepted, substitute the C particle with an A particle, create a new B particle whose radial position is drawn from the normalised distribution $g(r)$ and the angular position from the uniform distribution $\Omega(\theta, \varphi)$. If this leads to an overlap with another B particle, then reject the move.
   (c) If the trial reaction move is rejected, keep the identities and positions of particles.
5. Repeat step 2. and 3. or 4. M times, then increase the simulation time by Δt.

Keeping particle A and C fixed could mimic for example a system where one reactant is anchored to some rigid scaffold. A relevant biological example is the binding of proteins to DNA in a bacterial cell, particle A representing a binding site on the DNA, typically in proximity of some gene. In this case, the motion of A is only related to the fluctuations of the polymer, which happen on time scales much longer than the diffusion of proteins in the bacterial cytoplasm, and can therefore be neglected. The scheme could be extended to the situation in which the A particle also moves, or cases with more reactants.

5.3 Tests

In this Section we check the BD scheme against a series of testing procedures. As the algorithm that we have described obeys detailed balance, equilibrium quantities, such as the average time spent in the bound state, must be correctly reproduced. A BD algorithm cannot resolve the dynamics of a system at time scales below the time step Δt used in the simulations. However, dynamical quantities on long time scales should be reproduced, provided that the time step is not too large. In the case of two particles, we investigate then whether our BD algorithm correctly reproduces the survival probability [166] of a B particle, and its probability distribution in time and space. Finally, we compare the distribution of the association times with that obtained with a non-spatial stochastic simulation.

We believe that these are stringent tests and are sufficient to validate the correctness of the BD scheme introduced in the previous Section.

5.3.1 Irreversible Reactions

We begin by simulating the *irreversible* reaction $A+B \xrightarrow{k_f} C$, within the following setup: a single particle A is held fixed in an unbounded system, and a single particle B is positioned on a spherical surface at an initial distance r_0 from A, with a random angle. The particles have the same radius $R_A = R_B = R/2$. We run the algorithm for a time $t_{\rm sim}$ and we record the final radial position of the particle B. In the case that a reactive event happens before $t_{\rm sim}$, we stop the run. We repeat the run for a large number of times, and we collect the final positions of the B particle in a histogram, divided by the fraction of B particles which have survived until the final time. Such a histogram should reproduce the irreversible probability distribution $p_{\rm irr}(r,t_{\rm sim}|r_0,0)$. This quantity has been calculated analytically for 2 ideal particles [167], and represents the probability of finding the two particles at time $t_{\rm sim}$ separated by a distance r, given an initial separation of r_0 at $t_0=0$. We note that this probability distribution is not normalised: the integral over space of $p_{\rm irr}$ is nothing but the survival probability of the particle, that is the probability that the particle has not reacted at the final time. Formally:

$$\int_R^\infty p_{\rm irr}(r,t|r_0)\, r^2\, dr = S_{\rm irr}(t|r_0). \tag{5.23}$$

We are thus able to simultaneously test our algorithm twice: comparing the analytical curve with the profile of our histogram, and the area of the histogram with the analytical value of the survival probability.

Results are collected in Figure 5.1: we simulate the irreversible reaction for 4 different simulation times, from $t_{\rm sim} = 10^{-4}\tau$ to $t_{\rm sim} = 10^{-1}\tau$, where $\tau = R^2/D$ is the natural time scale of the system. Particles are initially positioned at contact: $r_0 = R$. We see that, with a time step 10^{-4} times smaller than $t_{\rm sim}$, the analytical curves perfectly overlap with the numerical data. This means that both the shape and the area of the irreversible probability is captured by our algorithm. For the largest $t_{\rm sim}$, we used a time step of $10^{-6}\tau$.

The fate of the particle is most probably decided within the first time steps of the simulation. For $t_{\rm sim} = 10^{-4}\tau$, we checked then whether the algorithm could yield a correct distribution for different values of $t_{\rm step}$, as shown in the Inset of Figure 5.1. As expected, the agreement for very small time steps is optimal, and it tends to slightly worsen when the time step grows. The Inset shows that the survival probability of the system is mildly underestimated when the time steps are a significant fraction of $t_{\rm sim}$. The dynamics of the system, in the case of an irreversible reaction is then correctly reproduced by the BD algorithm, provided that the time step is not too large.

5.3.2 Reversible Reactions

We extend now the dynamical test performed above to the case of the *reversible* reaction $A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C$. Even for this system, an analytical result is derived in [167], providing another radial probability distribution: $p_{\text{rev}}(r, t_{\text{sim}}|r_0, 0)$. In this test, we proceed similarly as we did for the irreversible case, except that we do not stop the run after a reaction, but we let the particle dissociate. At $t = t_{\text{sim}}$ we check whether the B particle is in the bound state or in the unbound. In the last case we record the final position. The histogram of final positions of the B particles will be normalised to the number of survivors at $t = t_{\text{sim}}$, which also yield an estimate for $S_{\text{rev}}(t|r_0)$. Again, we decide to initialise the B particle at contact $r_0 = R$, so that a large number reactions and dissociations can happen within t_{sim}. With this choice, the test will provide a sound check for the dynamics of the system as described by the BD algorithm. In Figure 5.2, we plot $p_{\text{rev}}(r, t_{\text{sim}}|r_0, 0)$ for 4 different values of t_{sim}, and again we find the BD algorithm to correctly reproduce both the shape and the area of the analytical probability distribution. Similarly to Figure 5.1, we show in the Inset p_{rev}, computed for $t_{\text{sim}} = 10^{-4}\tau$ and different t_{step}. Even in this case, simulations with large time steps slightly underestimate the survival probability.

The next tests will deal with a system with more than one B particles, all of them able to bind to the single A particle. The B particles do not interact among themselves and the system is placed in a box of volume V, endowed with reflecting walls. Particles B and C do not interact among themselves, although they are not allowed to overlap. To our knowledge, there are no analytical results for dynamical properties such as p_{rev} for a many-body problem.

We can check whether *equilibrium* properties of the system, such as the probability of being in the bound state C (p_{bound}), are correctly reproduced by the BD algorithm. Since we have designed the simulation scheme following detailed-balance prescriptions, we should find perfect agreement with the theory, independent of the simulation time step.

The probability p_{bound} can be evaluated by measuring the time when the C particle is present in the system, with respect to the total simulation time. The mean field value for this quantity can be obtained from the macroscopical rate equation in steady state:

$$p_{\text{bound}} = \frac{K_{\text{eq}} N_B}{K_{\text{eq}} N_B + V^*} , \tag{5.24}$$

where $K_{\text{eq}} = k_f/k_b$, and $V^* = V - \frac{4}{3}\pi(R_A + R_B)^3$.

We simulate the system with a varying number N_B of B particles, with a fixed time step $t_{\text{step}} = 10^{-4}\tau$. We choose $K_{\text{eq}} = V$, so that $p_{\text{bound}}(N_B = 1) = 0.5$. The box is cubic and measures $(20R \times 20R \times 20R)$. Figure 5.3 compares the results of our simulations with (5.24): we see a clear agreement between our simulations and the theoretical curve. We performed a check against the "conventional" way of treating dissociations in Brownian Dynamics algorithms: we positioned particles at contact after a reaction, that is, we considered a function $g(r) = \delta(r - R)$. This move does violate detailed balance and affects

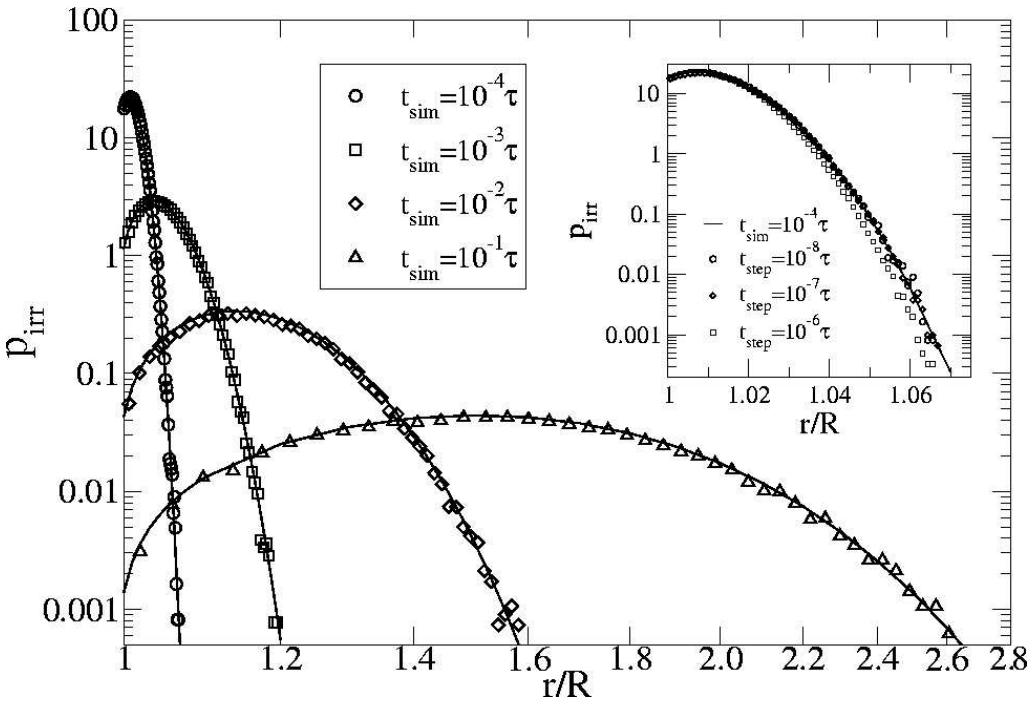


Figure 5.1: Radial probability distribution for an *irreversible* reaction. The four continuous lines represent the analytical solution of the problem and refer to different $t_{\rm sim}$. Symbols correspond to simualations and were obtained with time steps $t_{\rm step} = 10^{-4} t_{\rm sim}$, except for $t_{\rm sim} = 0.1\tau$ where we used $t_{\rm step} = 10^{-6}\tau$ ($\tau = R^2/D$, $R = R_A + R_B$). Particles were initially positioned at contact: $r_0 = R$. The intrinsic association constant is $k_{\rm f} = 1000R^3/\tau$. The numerical results are in perfect agreement with the analytical curves. In the Inset, we plot the probability distribution for $t_{\rm sim} = 10^{-4}\tau$ for several time steps. For large values of $t_{\rm step}$, the BD algorithm deviates from the analytical line and mildly underestimates the survival probability.

$p_{\rm bound}$, as shown in Inset A of Figure 5.3. The incorrect procedure overestimates the time the particle spends in the bound state, especially for a low number of Bs. Finally, we tested whether the equilibrium properties of the system do not depend on the chosen time step. To this end, we compute $p_{\rm bound}$ for $N_B = 1$ and different values of $t_{\rm step}$. As illustrated in the Inset B of Figure 5.3, we obtain a good agreement even for very large time steps, where probably the *dynamics* of the system is not entirely natural. In these runs, we varied $k_{\rm f}$ in order to have $P_{\rm acc,f} = 0.1$, and $k_{\rm b}$ to have $K_{\rm eq} = V$, so that $p_{\rm bound} = 0.5$. Runs with shorter time steps are computationally more intensive.

Finally, we compare our Brownian Dynamics algorithm with a Stochastic Simulation Algorithm (SSA), based on a Kinetic Monte Carlo scheme that propagates the system according to the solution of its chemical master equation [35], as described in Section 1.6.1.

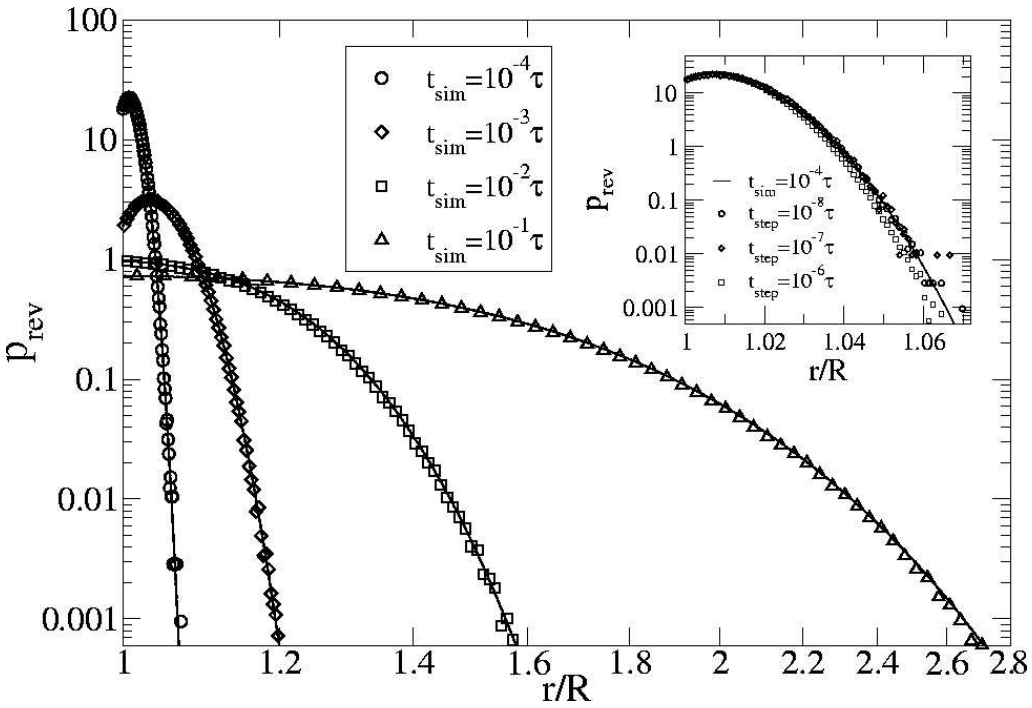


Figure 5.2: Radial probability distribution for a *reversible* reaction. The four continuous lines represent the analytical solution of the problem and refer to different $t_{\rm sim}$. Symbols correspond to simulations and were obtained with time steps $t_{\rm step} = 10^{-4} t_{\rm sim}$, $r_0 = R$. The association constant is $k_{\rm f} = 1000R^3/\tau$ ($\tau = R^2/D$. $R = R_A + R_B$), while the dissociation constant is set to $k_{\rm b} = 100\tau^{-1}$. We observe agreement between the numerical results and the analytical curves. In the Inset, the probability distribution for $t_{\rm sim} = 10^{-4}\tau$ is plotted for several values of $t_{\rm step}$. the BD algorithm deviates from the analytical line and mildly underestimates the survival probability.

This scheme accounts only for the stochasticity arising from the fluctuations in the number of particles; spatial fluctuations due to the diffusive motion of particles are completely neglected. The system is thus assumed to be well-stirred at all times. We consider the reversible reaction $A + B \underset{k_{\rm d}}{\overset{k_{\rm a}}{\rightleftharpoons}} C$ for $N_B = 1$: in the SSA, the association times follow a Poisson distribution, with mean $1/k_{\rm a}$, where $k_{\rm a}$ is the SSA forward rate. In the BD scheme, the association of particles is governed by the rate $k_{\rm f}$, *given* that the particles are already in contact (overlapping). In order to correctly compare the results with the SSA, we must account also for the time a particle needs to reach its reaction partner. We set therefore $1/k_{\rm a} = 1/k_{\rm f} + 1/k_{\rm D}$ [166], where $k_{\rm D} = 4\pi RD[B]$ is the Smoluchowski diffusion-limited association constant. In other words, the mean association time in SSA is the sum of the mean reaction time given the particles are overlapping and the mean time it takes for a particle to diffuse to the target.

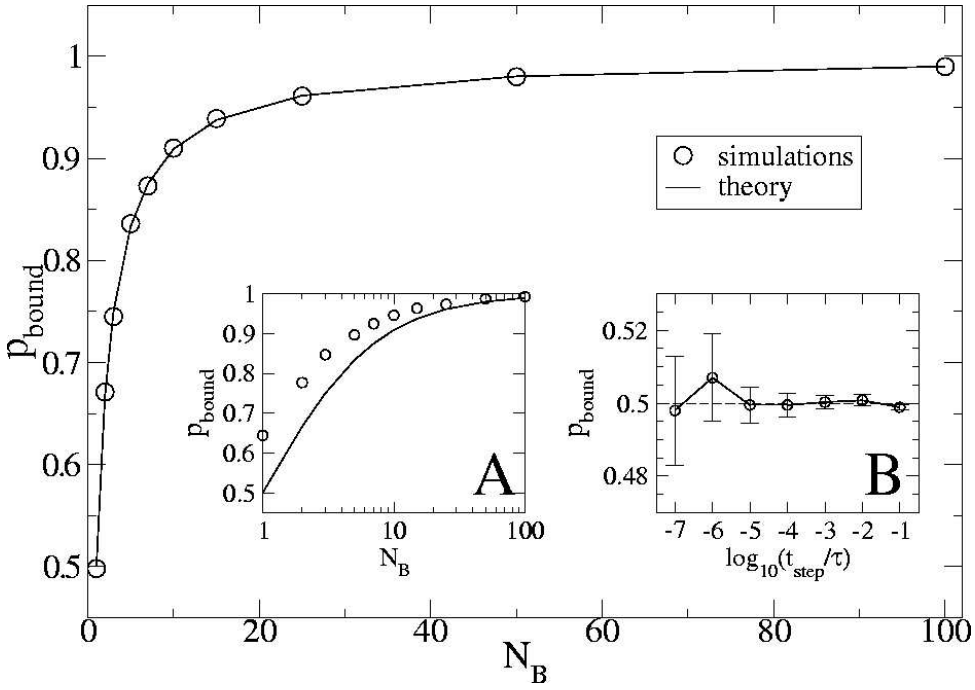


Figure 5.3: Probability of having an A particle bound to a B particle, as a function of the number of B particles. The time step is set to $t_{\text{step}} = 10^{-4}\tau$ ($\tau = R^2/D$, $R = R_A + R_B$), the intrinsic association constant is $k_{\text{f}} = 71R^3/\tau$ so that $P_{\text{acc,f}} = 0.1$. k_{b} is chosen so that $K_{\text{eq}} = k_{\text{f}}/k_{\text{b}} = V$ ($V = 8000R^3$) and therefore $p_{\text{bound}}(N_B = 1) = 0.5$. The numerical data obtained with BD are in agreement with the mean-field values. The error bars of the numerical results are smaller than the size of the circles. In Inset A, the simulations are performed positioning dissociated particles at contact. This move violates detailed balance and yield an incorrect p_{bound} for low number of particles. In Inset B, p_{bound}, for $N_B = 1$ is plotted against the time step used in the simulations. To keep $P_{\text{acc,f}} = 0.1$, we varied k_{f} from $2242R^3/\tau$ ($t_{\text{step}} = 10^{-7}\tau$) to $0.00026R^3\tau$ ($t_{\text{step}} = 10^{-1}\tau$). As expected for an equilibrium quantity, p_{bound} does not depend on the chosen time step.

We collect the association times for a BD run with $V = 64000R^3, D = R^2/\tau, t_{\text{step}} = 10^{-4}\tau, k_{\text{f}} = 100R^3/\tau, k_{\text{d}} = 1000\tau^{-1}$, and we compare it with an SSA run obtained in the same conditions, apart for the modified association rate. Figure 5.4 compares the two distributions: the BD line shows a marked increase in the region of short association times over the expected Poissonian distribution with mean k_{a}, expected for the SSA. This effect has a purely spatial origin and has been previously observed ([41, 168], and Chapter 6): when particles dissociate in space, their distance is still very small, therefore the probability of an immediate rebinding in next few times steps is very high. Long association times, in a BD simulation, are related to particles which have wandered diffusively in the box, and have finally found the target. The distribution of such times is again exponential, with a constant k_{a}. This test indicates that the algorithm correctly propagates the system

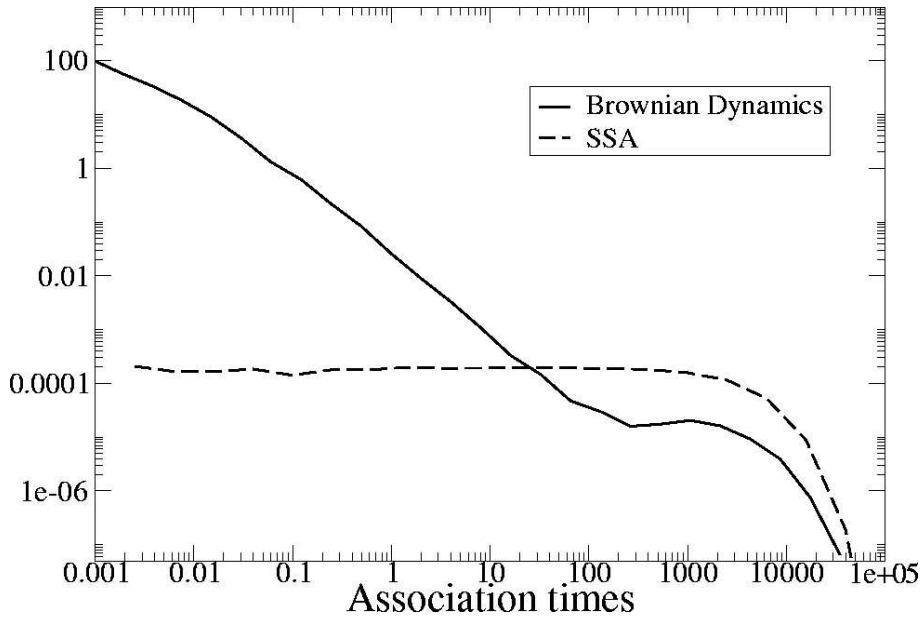


Figure 5.4: Distribution of association times, for the reaction $A + B \leftrightarrow C$, obtained with the Brownian Dynamics algorithm (solid line) and with a Stochastic Simulation Algorithm (dashed line) neglecting spatial effects. The data are obtained for $V = 64000R^3, D = R^2/\tau, t_{\text{step}} = 10^{-4}\tau, k_{\text{f}} = 100R^3/\tau, k_{\text{d}} = 1000\tau^{-1}$ ($\tau = R^2/D$, $R = R_A + R_B$). Spatial simulations account for immediate rebindings after a dissociation event, and show a higher probability for short association times. The two curves decay exponentially to zero with the same rate $k_{\text{a}}^{-1} = k_{\text{f}}^{-1} + k_{\text{D}}^{-1}$.

in time (for $t_{\text{sim}} \gg t_{\text{step}}$) and properly accounts for detailed balance.

5.4 Application: the push-pull model

In this Section, we show an illustrative application of the Brownian Dynamics scheme. We consider a simple model system aimed, with notable oversimplifications, to represent the dynamics of a substrate molecule S under the action of two antagonistic enzymes. The model was first introduced in 1981 by Goldbeter *et al.* [164]. The two enzymes covalently modify the substrates, as in the widely-found biological case of attachment/detachment of phosphate groups to proteins. The first enzyme converts a substrate molecule into an "active" state: bearing in mind the phosphorylation example, we call this active substrate S_p and the enzyme K (kinase). A molecule in the active state S_p can be brought back to the original state by reacting with the second enzyme, P (phosphatase). The model is nicknamed "push-pull", as the substrates are continuously switching between the two states, while consuming energy. The reactions with the enzymes are described according to Michaelis-Menten kinetics: the two reactants form first an intermediate bound state,

which can lead either to a dissociation or to the release of a converted molecule. In [164], the model is solved at the level of the Macroscopical Rate Equation at steady state, which yields the average behavior of the system.

The push-pull model was originally introduced to show that such a system can display an ultrasensitive behavior (that is, a sensitivity curve steeper than the conventional response showed by the Michaelis-Menten mechanism) without the need of introducing cooperative interactions. More precisely, the interplay between two converter enzymes operating in opposite directions on a target whose quantity is conserved can give rise to a switch-like response in the steady-state fraction of modified molecules, when the conversion rates k_1 and k_2 are varied. The requirement for such a sharp transition is the saturation of the enzymes: the effective conversion rates then become independent on the number of substrate molecules, thus attaining a quasi zero-order regime.

The above-mentioned analysis does not however account for any kind of fluctuations that may arise from the low number of reactants, the stochastic behavior of the chemical reactions, or the diffusion of the molecules in space. In [169], the same model is studied at the level of the chemical master equation, taking into account finite-size effects in real systems, that is the discreteness and the low copy number of enzymes and substrate. In order to achieve ultrasensitivity, the enzymes must be saturated, and therefore their concentration is likely to be very low. Large fluctuations are then observed around their average behavior: the authors show that the results obtained with a mesoscopic approach reduce to those of the macroscopic analysis of [164] *only* when the number of molecules is sufficiently large. If this is not the case, as it can easily happen in a bacterial cell where some species are present only in few dozens of copies, the increased sensitivity of the system is reduced, and the response is less steep than the macroscopic theory would predict. This deviation can be easily understood when one realises that high sensitivity corresponds to highly saturated enzymes. In this regime, the reaction rates do not depend on the number of substrate molecules (hence the name "zeroth-order ultrasensitivity"). The system then performs a random walk in the number of S molecules and it is thus subject to large fluctuations.

The system we consider for our simulations is defined by the following set of reactions:

Reaction	Rate	
$S+K \rightleftharpoons KS$	$k_\mathrm{a}, k_\mathrm{b}$	(5.25a)
$KS \to K+S_p$	k_1	(5.25b)
$S_p+P \rightleftharpoons PS_p$	$k_\mathrm{a}, k_\mathrm{b}$	(5.25c)
$PS_p \to P+S$	k_2.	(5.25d)

It will be simulated with the BD algorithm in a rectangular box of dimensions $x_\mathrm{box} = 20R, y_\mathrm{box} = 10R, z_\mathrm{box} = 10R$, with a single kinase and phosphatase enzyme, held fixed at distance Δ on the central axis of the box, as depicted in Figure 5.5, which represents a snapshot of the simulation for 50 total substrate molecules. The system is initially pre-

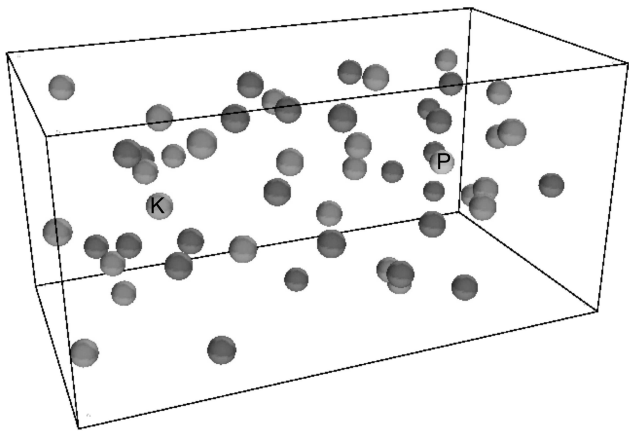


Figure 5.5: Snapshot of the push-pull system. The kinase and the phosphatase molecule are marked by the tag "K" or "P" respectively, and are held fixed along the main axis of the box. The other spheres represent S and S_p molecules, which are free to diffuse in the box. The system is represented for $N_{S\text{tot}} = 50$ and a box of $20R \times 10R \times 10R$.

pared with $N_{S\text{tot}}$ particles, distributed in the two states according to the solution of the macroscopical rate equation. In the following, we investigate the effect of fluctuations on some properties of the system, and we compare the results obtained with the mean-field and master equation approach. In particular, we focus on effects arising from the distribution of reacting sites in space, which can not be captured even by the SSA described in 1.6.1.

We start by computing the sensitivity curve, *i.e.* the mean fraction of phosphorylated substrate molecules $\langle S_p \rangle / N_{S\text{tot}}$ as a function of k_1/k_2. We set 80 substrate molecules in the simulation box, in order to meet the requirement $N_S \gg N_K$ and $N_{S_p} \gg N_P$. The parameters governing the steepness of the sigmoid curves are $K_1 = (k_\text{b} + k_1)/(k_\text{f}[S])$ and $K_2 = (k_\text{b} + k_2)/(k_\text{f}[S_p])$, where $1/k_\text{f} = 1/k_\text{a} + 1/k_\text{D}$. In all our simulations we set $K_1 = K_2 = K_\text{M}$. When $K_\text{M} \ll 1$ the enzymes are totally saturated and the change in the fraction of modified proteins is abrupt; on the other hand, when $K_\text{M} \geq 1$, the rise of the curve is closer to the hyperbolic Michaelis-Menten shape.

Figure 5.6 shows the sensitivity curve, obtained with three different methods: with the analytic macroscopic curve of [164], with SSA simulations as in [169] and with the Brownian Dynamics algorithm. In the SSA runs, the forward rate was set to $k_\text{f} = (1/k_\text{a} + 1/k_\text{D})^{-1}$, and the backward rate was accordingly renormalised: $k'_\text{b} = k_\text{b} k_\text{f}/k_\text{a}$. With this rescaling, we lump the immediate rebindings happening in a spatial simulation into a lower dissociation rate, as it was done in [168]. Since a converted particle must travel to the other reaction site before being brought back to its initial state, we do not need to renormalise k_1 and k_2. Panel A of Figure 5.6 shows the data for $K_\text{M} = 1$ and $K_\text{M} = 0.01$: in the first case, the BD results (circles) show a mild deviation from the analytical curve, while the SSA data closely coincide with it (diamonds). Instead, for $K_\text{M} = 0.01$, the BD algorithm can not reach the steep response predicted at a macroscopical level, and even the SSA displays a slightly lower sensitivity. We repeated the analysis for $K_\text{M} = 200$ (data not shown) and in that case we found perfect agreement between the three sets of data. In this last case both reactions are first-order regime, which means that their rates are proportional to the number of substrate molecules. As a result, when this number

changes, the rates of conversion in the opposite direction changes immediately. This counteracts the modification and reduces the effect of fluctuations. Furthermore, we find another deviation from the mean-field averages: panels B through E in Figure 5.6 show the data obtained by decreasing the diffusion coefficient of the system, and keeping all the other parameters fixed. In panel B, the associations of the substrate molecules to the two enzymes are reaction-limited, because diffusion is very rapid; in this case, we find a situation analogous to that shown in panel A, with SSA perfectly reproducing the data and BD showing a mild deviation. However, when the diffusive motion slows down, the associations become diffusion-limited (K_{M} varies accordingly) and the BD results deviate more and more from the mean-field line, whereas the agreement between the SSA results and the mean-field analysis remains. These results demonstrate that slow diffusion can strongly reduce the sensitivity of the system.

Figure 5.6 confirms and extends then what was found in [169]: the stochastic fluctuations of the system dampen the ultrasensitivity, which could be obtained only in an infinitely large, well-stirred system. In particular, diffusion of substrate molecules is a seriously limiting factor, which can strongly reduce the sensitivity of system, bringing it below the Michaelis-Menten curve. The discrepancy between the numerical data and the macroscopic prediction drastically increases when spatial fluctuations are considered, in particular when the diffusion of molecules is slow. The SSA, instead, deviates from mean-field only when the system is in the ultrasensitive regime (low K_{M}) and the number of molecules present in the system is low. It however predicts a good agreement when K_{M} is high. Brownian Dynamics is able to show that, even in this last regime, the response of the system can be much less sensitive if the species move slowly enough in space, *i.e.* when the system is not well-stirred.

Brownian Dynamics allows us to directly measure spatial properties of the system, such as the spatial density of particles. In Figure 5.7 the density of particles along the main axis of the box is shown, for $K_{\mathrm{M}}=0.2$ and $K_{\mathrm{M}}=0.01$, corresponding to the substrate molecules bound to the enzymes. The S particles spend on average more time closer to the phosphatase enzyme, where they are produced, and less time close to the kinase enzyme, where they are converted to S_p; however, since it is very probable to find an S molecule bound to a kinase, we observe a high peak at the kinase location. The profile for S_p is completely symmetric, as these simulations are obtained for $k_1 = k_2$. When the enzymes are completely saturated (low K_{M}), the fraction of time spent on the enzymes increases, and the regions around the peaks experiences a stronger depletion on the respective substrates.

5.5 Summary

Brownian Dynamics algorithms are widely-utilised techniques in the field of soft condensed matter and biochemistry. Recently, they have been applied to coarse-grained models of cellular processes, which can be viewed as reaction-diffusion systems. A prominent

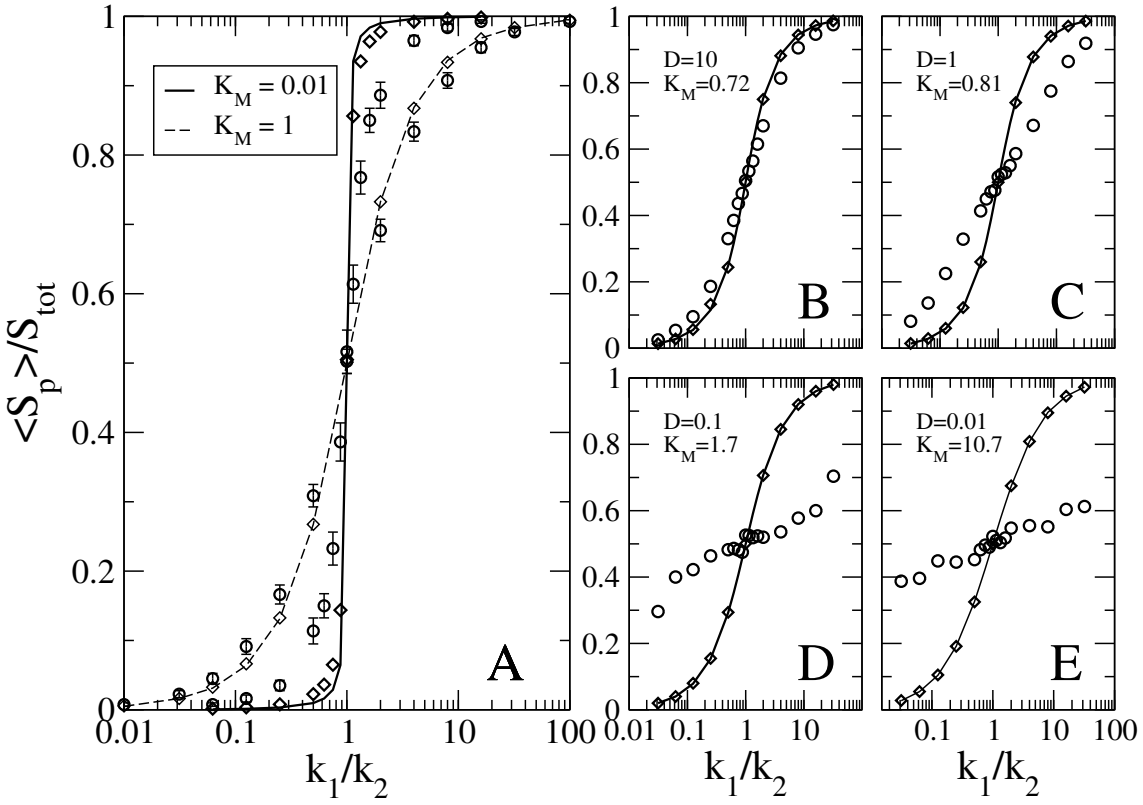


Figure 5.6: Fraction of converted molecules as a function of k_1/k_2. A) The continuous and dashed lines are obtained with Macroscopical Rate Equation, for $K_{\mathrm{M}}=0.01$ (full saturation of enzymes) and $K_{\mathrm{M}}=1$ (first-order regime), respectively. The numerical solutions of the master equation (SSA, diamonds) show a mild deviation only when the system displays an ultrasensitive behavior. Conversely, Brownian Dynamics simulations (circles), do not yield a perfect agreement for $K_{\mathrm{M}}=1$, and markedly deviate from the ultrasensitive line. Methods accounting for the stochastic behavior of the system show thus a reduction in sensitivity for $K_{\mathrm{M}}=0.01$. B) to E) The system is simulated for a decreasing diffusion coefficient. When the diffusion-limited regime is approached, BD simulations are not able to reproduce the sensitivity predicted in the Michaelis-Menten kinetics. The SSA results, instead, closely coincide with mean-field. Accounting for diffusion of substrate molecules thus drastically reduces the sensitivity of the system, and yields a pronounced deviation from mean-field results, even for high K_{M}.

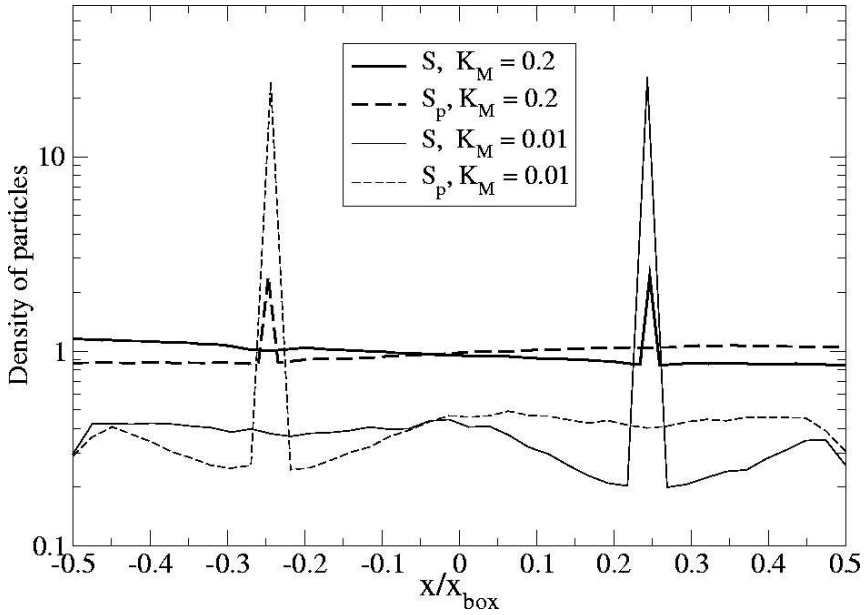


Figure 5.7: The spatial density profiles for S, S_p ($k_1/k_2 = 1$) show clear symmetric gradients. Particles S (continuous line) tend to have a higher concentration around the phosphatase enzyme, where they are produced. However, as the enzymes are saturated, we find a peak closer to the kinase enzyme, which is most the time bound to a S molecule. The profile for S_p (dashed line) is completely symmetric. When the enzymes are completely saturated ($K_\mathrm{M} = 0.01$, thinner lines), they are more depleted in the proximity of reaction sites, where they spend a longer amount of time.

example is represented by biochemical networks, which allow a cell to detect and respond to changes in its environment. These networks are composed of proteins and DNA that chemically and physically interact with each other, and are generally prone to fluctuations. The main sources of these fluctuations can be of two different kinds. First, they can derive from the intrinsic stochasticity of the reactive events, this effect being exacerbated by a low number of molecules in the system. These "temporal" fluctuations are correctly captured even when the system is described at the level of the chemical master equation, and simulated with the SSA algorithm described in Section 5.3. Second, they can be due to the erratic behaviour of diffusing molecules, which causes the association-time distribution to deviate from the poissonian form typical of well-stirred systems [41, 168]. Brownian Dynamics algorithms are able to account for this second source of stochasticity,whereas the SSA is not.

However, in order to avoid the introduction of systematic errors in the simulations, a Brownian Dynamics algorithm must obey the detailed-balance rule. In case of second-order reactions, this is not a trivial requirement. In this Chapter, we have designed a Brownian Dynamics algorithm for reaction-diffusion systems which rigorously obeys detailed balance. In Section 5.3 we have checked that our scheme is able to reproduce the

equilibrium properties of an elementary system. Dynamic properties are also correctly reproduced, provided that the time step of the simulation is small enough.

As an illustrative example, we have applied our BD scheme to a model representing the dynamics of a substrate molecule subject to two antagonistic enzymes. This model was previously analysed with deterministic methods [164], which revealed an ultrasensitive behavior in the response of the system when the enzymes are fully saturated. A study conducted at the level of the Chemical Master Equation [169], thus accounting for the low copy number of the substrate molecules, highlighted that the ultrasensitivity predicted in Ref. [164] cannot be achieved when the concentration of the substrate is very low. Temporal fluctuations limit then the sensitivity of the system. We repeated the analysis of Ref. [169] simulating the system with the SSA, and confirmed their findings. Furthermore, we have investigated the role of spatial fluctuations on the system with BD simulations. Our analysis shows that the sensitivity of the response curve is further reduced. In particular, when diffusion of particles is slow and the system is far from well-stirred, spatial fluctuations are the dominant source of noise, and the reduction of the gain is significant.

Chapter 6

Spatial fluctuations of transcription factors enhance noise in gene expression

But I, being poor, have only my dreams;
I have spread my dreams under your feet.
Tread softly because you tread on my dreams.
William Butler Yeats

We study by Green's Function Reaction Dynamics the effect of the diffusive motion of repressor molecules on the noise in mRNA and protein levels for a gene that is under the control of a repressor. We find that spatial fluctuations due to diffusion can drastically enhance the noise in gene expression. After dissociation from the operator, a repressor can rapidly rebind to the DNA. Our results show that the rebinding trajectories are so short that, on this time scale, the RNA polymerase (RNAP) cannot effectively compete with the repressor for binding to the promoter. As a result, a dissociated repressor molecule will on average rebind many times, before it eventually diffuses away. These rebindings thus lower the effective dissociation rate, and this increases the noise in gene expression. Another consequence of the time scale separation between repressor rebinding and RNAP association is that the effect of spatial fluctuations can be described by a well-stirred, zero-dimensional, model by renormalising the reaction rates for repressor-DNA (un) binding. Our results thus support the use of well-stirred, zero-dimensional models for describing noise in gene expression. We also show that for a fixed repressor strength, the noise due to diffusion can be minimised by increasing the number of repressors or by decreasing the rate of the open complex formation. Lastly, our results emphasise that power spectra are a highly useful tool for studying the propagation of noise through the different stages of gene expression.

6.1 Introduction

Cells process information from the outside and regulate their internal state by means of proteins and DNA that chemically and physically interact with one another. These biochemical networks are often highly stochastic, because in living cells the reactants often occur in small numbers [170, 171]. This is particularly important in gene expression [172, 173, 174, 17, 21, 175, 176, 177], where transcription factors are frequently present in copy numbers as low as tens of molecules per cell. While it is generally believed that biochemical noise can be detrimental to cell function [175], it is increasingly becoming recognised that noise can also be beneficial to the organism [176]. Understanding noise in gene expression is thus important for understanding cell function, and this observation has recently stimulated much theoretical and experimental work in this direction [175, 176, 177]. However, the theoretical analyses usually employ the zero-dimensional chemical master equation [177, 31, 66], which is solved by the Stochastic Simulation Algorithm described in Section 1.6.1. This approach takes into account the discrete character of the reactants and the probabilistic nature of chemical reactions. It does assume, however, that the cell is a 'well-stirred' reactor, in which the particles are uniformly distributed in space at all times; the reaction rates only depend upon the global concentrations of the reactants and not upon the spatial positions of the reactant molecules. Yet, in order to react, reactants first have to move towards one another. They do so by diffusion, or, in the case of eukaryotes, by a combination of diffusion and active transport. Both processes are stochastic in nature and this could contribute to the noise in the network. Here, we study by computer simulation the expression of a single gene that is under the control of a repressor R in a spatially-resolved model. We find that at low repressor concentration, *i.e.* $[R] < 50$nM, the noise in gene expression is dominated by the noise arising from the diffusive motion of the repressor molecules. Our results thus show that spatial fluctuations of the reactants can be an important source of noise in biochemical networks. Our analysis also reveals that the effects of diffusion can nevertheless be described by a well-stirred model, provided that the reaction rates of repressor-DNA (un)binding are properly rescaled.

The simulations show that in gene expression significant fluctuations occur on both short and long length and time scales. As expected from earlier work [109, 178, 18], the fluctuations on long time scales are predominantly due to protein degradation; we assume that proteins are degraded by dilution, which means that the half-time of this process is on the order of an hour. Our results, however, also elucidate an important process on much shorter length and time scales. It is associated with the competition between the repressor and RNA polymerase (RNAP) for binding to the promoter. When a repressor molecule dissociates from the DNA, it can rebind very rapidly: in our model, which neglects 1-dimensional diffusion along the DNA, it can rebind on a time scale of milliseconds, or less. This time scale is much shorter than that with which the RNAP binds to the promoter, which is on the order of $0.01 - 0.1$ seconds. Hence, when a repressor molecule has just dissociated, the probability that an RNAP molecule will bind before the repressor molecule rebinds, is very small. This has two important consequences. The

first is that a repressor molecule will on average rebind many times, before it eventually diffuses away from the promoter and an RNAP molecule, or another repressor molecule, can bind to the promoter. This decreases the effective dissociation rate, which introduces long time scales fluctuations in transcription and therefore increases the noise in gene expression.

The second consequence of the rapidity of the rebindings is that noise propagation during gene expression can be described by a well-stirred, zero-dimensional, model. In the commonly used zero-dimensional models, chemical reactions are separated by exponentially distributed waiting times. In a spatially-resolved model, the distribution of repressor-DNA association times deviates markedly from Poisson statistics. While at long time scales the distribution is exponential, at short time scales it is algebraic, due to the diffusive nature of the rebinding trajectories. However, these repressor rebindings are so fast, that they do not significantly affect the dynamics of RNAP-DNA association; the latter is only affected by the repressor-DNA (un)binding dynamics at longer time scales, which obey Poisson statistics. The reason that the effect of spatial fluctuations on noise in gene expression can be described by a zero-dimensional model is thus a separation of time scales. In fact, it is conceivable that in a more realistic model of gene expression, which includes 1-dimensional sliding along the DNA, the time scale of repressor rebinding is not separated from that of the RNAP dynamics. Under these conditions, the effect of spatial fluctuations might be detected in the statistics of mRNA production. However, as we discuss in the Discussion and Outlook Section, the noise strength (variance) of the mRNA level can probably still be described by a zero-dimensional model, because the time scale of the spatial fluctuations, even in those more refined models, is expected to be still shorter than the typical life time of an mRNA molecule.

Since fluctuations in the rate of gene expression span orders of magnitude in length and time scales, the simulation technique should be sufficiently detailed to resolve the events at short length and time scales, yet also efficient enough to access the long length and time scales. Recently, several simulation techniques have been developed for the stochastic modeling of reaction-diffusion systems [38, 39]. These techniques, however, do not satisfy both criteria: they either describe the system in a coarse-grained way, *i.e.* on the level of local concentrations rather than single particles [38, 39], or are too slow to accurately model the dynamics on the long time scales [159]. Our simulations have been made possible via the use of our recently developed Green's Function Reaction Dynamics (GFRD) algorithm, described in Section 1.6.3 and in Refs. [40, 41]. GFRD is an event driven algorithm that uses Green's functions to combine in one step the propagation of the particles in space with the reactions between them. The event-driven nature of the algorithm makes it particularly useful for problems, such as gene expression, in which the events are distributed over a wide range of length and time scales: the algorithm takes small steps when the reactants are close to each other – such as when a repressor molecule has just dissociated from the DNA – while it takes large jumps in time and space when the molecules are far apart from each other – like when the repressor molecule has eventually diffused away from the promoter. The event-driven nature of GFRD makes

it orders of magnitude more efficient than brute-force particle-based algorithms [41] and this has allowed us to simulate gene expression on the relevant biological time scale of hours.

Several publications [179, 180, 52, 181, 182, 183, 184, 185, 186] have discussed the effect of fluctuations in the binding of transcription factors to their site on the DNA (called operator) on the noise in gene expression. Most of these models are relatively simple, ignoring, for instance, production of mRNA [52, 182, 183, 186, 181]. Moreover, all these studies, with the exception of [181, 185], ignore the role of the spatial fluctuations of the transcription factors. Our aim is to study gene expression in a biologically meaningful model. We have therefore constructed a rather detailed model, although we will also use minimal models that can be studied analytically, in order to interpret the simulation results. The full model, which is described in the next Section, contains the diffusive motion of repressor molecules, open complex formation, promoter clearance, transcription elongation and translation [65].

In Section 6.4, we discuss the simulation results for both the noise in mRNA and in protein level. The results reveal that for $[R] < 50$nM, the noise in the spatially-resolved model can be more than five times larger than the noise in the well-stirred model. We also show that a cell could minimise the effect of spatial fluctuations, either by tuning the open complex formation rate or by changing the number of repressors and their affinity for the binding site on the DNA. In Section 6.5, we elucidate the origin of the enhanced noise in the spatially resolved model. In the subsequent Section, we show that in the model employed here the effect of spatial fluctuations can be quantitatively described by a well-stirred model in which the reaction rates for repressor binding and unbinding are appropriately renormalised; however, as alluded to above, and as we will discuss in more detail in the last Section, we expect that in a more refined model the effect of diffusion will be more complex, impeding such a simplified description. In Section 6.7, we discuss how the operator state fluctuations propagate through the different stages of gene expression using power spectra for the operator state, the elongation complex, the mRNA and the protein. The results show that these power spectra are highly useful for unraveling the dynamics of gene expression. We hope that this stimulates experimentalists to measure power spectra of not only mRNA and protein levels [30], but also of the dynamics of transcription initiation and elongation using *e.g.* magnetic tweezers [187]. As we argue in the last Section, such experiments should make it possible to determine the importance of spatial fluctuations for the dynamics of gene expression.

6.2 Model

6.2.1 Diffusive motion of repressors

We explicitly simulate the diffusive motion of the repressor molecules in space. However, since the experiments of Riggs *et al.* [188] and the theoretical work of Berg, Winter, and Von Hippel [189], it is well known that proteins could find their target sites via a

combination of 1D sliding along the DNA and 3D diffusion through the cytoplasm – "hopping" or "jumping" from one site on the DNA to another. This mechanism could speed up the search process and make it faster than the rate at which particles find their target by free 3D diffusion; this rate is given by $k = 4\pi\sigma D_3[R]$, where σ is the interaction radius, which is on the order of a protein diameter or DNA diameter, D_3 is the diffusion constant of the protein in the cytoplasm, and $[R]$ is the concentration of the (repressor) protein. However, while it is clear that the mechanism of 3D diffusion and 1D sliding could potentially speed up the search process, whether this mechanism in living cells indeed drastically reduces the search time is still under debate [190]. In this context, it is instructive to discuss the two main results of recent studies on this topic [191, 192, 193, 190, 194, 195]. The first is that the mean search time τ is given by [195]

$$\tau \sim \frac{L}{\lambda}\left[\frac{\lambda^2}{D_1} + \frac{r^2}{D_3}\right], \tag{6.1}$$

where L is the total length of the DNA, λ is the average distance over which the protein slides along the DNA before it dissociates, D_1 is the diffusion constant for sliding, r is the typical mesh size in the nucleoid (the characteristic distance between two segments on the DNA [195]), and D_3 is the diffusion constant in the cytoplasm. This formula has a clear interpretation [195]: λ^2/D_1 is the sliding time, r^2/D_3 is the time spent on 3D diffusion, the sum of these terms is thus the time to perform one round of sliding and diffusion, and L/λ is the total number of rounds needed to find the target. The other principal result is that the search time is minimised when the sliding distance λ is

$$\lambda = \sqrt{\frac{D_1}{D_3}}r. \tag{6.2}$$

Under these conditions, a protein spends equal amounts of time on 3D diffusion and 1D sliding (a protein is thus half of the time bound to the DNA). Eq. (6.2) is a useful result, because it shows that the average sliding distance λ depends upon the ratio of diffusion constants and on the typical mesh size in the nucleoid. If we now assume that D_1 and D_3 are equal (which is not obvious given that proteins bind relatively strongly to DNA – D_1 could thus very well be much smaller than D_3) and if we take the mesh size to be given by $r \sim \sqrt{v/L}$ [195], where $v \approx 1\mu\text{m}^3$ is the volume of an *E. coli* cell and $L \approx 10^3\mu$m, we find that λ is in the order of 10nm (30 bp). This corresponds to the typical diameter of a protein or DNA double helix and is thus not very large. Interestingly, recent experiments seem to confirm this: experiments from Halford *et al.* on restriction enzymes (EcoRV and BbcCI) with a series of DNA substrates with two target sites and varying lengths of DNA between the two sites, suggest that under the *in vivo* conditions, sliding is indeed limited to relatively short distances, *i.e.* to distances less than 50 bp ($\approx$ 16nm) [196, 197].

Now, it should be realised that on length scales beyond the sliding length, the motion is essentially 3D diffusion: the sliding/hopping mechanism corresponds to 3D diffusion with a jump distance given by the sliding distance [191]. Moreover, since the sliding distance is only on the order of a particle diameter, as discussed above, we have therefore

decided to model the motion of the repressor molecules as 3D diffusion. But it should be remembered that on length scales shorter than $10-30$nm, this approach is not correct. As we discuss in the Discussion Section, this might have significant implications for the importance of spatial fluctuations for the noise in gene expression.

6.2.2 Transcription and Translation

Most repressors bind to a site that (partially) overlaps with the core promoter – the binding site of the RNA polymerase (RNAP). When a repressor molecule is bound to its operator site, it prevents RNAP from binding to the promoter, thereby switching off gene expression. Only in the absence of a repressor on the operator site, can RNAP bind to the promoter and initiate transcription and translation, ultimately resulting in the production of a protein. We model this by the following reaction network:

$$O + R \underset{k_{\mathrm{bR}}}{\overset{k_{\mathrm{fR}}}{\rightleftarrows}} OR \tag{6.3}$$

$$O \underset{k_{\mathrm{bRp}}}{\overset{k_{\mathrm{fRp}}}{\rightleftarrows}} ORp \tag{6.4}$$

$$ORp \xrightarrow[k_{\mathrm{OC}}]{} ORp^* \tag{6.5}$$

$$ORp^* \xrightarrow[t_{\mathrm{clear}}]{} T + O \tag{6.6}$$

$$T \xrightarrow[t_{\mathrm{elon}}]{} M \tag{6.7}$$

$$M \xrightarrow[k_{\mathrm{dm}}]{} \emptyset \tag{6.8}$$

$$M \xrightarrow[k_{\mathrm{ribo}}]{} M + M_{\mathrm{ribo}} \tag{6.9}$$

$$M_{\mathrm{ribo}} \xrightarrow[t_{\mathrm{trans}}]{} P \tag{6.10}$$

$$P \xrightarrow[k_{\mathrm{dp}}]{} \emptyset \tag{6.11}$$

Eqs. (6.3) and (6.4) describe the competition between the binding of the repressor R and the RNAP molecules Rp to the promoter (O is the operator site). In our simulation we fix the binding site O in the center of a container with volume $V = 1\mu\mathrm{m}^3$, comparable to the volume of a single *E. coli* cell. We simulate both the operator site O and the repressor molecules as spherical particles with diameter $\sigma = 10$nm. The operator site O is surrounded by N_R repressor molecules that move by free 3D diffusion (see previous Section) with an effective diffusion constant $D = 1\mu\mathrm{m}^2\mathrm{s}^{-1}$, as has been reported for proteins of a similar size [93]. The intrinsic forward rate $k_{\mathrm{fR}} = 6 \cdot 10^9 \mathrm{M}^{-1}\mathrm{s}^{-1}$ for the repressor particles R at contact is estimated from the Maxwell-Boltzmann distribution [40]. The backward rate k_{bR} depends on the interaction between the DNA binding site of the repressor and the operator site on the DNA and varies greatly between different operons, with stronger

repressors having a lower k_{bR}. In our simulations, we vary k_{bR} between $1-0.01\ \mathrm{s}^{-1}$, as discussed in more detail below. The concentration of RNAP is much higher than that of the repressor [198]. Because of this we treat the RNAP as distributed homogeneously within the cell and we do not to take diffusion of RNAP into account explicitly. Instead, RNAP associates with the promoter with a diffusion-limited rate $k_{\mathrm{fRp}} = 4\pi\sigma D[Rp]$. In our simulations, the concentration of free RNAP is $[Rp] = 0.5\mu\mathrm{M}$ [198], leading to a forward rate $k_{\mathrm{fRp}} = 38\mathrm{s}^{-1}$. Finally, the backward rate $k_{\mathrm{bRp}} = 0.5$ is determined such that $K_{\mathrm{eq}} = 4\pi\sigma D/k_{\mathrm{bRp}} = 1.4\cdot 10^{9}\mathrm{M}^{-1}$ [80].

Transcription initiation is described by Eqs. (6.5) and (6.6). Before productive synthesis of RNA occurs, first the RNAP in the RNAP-promoter complex ORp unwinds approximately one turn of the promoter DNA to form the open complex ORp^{*}. The open complex formation rate k_{OC} has been measured to be on the order of $0.3-3\mathrm{s}^{-1}$ [187]. We approximate open complex formation as an irreversible reaction. Some experiments find this step to be weakly reversible [187]. However, adding a backward reaction to the model did not change the dynamics of the system in a qualitative way, as long as the backward rate is smaller than k_{OC}, which is in agreement with experimental results. After open complex formation, RNAP must first escape the promoter region before another RNAP or repressor can bind. Since elongation occurs at a rate of $50-100$ nucleotides per second and between $30-60$ nucleotides must be cleared by RNAP before the promoter is accessible, a waiting time of $t_{\mathrm{clear}} = 1\mathrm{s}$ is required before another binding can occur. Since promoter clearance consists of many individual elongation events that obey Poisson statistics individually, we model the step as one with a fixed time delay t_{clear}, not as a Poisson process with rate $1/t_{\mathrm{clear}}$.

Eqs. (6.7)-(6.11) describe the dynamics of mRNA and protein numbers. After clearing the promoter region, RNAP starts elongation of the transcript T. As for clearance, the elongation step is modeled as a process with a fixed time delay $t_{\mathrm{elon}} = 30\mathrm{s}$, corresponding to an elongation rate of $50-100$ nucleotides per second [199] and a 1500 bp gene. When an mRNA M is formed, it can degrade with a rate k_{dm}. Here, the mRNA degradation rate is determined by fixing the average mRNA concentration in the unrepressed state, as described below. Furthermore, an mRNA molecule can form an mRNA-ribosome complex M_{ribo} and start translation. We assume that $b = 5$ proteins are produced on average from a single mRNA molecule [21], so that the start of translation occurs at a rate $k_{\mathrm{ribo}} = b\, k_{\mathrm{dm}}$. Assuming a translation speed of about 50 nucleotides/s [107], after a fixed time delay $t_{\mathrm{trans}} = 30\mathrm{s}$ a protein P is produced. The mRNA is available for ribosome binding immediately after the start of translation. Due to the delay in protein production, M can start to be degraded, while the mRNA-ribosome complex M_{ribo} is still present; M thus represents the mRNA leader region rather than the entire mRNA molecule. Finally, the protein P degrades at a rate k_{dp}, which is determined by the requirement that the average protein concentration in the unrepressed state has a desired value, as we describe now.

We vary the free parameters in the reaction network described in Eqs. (6.3)-(6.11) – N_R, k_{bR}, k_{dm}, k_{dp} – in the following way: first, we choose the concentration of mRNA and protein in the absence of repressor molecules. In this case, tuning of the concentrations

is most straightforward by adjustment of the mRNA and protein decay rates k_{dm} and k_{dp}. For the above reaction network one can show that the average mRNA number $N_M$and protein number N_P is given by

$$N_M = \frac{K_4 K_1 V}{K_2 N_R + V(1+K_1(1+K_3))}, \tag{6.12}$$

$$N_P = K_5 N_M, \tag{6.13}$$

where $K_1 = k_{\mathrm{fRp}}/(k_{bRp} + k_{\mathrm{OC}})$, $K_2 = k_{\mathrm{fR}}/k_{\mathrm{bR}}$, $K_3 = k_{\mathrm{OC}} t_{\mathrm{clear}}$, $K_4 = k_{\mathrm{OC}}/k_{\mathrm{dm}}$ and $K_5 = k_{\mathrm{ribo}}/k_{\mathrm{dp}}$ are equilibrium constants, V is the volume of the cell and N_R is the total number of repressors. The unrepressed state corresponds to $N_R = 0$. In our simulations, we fix the mRNA and protein numbers in the unrepressed state at $N_M = 50$ and $N_P = 2 \cdot 10^5$. The mRNA and protein decay rates then follow straightforwardly from Eqs. (6.12) and (6.13): the mRNA degradation rate is $k_{\mathrm{dm}} = 0.019\mathrm{s}^{-1}$ [200] and the protein degradation rate is $k_{\mathrm{dp}} = 2.4 \times 10^{-4}\mathrm{s}^{-1}$; the latter corresponds to protein degradation by dilution with a doubling time of around 1h.

Next, we determine by what factor these concentrations should decrease in the repressed state. This can be done by changing the number of repressors N_R and the repressor backward rate k_{bR}. We define the repression level f as the transcription initiation rate in the absence of repressors, divided by the initiation rate in the repressed state [146]. For a repression level f, the concentration of mRNA and proteins in the repressed state is a fraction $1/f$ of the concentration in the unrepressed state and it follows that

$$\frac{N_R}{k_{\mathrm{bR}}} = (f-1)\frac{V(1+K_1(1+K_3))}{k_{\mathrm{fR}}}. \tag{6.14}$$

Thus, a fixed repression level f does not specify a unique combination of N_R and k_{bR}: increasing the number of repressors twofold, while also increasing the repressor backward rate by the same factor, gives the same repression level. This means that the cell can control mRNA and protein levels in the repressed state either by having a large number of repressors that stay on the DNA for a short time or by having a small number of repressors, possibly even one, that stay on the DNA for a long time. Even though it is conceivable that the latter is preferable for economic reasons, there is no difference between the two extremes in terms of the average gene expression. In our simulations, we vary N_R and k_{bR}, but use a fixed repression level $f = 100$. Consequently, in the repressed state, on average $N_{\mathrm{M}} = 0.5$ and $N_{\mathrm{P}} = 200$.

Lastly, we would like to emphasise that, while all reaction rates were, as much as possible, taken from experiments, it should be realised that the measured rates might not be very precise. However, we believe that this does not affect the main conclusions of our work.

6.3 Simulation Technique

We simulate the above reaction network using Green's Function Reaction Dynamics (GFRD) [40, 41]. GFRD is an event-driven algorithm, which combines in one step the propagation of the particles in space with the reactions between them. The main idea is to determine at each iteration of the simulation, a maximum time step, such that only single particles or pairs of particles have to be considered. For these particles, the Smoluchowski equation [201] can be solved exactly using Green's functions. For each single particle, the Green's function is just the Gaussian distribution function $p_1(\mathbf{r},t|\mathbf{r}_0,t_0)$, which yields the probability that, given that the particle is at point $\mathbf{r}_0$ at time t_0, it is at position $\mathbf{r}$ at a later time t. For each pair of particles, two Green's functions are obtained: one for their center-of-mass, and one for their inter-particle vector $\mathbf{r}$; the latter, $p_2(\mathbf{r},t|\mathbf{r}_0,t_0)$, yields the probability that the inter-particle vector $\mathbf{r}_0$ at time t_0 becomes $\mathbf{r}$ at a later time t. Importantly, the inter-particle Green's function does not only take into account the diffusion of the particles, but also the reactions between them. This makes it possible to derive for each pair of particles the propensity function $q(t|r_0)$, which yields the probability that the pair will react for the first time at time t, given that the particles were separated by a distance r_0 initially. The propensity functions, then, can be used to set up an event-driven algorithm, quite analogous to kinetic Monte Carlo algorithms for zero-dimensional master equations, such as the Gillespie algorithm [35] (see Section 1.6.1). The event-driven nature allows GFRD to make large jumps in time and space when the particles are far apart from each other, making it up to five orders of magnitude more efficient than brute-force Brownian Dynamics. For details of the algorithm, in particular on how the Green's functions and the propensity functions are derived, we refer to Refs. [40, 41].

As discussed above, only the operator site O and the repressor particles R are simulated in space. All other reactions are assumed to occur homogeneously within the cell and are simulated according to the well-stirred model [35] or with fixed time delays for reaction steps involving elongation. A few modifications with respect to the algorithm described in [40, 41] are implemented to improve simulation efficiency. First, we neglect excluded volume interactions between repressor particles mutually, as the concentration of repressor is very low. This means that the only potential reaction pairs we consider are operator-repressor pairs. Secondly, we use periodic boundary conditions instead of a reflecting boundary, which leads to a larger average time step. As the operator site O is both small compared to the volume of the cell and is far removed from the cell boundary, this has no effect on the dynamics of the system. Moreover, when the reaction times drawn by GFRD are exceedingly small, the Green's Function for the pairs of particles become hard to evaluate numerically. In this case, we prefer to break down this reaction time in 10^4 substeps and run the Brownian Dynamics algorithm described in Chapter 5 on the system. Finally, as the repressor backward rate k_{bR} is rather small, the operator site can be occupied by a repressor for a time long compared to the average simulation time step. If the repressor is bound to the operator site longer than a time $L^2/6D$, where L is the length of the sides of our container, the other repressor molecules diffuse on average from one side of the box to the other. Consequently, when the repressor eventually dissociates from

the operator site, the other repressor molecules have lost all memory of their positions at the time of repressor binding. Here, when a repressor will dissociate after a time longer than $L^2/6D$, we do not propagate the other repressors with GFRD, but we only update the master equation and fixed delay reactions. We update the positions of the free repressors at the moment that the operator site becomes accessible again, by assigning each free repressor molecule a random position in the container; the dissociated repressor is put at contact with the operator site. We see no noticeable difference between this scheme and results obtained by the full GFRD algorithm described in Refs. [40, 41] and Section 1.6.3.

In order to obtain accurate statistics, especially for notoriously difficult quantities such as noise and power spectra, very long simulations were performed. A total number of 24 simulations were performed, one for each combination of parameter values (N_{R}, k_{OC}). A single simulation took on average 24 hrs of CPU time on a Pentium IV 3.0GHz processor.

6.4 Simulation results: dynamics and noise

To study the effect of spatial fluctuations on the repression of genes, we simulate the reaction network described in Eqs. (6.3-6.11) both by GFRD, thus explicitly taking into account the diffusive motion of the repressor particles, and according to the well-stirred model, where the repressor particles are assumed to be homogeneously distributed in space and the dynamics depends only on the concentration of repressor. In Figure 6.1 we show the behaviour of mRNA and protein numbers for a system with open complex formation rate $k_{\mathrm{OC}} = 30\mathrm{s}^{-1}$ and with varying numbers of repressors N_R. We keep the repression factor fixed at $f = 100$ so that with increasing N_R the repressor backward rate k_{bR} is also increased, *i.e.* repressor particles are bound to the DNA for a shorter time.

It is clear from Figure 6.1 that there is a dramatic difference between the behaviour of mRNA and protein numbers between the GFRD simulation and the well-stirred model. When spatial fluctuations of the repressor molecules are included, mRNA is no longer produced in a continuous fashion, but instead in sharp, discontinuous bursts during which the mRNA level can reach levels comparing to those of the unrepressed state. These bursts in mRNA production consequently lead to peaks in protein number. As the protein decay rate is much lower than that of mRNA, these peaks are followed by periods of exponential decay over the course of hours. Due to these fluctuations, protein numbers often reach levels of around $5-10\%$ of the protein levels in the unrepressed state. In contrast, in the absence of repressor diffusion, the fluctuations around the average protein number are much lower. For both cases, however, the average behaviour is identical: even though the dynamics is very different, we always find that on average $\langle N_{\mathrm{mRNA}} \rangle = 0.5$ and $\langle N_P \rangle = 200$. Also, in all cases the fluctuations in mRNA number are larger than those in protein number. This means that the translation step functions as a low-pass filter to the repressor signal.

When we increase the number of repressors N_R and change k_{bR} in such a way that the repression level f remains constant, we find that both for GFRD and the well-stirred model the fluctuations in mRNA and protein number decrease. In the absence of spatial

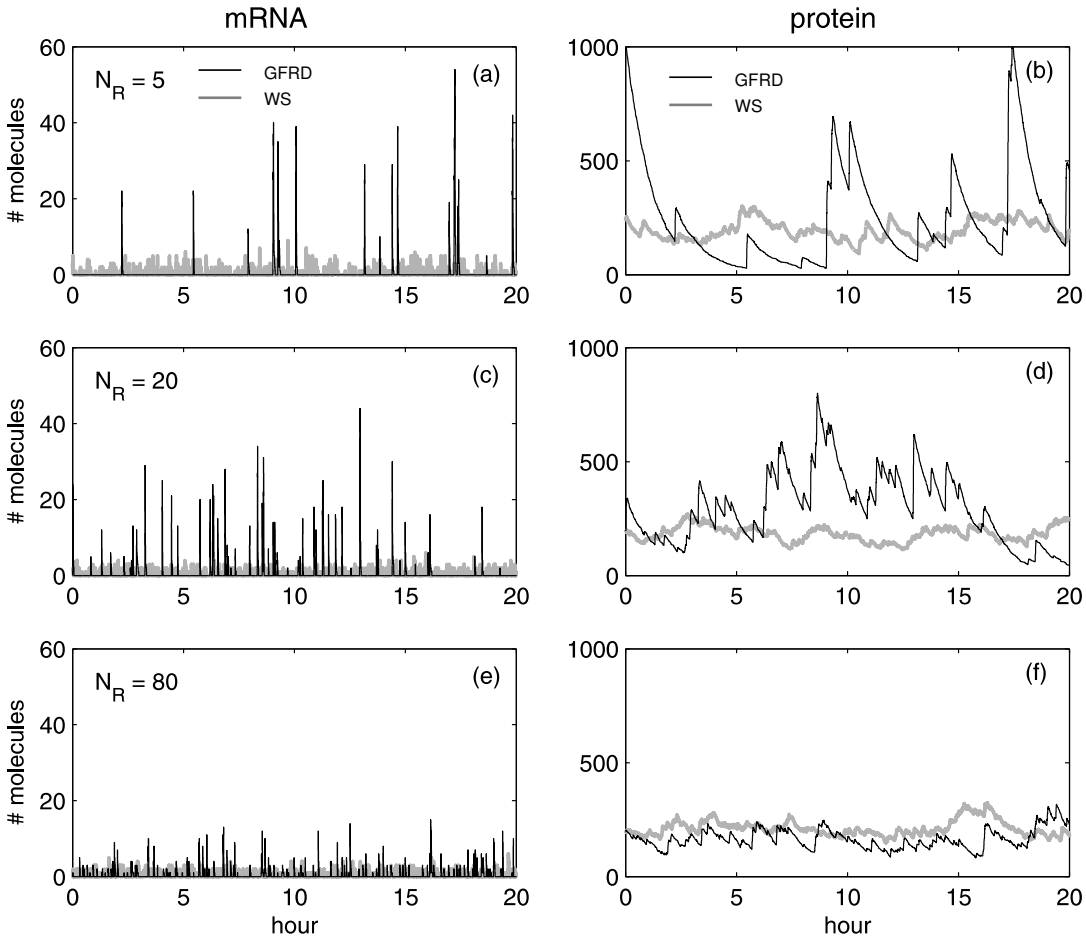


Figure 6.1: Dynamics of mRNA and protein numbers in the repressed state for different number of repressors N_R. The number of mRNA and protein molecules is shown for simulations with GFRD (black line) and according to the zero-dimensional master equation, the well-stirred model (WS, gray line). In the GFRD simulation, diffusion of repressor particles is explicitly included. (a) and (b) $N_R = 5$. (c) and (d) $N_R = 20$. (e) and (f) $N_R = 80$. In general, there is a dramatic difference in dynamics due to the spatial fluctuations of the repressor molecules. This difference becomes more pronounced as the number of repressors decreases. However, we find that in all cases $\langle N_{\mathrm{M}} \rangle = 0.5$ and $\langle N_P \rangle = 200$, on average. The rate of open complex formation $k_{\mathrm{OC}} = 30\mathrm{s}^{-1}$ and, when N_R varies, the repression strength is kept constant at $f = 100$ by changing k_{bR} (see Eq. (6.14)) .

fluctuations this effect is minor, but for GFRD this decrease is sharp: for large number of repressors, the burst in mRNA become both weaker and more frequent. This in turn leads to smaller peaks and shorter periods of exponential decay in protein numbers. In fact, as N_R is increased both approaches converge to the same behaviour. At around $N_R \approx 100$, the dynamics of the protein number is similar for the well-stirred model and the spatially resolved model. The same happens for mRNA number when $N_R \approx 500$.

In Figure 6.2, we quantify the noise in mRNA and protein number, defined as standard deviation divided by the mean, while we change the number of repressors N_R. As we keep the amount of repression fixed at $f = 100$, we simultaneously vary the backward rate k_{bR} according to Eq. (6.14). When all parameters are the same, the noise for the GFRD simulation, including the diffusive motion of the repressors, is always larger than the noise for the well-stirred model, where the diffusive motion is ignored. In both cases, the noise decreases when the number of repressors is increased and the repressor backward rate becomes larger. This is consistent with the mRNA and protein tracks shown in Figure 6.1. We also investigated the effect of changing the open complex formation rate k_{OC}. In nature, this rate can be tuned by changing the base pair composition of the promoter region on the DNA. When we change k_{OC}, we change the mRNA decay rate k_{dm} so that the average mRNA and protein concentrations remain unchanged (see Section 6.2.2). We find that when k_{OC} is lowered, the fluctuations in mRNA and protein levels are sharply reduced. When k_{OC} is much larger than the RNAP backward rate $k_{\mathrm{bRp}} = 0.5\mathrm{s}^{-1}$, almost every RNAP binding to the promoter DNA will result in transcription of an mRNA. For k_{OC} smaller than k_{bRp}, RNAP binding will lead to transcription only infrequently. As a consequence, the operator filters out part of the fluctuations in RNAP binding due to the diffusive motion of the repressor particles, leading to the decrease in noise observed in Figure 6.2. This shows that the open complex formation rate plays a considerable role in controlling noise in gene expression.

6.5 Simulations results: operator binding

To understand how the diffusive motion of repressor molecules leads to increased fluctuations in mRNA and protein numbers, it is useful to look in some detail at the dynamics of repressor-DNA binding. In Figure 6.3A, we show the *OR* bias for both GFRD and the well-stirred model. The *OR* bias is a moving time average over $OR(t)$ with a 50s time window and should be interpreted as the fraction of time the operator site was bound by repressor particles over the last 50 seconds. The results we show here are for $N_R = 5$ repressors and a repression factor $f = 2$. At this repression factor, k_{bR} is such that the repressor molecules are bound to the operator only fifty percent of the time, making it easier to visualise the operator dynamics than in the case of $f = 100$ as used above.

The *OR* bias for the well-stirred model fluctuates around the average value $\langle OR \rangle = 0.5$, indicating that on the timescale of 50s several binding and unbinding events occur, in agreement with $k_{\mathrm{bR}} = 1.26\mathrm{s}^{-1}$ for $f = 2$. On the other hand, when including spatial fluctuations, the *OR* bias switches between periods in which repressors are bound to the DNA

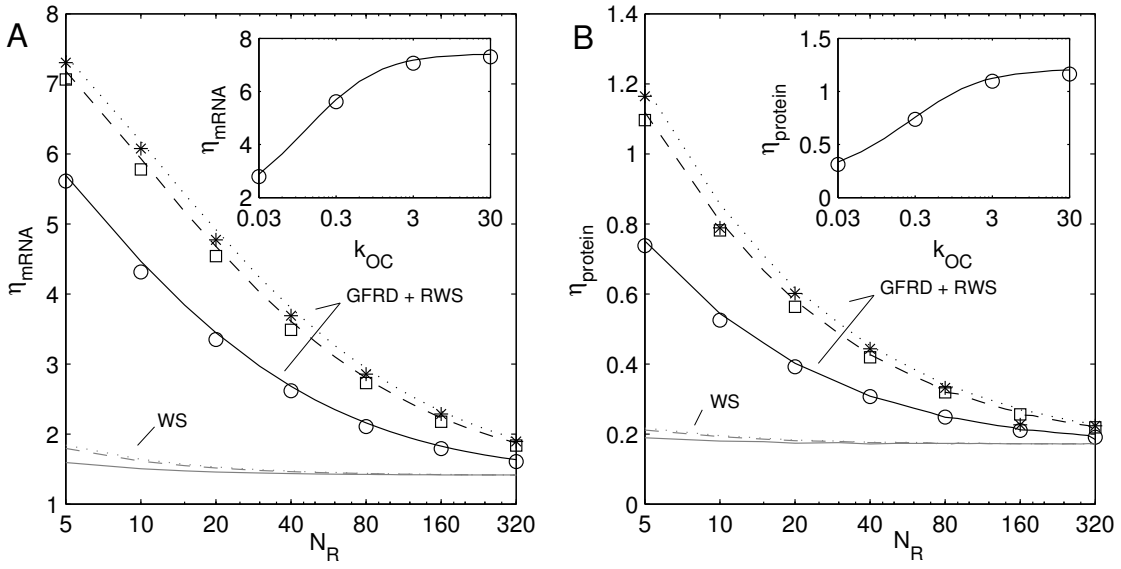


Figure 6.2: Noise in (a) mRNA number and (b) protein number as a function of the number of repressors N_R and for constant repression factor $f = 100$. Data obtained by GFRD simulation is shown for $k_{OC} = 0.3(\circ), 3(\square)$ and $30(*)\text{s}^{-1}$. Noise levels for the well-stirred model (WS) are shown as grey lines and those for the well-stirred model with reaction rates renormalised according to Eqs. (6.16) and (6.17) (the renormalised well-stirred (RWS) model) are shown as black lines, both for $k_{OC} = 0.3$ (solid lines), 3 (dashed lines) and 30 (dotted lines) s^{-1}. Only when the reaction rates are properly renormalised does the noise in the well-stirred model agree well with the noise in the GFRD simulations, which include the effect of diffusion. (Insets) Noise levels as a function of k_{OC} for $N_R = 5$. Symbols indicate results for GFRD and lines are results for the chemical master equation with renormalised reaction rates (RWS model).

continuously and periods in which the repressors are virtually absent, both on timescales much longer than the 50s time window. How is it possible that repressors are bound to the operator site for times much longer than the timescale set by the dissociation rate from the DNA? The answer to that question can be found in Figures 6.3B and C, where a time trace is shown of the operator occupancy by the repressor for both GFRD and the well-stirred model. The time trace for the simulation of the well-stirred model in Figure 6.3C shows a familiar picture: binding and dissociation of the repressor from the operator occurs irregularly, the time between events given by Poisson distributions. The time trace for GFRD in Figure 6.3B looks rather different. Here, in general a dissociation event is followed by a rebinding very rapidly. Only occasionally does a dissociation result in the operator being unbound by repressors for a longer time. When this happens, repressors stay away from the operator for a time much longer than the typical time separating binding events in Figure 6.3C. These series of rapid rebindings followed by periods of prolonged absence

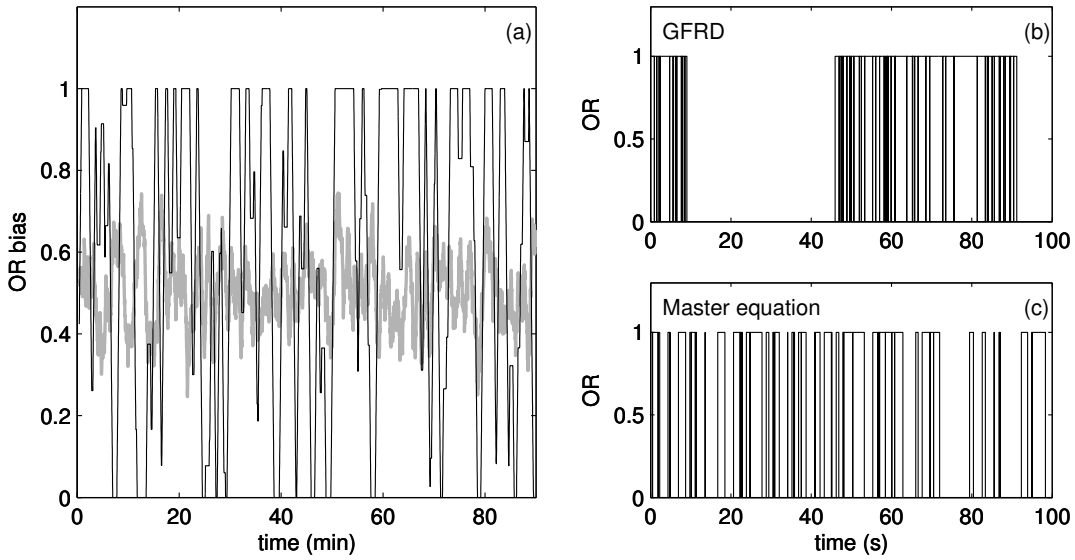


Figure 6.3: Dynamics of repressor binding for a repression factor of $f=2$ and $N_R=5$. (a) The *OR*-bias for GFRD (black line) and the well-stirred model (gray line). The *OR*-bias is defined as the fraction of time a repressor is bound to the operator site in the last 50 seconds. When the diffusive motion of repressor molecules is included (black line), the *OR*-bias switches between periods where repressors are continuously bound to or absent from the DNA for long times. (b) and (c) Time trace of the occupancy of the operator site by repressor molecules. When $OR=1$ a repressor is bound to the operator site and $OR=0$ indicates either a free operator site or one with RNAP bound. For the GFRD simulations, an initial binding is followed by several rapid rebindings, whereas for the well-stirred model binding and rebinding is much more unstructured. Note that here, for reasons of clarity, $f=2$ instead of $f=100$ as used in the text and Figures 6.1 and 6.2.

from the operator result in the aberrant *OR* bias shown in Figure 6.3A.

The occurrence of rapid rebindings is intimately related to the nature of diffusion. When diffusion and the positions of the reactants are ignored all dynamics is based only on the average concentration of the reactants. As a consequence, when in this approach a repressor dissociates from the operator site, the probability of rebinding depends only on the concentration of repressor in the cell. On the level of actual positions of the reactants, this amounts to placing the repressor at a random position in the container. The situation is very different for the GFRD approach, where the positions of the reactants are taken into account. After a dissociation from the operator site, the repressor particle is placed at contact with the operator site. Because of the close proximity of the repressor to its binding site, it has a high probability of rapidly rebinding to, and only a small probability of diffusing away from, the binding site. At the same time, when the repressor eventually diffuses away from the operator site, the probability that the same, or more likely, another

repressor diffuses to and binds the operator site is much smaller than the probability of binding in the well-stirred model, as will be shown quantitatively in Section 6.6. This results in the behaviour observed in Figure 6.3B.

It can now be understood that the bursts in mRNA production correspond to the prolonged absence of repressor from the operator site compared to the well-stirred model. Especially for low repressor concentrations, these periods of absence can be long enough that the concentration of mRNA reaches values comparable to those in the unrepressed state for brief periods of time. When a repressor binds to the operator site, due to the rapid rebindings it will remain bound effectively for a time much longer than the mRNA lifetime, leading to long periods where mRNA is absent in the cell. This shows that under these conditions spatial fluctuations and not stochastic chemical kinetics are the dominant contribution to the noise in mRNA and protein numbers in the repressed state.

6.6 Two-step kinetic scheme

In the current model, the average repressor concentration profile is uniform. It is therefore natural to investigate to what extent the effect of diffusion on the repressor dynamics can be described by a zero-dimensional, well-stirred, model, via the following two-step kinetic scheme [202, 203]:

$$O + R \underset{k_-}{\overset{k_+}{\rightleftarrows}} O\cdots R \underset{k_d}{\overset{k_a}{\rightleftarrows}} OR. \tag{6.15}$$

The first step in Eq. (6.15) describes the diffusion of repressor to the operator site resulting in the encounter complex $O\cdots R$, with the rates k_+ and k_- depending on the diffusion coefficient D and the size of the particles. The next step describes the subsequent binding of repressor to the DNA. In this case the rates are related to the microscopic rates defined in Eq. (6.3). When the encounter complex is assumed to be in steady state, the two-step kinetic scheme can be mapped onto the reaction described in Eq. (6.3), but with effective rate constants $k'_{fR} = k_+ k_a/(k_- + k_a)$ and $k'_{bR} = k_- k_d/(k_- + k_a)$ [202]. The two-step kinetic scheme should yield the same average concentrations as the scheme in Eq. (6.3), so that the equilibrium constant $K = k_a/k_d = k'_{fR}/k'_{bR} = k_{fR}/k_{bR}$, where k_{fR} and k_{bR} are the reaction rates defined in Eq. (6.3).

It is possible to express the effective rate constants k'_{fR} and k'_{bR} in terms of the microscopic rate constants k_{fR} and k_{bR}. For the setup used here, where a single operator O is surrounded by a homogeneous distribution of repressor R, the rate k_+ follows from the solution of the steady state diffusion equation with a reactive boundary condition with rate $k = k_a$ at contact [201, 203] and is given by the diffusion-limited reaction rate $k_D = 4\pi\sigma D$. The rates k_- and k_a depend on the exact definition of the encounter complex $O\cdots R$. It is natural to identify the rate k_d with the intrinsic dissociation rate k_{bR}, thus $k_d = k_{bR}$. From these expressions for k_+ and k_d and the requirement that the equilibrium constant should remain unchanged, one finds that $k_a/k_- = k_{fR}/k_D$. Using this result one obtains $k'_{fR} = k_D k_{fR}/(k_D + k_{fR})$ and $k'_{bR} = k_D k_{bR}/(k_D + k_{fR})$.

These renormalised rate constants have a clear interpretation. For the effective forward rate it follows, for instance, that: $1/k'_{fR} = 1/k_D + 1/k_{fR}$: that is, on average, the time required for repressor binding is given by the time needed to diffuse towards the operator plus the time for a reaction to occur when the repressor is in contact with the operator site [203]. The effective backward rate has a similar interpretation. The probability that after dissociation the repressor diffuses away from the operator site and never returns is given by $S_{\rm irr}(t \to \infty|\sigma)$, where $S_{\rm irr}(t, r_0)$ is the irreversible survival probability for two reacting particles [204]. Using that $S_{\rm irr}(t \to \infty|\sigma) = k_D/(k_{fR} + k_D)$, the expression for k'_{bR} can be written as $k'_{bR} = k_{bR} S_{\rm irr}(t \to \infty|\sigma)$: that is, the effective dissociation rate is the microscopic dissociation rate multiplied by the probability that after dissociation the repressor escapes from the operator site [203].

For diffusion limited reactions, such as the reaction considered here, we have that $k_{fR} \gg k_D$. Now, the renormalised rate constants reduce to:

$$k'_{fR} = k_D, \tag{6.16}$$

$$k'_{bR} = k_D k_{bR}/k_{fR}. \tag{6.17}$$

In Figure 6.2, we compare the noise profiles for the GFRD algorithm with those obtained by a simulation of the well-stirred model, where instead of the microscopic rates k_{fR} and k_{fB} we use the renormalised rates from Eqs. (6.16) and (6.17). Surprisingly, we find complete agreement. One of the main reasons why this is unexpected, is that for the master equation the time between events is Poisson-distributed, whereas after a dissociation the time to the next rebinding is distributed according to a power-law distribution when diffusion is taken into account [204]. The reason that this power-law behaviour of rebinding times does not influence the noise profile, is that the time scale of rapid rebinding is much shorter than any of the other relevant time scales in the network. Specifically, rebinding times are so short that the probability that an RNAP will bind before a rebinding event occurs is negligible. As a consequence, the transcription network is not at all influenced by the brief period the operator site is accessible before rebinding: for the transcription machinery the series of consecutive rebindings, albeit distributed algebraically in time individually, is perceived as a single event. And on much longer time scales, when a repressor diffuses in from the bulk towards the operator site, the distribution of arrival times is expected to be Poissonian, because on these time scales the repressors are distributed homogeneously in the bulk. This is succinctly summarised in Figure 6.4, which shows the distribution of association times. It is seen that at short time scales, the association events are algebraically distributed in time – these arise from the rapid rebindings – while at long time scales, they are distributed exponentially in time. For comparison, we also show the distribution of the repressor-DNA association times in the well-stirred model, with appropriately renormalised rate constants for repressor (un)binding (Eqs. (6.16) and (6.17)). As expected, the number of association events is much smaller at short time scales, but follows the same distribution as that of the spatially resolved model at long time scales. As described quantitatively in Section 6.7.2, the rate constant for the exponential relaxation is given not only by the diffusion-limited rate of repressor-DNA association, but also by

the RNAP promoter occupancy.

It is possible to reinterpret the effective rate constants in Eq. (6.16) and (6.17) in the language of rapid rebindings. The probability p that a rebind will occur after a dissociation from the DNA is given by $p = 1-S_\infty$, where $S_t = S_{\rm irr}(t, r_0 = \sigma)$. The probability that n consecutive rebindings occur before the repressor diffuses away from the operator site is then given by $p_n = (1-S_\infty)^n S_\infty$. From this follows that the average number of rebindings is $N_{RB} = (1-S_\infty)/S_\infty$. Using again that $S_\infty = k_D/(k_{fR}+k_D)$, we find that $N_{RB} = k_{fR}/k_D$. Combining this with Eqs. (6.16) and (6.17), we get:

$$k'_{fR} = k_{fR}/N_{RB}, \tag{6.18}$$

$$k'_{bR} = k_{bR}/N_{RB}. \tag{6.19}$$

In words, after an initial binding the repressor spends N_{RB} times longer on the DNA than expected on the basis of the microscopic backward rate, as it rebinds on average N_{RB} times. Because the average occupancy should not change, the forward rate should be renormalised in the same way. In conclusion, in this model the effects of diffusion can be properly described by a well-stirred model when the reaction rates are renormalised by the average number of rebindings.

6.7 Power Spectra

In this Section, we study how the noise due to the stochastic dynamics of the repressor molecules propagates through the different steps of gene expression for both the spatially resolved model and the well-stirred model, as it was sketched in Section 1.6.1. This analysis will also provide further insight into why the well-stirred model with renormalised rate constants for the (un)binding of the repressor molecules works so well.

In biochemical networks, the noise in the output signal depends upon the noise in the biochemical reactions that constitute the network, the so-called intrinsic noise, and on the noise in the input signal, called extrinsic noise [17, 19, 205, 178, 120, 206]. In our case, the output signal is the protein concentration, while the input signal is provided by the repressor concentration. The intrinsic noise arises from the biochemical reactions that constitute the transcription and translation steps. Moreover, we consider the noise in the protein concentration that is due to the (un)binding of the RNAP to (from) the DNA to be part of the intrinsic noise. The extrinsic noise is provided by the fluctuations in the binding of the repressor to the operator, *i.e.* in the state OR. Since the total repressor concentration, $[R_{\rm T}] = [R] + [OR]$, is constant, the extrinsic noise is also given by the fluctuations in the concentration of unbound repressor.

The noise properties of biochemical networks are most clearly elucidated via the power spectra of the time traces of the copy numbers of the components. Recently, we have shown that if the fluctuations in the input signal are uncorrelated with the noise in the biochemical reactions that constitute the processing network, the power spectrum of the output signal is given by [206]

$$S_{\rm P}(\omega) = S_{\rm int}(\omega) + g(\omega) S_{\rm ext}(\omega). \tag{6.20}$$

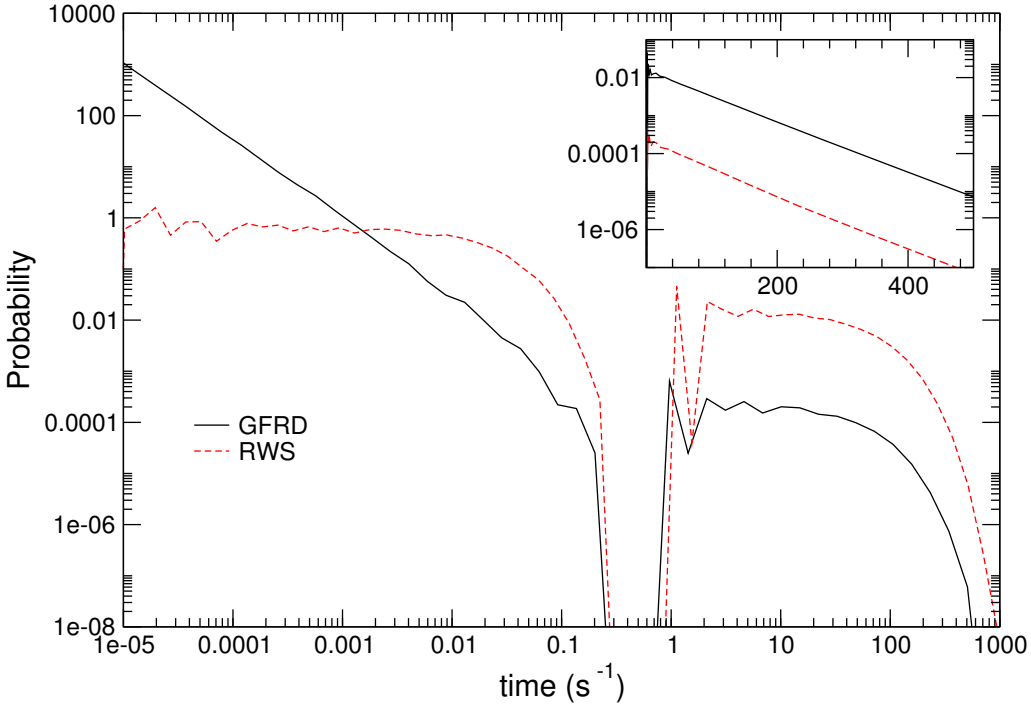


Figure 6.4: Distribution of repressor-DNA association times for the spatially resolved model (GFRD, solid line) and for the well-stirred, zero-dimensional model in which the rate constants for repressor-DNA (un)binding are given by the intrinsic (un)binding rates divided by the number of repressor rebindings in the absence of RNAP (see Eqs. (6.18) and (6.19)) – the renormalised well-stirred model (RWS, dashed line). It is seen that for the spatially resolved model, at short times the distribution follows a power-law, while at long times it is exponential. Moreover, at long times, the distribution of the spatially resolved model (GFRD) follows that of the renormalised well-stirred model (RWS). To a good approximation, the relaxation rate at long times in both the spatially resolved model and in the renormalised well-stirred model is given by $k'_{fR}[R_{\mathrm{T}}][O]'$, where k'_{fR} is the effective repressor association rate (see Eq. (6.18)), $[R_{\mathrm{T}}]$ is the total repressor concentration, and $[O]'$ is the probability that the promoter is not occupied by RNAP, given that it is not occupied by repressor (see Eq. (6.28)). The dip in the distribution at $t \approx 0.1-1.0\mathrm{s}^{-1}$ is due to the competition with the RNAP for binding to the promoter. The model parameters are $f = 100$, $k_{\mathrm{OC}} = 30\mathrm{s}^{-1}$, $N_R = 5$.

Here, $S_{\mathrm{P}}(\omega)$ is the power spectrum of the output signal, the protein concentration. The spectrum $S_{\mathrm{int}}(\omega)$ denotes the intrinsic noise of the processing network; it is defined as the noise in the output signal in the absence of noise in the input signal. Here, the intrinsic noise is due to the biochemical reactions of transcription and translation. The spectrum $S_{\mathrm{ext}}(\omega)$ is the power spectrum of the input signal, which, in this case, is given by the noise in the concentration of unbound repressor: $S_{\mathrm{ext}}(\omega) = S_{\mathrm{R}}(\omega)$; because the total repressor concentration is constant this power spectrum is also directly related to that of the repressor-bound state of the operator, $S_{\mathrm{OR}}(\omega)$. The function $g(\omega)$ is a transfer function, which indicates how fluctuations in the input signal are transmitted towards the output signal. If the extrinsic noise is uncorrelated with the intrinsic noise, then $g(\omega)$ is an intrinsic quantity that only depends upon properties of the processing network, and not upon properties of the incoming signal [206]. However, for the network studied here, the noise in the input signal is not uncorrelated with the intrinsic noise [206]. As we have shown recently, this means that Eq. (6.20) is not strictly valid [206]; the extrinsic contribution to the power spectrum of the output signal can no longer be factorised into a function that only depends upon intrinsic properties of the network, $g(\omega)$, and one that only depends upon the input signal, $S_{\mathrm{ext}}(\omega)$. This relation is nevertheless highly instructive. Indeed, Eq. (6.20) could be interpreted as a heuristic definition of the transfer function $g(\omega)$.

The diffusive motion of the repressor molecules impede an analytical evaluation of the power spectrum for the extrinsic noise. Moreover, while power spectra can be calculated analytically for linear reaction networks [34], the delays in transcription resulting from promoter clearance and elongation, preclude the derivation of an analytical expression for the power spectrum of the intrinsic noise. We have therefore obtained the power spectra $S_{\mathrm{P}}(\omega)$, $S_{\mathrm{ext}}(\omega)$, and $S_{\mathrm{int}}(\omega)$, directly from the time traces of the copy numbers. The power spectrum of a component X is given by $S_{\mathrm{X}}(\omega) = \langle |\tilde{X}(\omega)|^2 \rangle$, where $\tilde{X}(\omega)$ is the Fourier Transform of the concentration $X(t)$ of component X. Conventional FFT algorithms are not convenient for computing the power spectra, because our signals vary over a wide range of time scales. We therefore adopted a novel and efficient procedure, which is described in Appendix C. This procedure should prove useful for computing the power spectra of time traces of copy numbers of species in biochemical networks, as obtained by Kinetic Monte Carlo simulations.

As indicated above, the intrinsic noise, $S_{\mathrm{in}}(\omega)$, is defined as the noise in the output signal in the absence of fluctuations in the input signal. In order to determine the intrinsic contribution to the noise in the protein concentration, we discarded the (un)binding reaction of the repressor to the DNA (Eq. (6.3)), while rescaling the rate k_{bRp} for the dissociation reaction of the RNAP from the DNA (Eq. (6.4)) in such a way that the average concentration of the protein P remains unchanged. This eliminates the extrinsic noise arising from the repressor dynamics, thereby allowing us to obtain the intrinsic noise of the reactions in Eqs. (6.4-6.11). The rescaled backward rate k^*_{bRp} is given by

$$k^*_{\mathrm{bRp}} = k_{\mathrm{bRp}}(1 + K_2 N_R/V) + k_{\mathrm{OC}} K_2/V \tag{6.21}$$

where $K_2 = k_{\mathrm{fR}}/k_{\mathrm{bR}}$.

For the interpretation of the power spectra of the mRNA and protein concentration, as discussed below, it is instructive to recall the power spectrum of a linear birth-and-death process,

$$\emptyset \xrightarrow{k} A \xrightarrow{\mu} \emptyset, \tag{6.22}$$

with rate constants k and μ. For the interpretation of the spectra of repressor binding to the DNA, it is useful to recall the spectrum of a two-state model,

$$O \underset{k_2}{\overset{k_1}{\rightleftarrows}} O^*, \tag{6.23}$$

with rate constants k_1 and k_2. For both models, the power spectrum is a Lorentzian function of the form [31]:

$$S(\omega) = \frac{2\sigma^2\mu}{\mu^2+\omega^2}. \tag{6.24}$$

For the birth-and-death process, the variance in the concentration of A, σ^2, is k/μ, while for the two state system, the variance σ^2 in the occupancy n is $n(1-n)$; the decay rate in the two-state model is $\mu = k_1 + k_2$. The corner frequency μ (in both models) yields the time scale on which fluctuations relax back to steady state. We also note that the noise strength σ^2 is given by the integral of the power spectrum $S(\omega)$: $\sigma^2 = 1/(2\pi)\int_{-\infty}^{\infty} d\omega S(\omega)$. The noise strength is thus dominated by those frequencies at which the power spectrum is largest.

In the next Subsection, we discuss the effect of spatial fluctuations on the noise in gene expression and explain why a well-stirred model with renormalised rate constants for repressor (un)binding can capture its effect. In the subsequent Section, we discuss how the noise is propagated through the different stages of gene expression.

6.7.1 Spatial Fluctuations

In Figure 6.5, we show the power spectra for the input and output signals, for both the spatially resolved model and the well-stirred model with renormalised rate constants for repressor (un)binding (see previous Section). We recall that the output signal is the protein concentration, while the input signal is the concentration of unbound repressor (the extrinsic noise). Figure 6.5 also shows the power spectrum of the intrinsic noise. This is the noise in the protein concentration (the output signal), when the noise in the input signal resulting from the repressor dynamics has been eliminated by the procedure outlined above. The power spectra have been obtained in a parameter regime where the diffusing repressors have a large effect on the noise: $k_{\mathrm{OC}} = 30\mathrm{s}^{-1}$, $N_R = 5$ (see Figure 6.2).

Figure 6.5 shows that the power spectrum of the protein concentration in the spatially resolved model is identical to that in the well-stirred model for the entire range of frequencies observed. This confirms the observation in Section 6.6 that the effect of the spatial fluctuations of the repressor molecules on the noise in the protein concentration can by

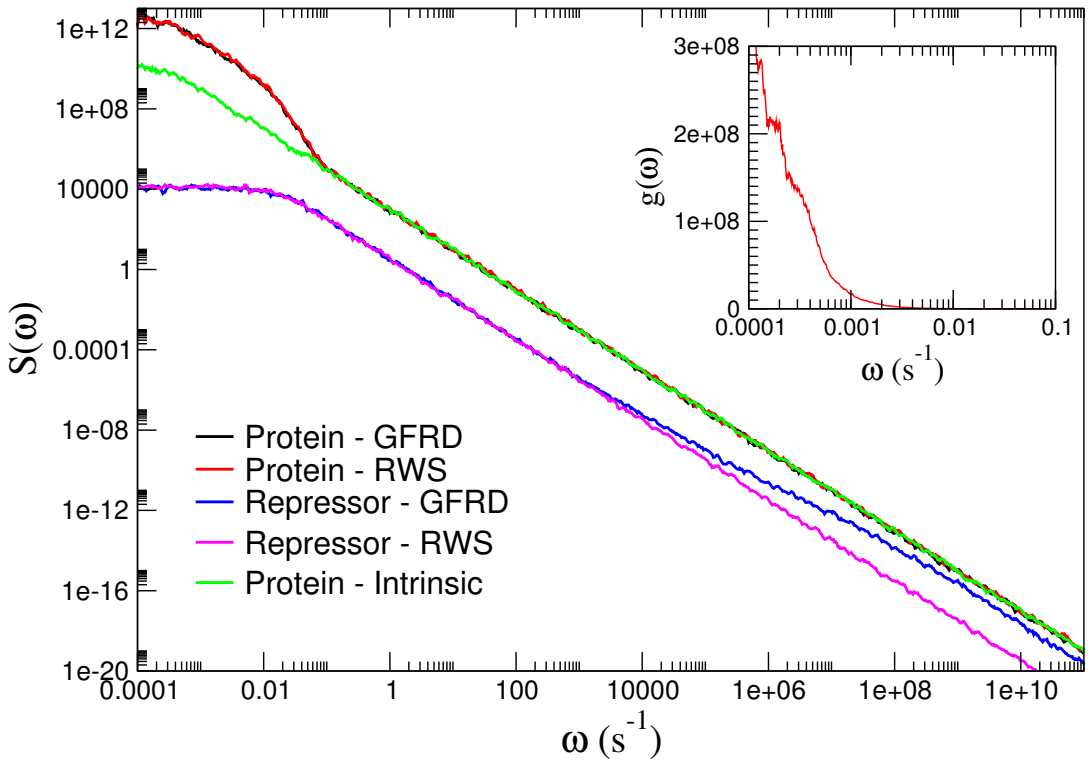


Figure 6.5: Power spectra of the repressor and protein concentrations for $f = 100$, $k_{OC} = 30\text{s}^{-1}$, $N_R = 5$. Data are shown both for the renormalised well-stirred model (RWS) with reaction rates renormalised according to Eqs. (6.16-6.17), and for GFRD, taking into account the spatial fluctuations of the repressor molecules explicitly. Also shown is the power spectrum of the intrinsic noise, which is the power spectrum of the protein concentration in the absence of fluctuations in the repressor concentration (extrinsic noise). For large ω, the repressor spectrum (extrinsic noise) differs between the well-stirred and the spatially resolved model. However, this difference does not appear in the power spectra of the protein concentration. The inset shows the frequency-dependent gain $g(\omega)$ (see Eq. (6.20)).

described by a well-stirred model in which the reaction rates for repressor (un)binding to the DNA are properly renormalised.

Figure 6.5 also elucidates the reason why a well-stirred model with properly renormalised rate constants for repressor (un)binding can successfully describe the effect of the diffusive motion of the repressor molecules on the noise in gene expression. It is seen that the repressor spectrum for the renormalised well-stirred model is accurately described by a Lorentzian function with a corner frequency $\mu = 0.02\text{s}^{-1}$ (see also Eq. (6.24)), as expected for the dynamics of repressor (un)binding dynamics (see next Section). The repressor spectrum of the spatially resolved model fully overlaps with that of the well-stirred model up to a frequency of $\omega \approx 10^6\text{s}^{-1}$, but for higher frequencies it shows a clear

deviation from the ω^{-2} behaviour. This deviation is caused by the diffusive motion of the repressor molecules. Indeed, the deviation occurs at frequencies comparable to the inverse of the typical time scale for rapid rebindings ($\sim \mu$s). However, this difference between the spectrum of the repressor dynamics in the spatially resolved model and that in the well-stirred model does not lead to a difference in the noise strength of the protein concentrations of the two respective models (see Figure 6.2), for two reasons: 1) the difference only occurs at high frequencies, *i.e.* in a frequency regime where the fluctuations only marginally contribute to the noise strength (the difference in area under the curves of the repressor power spectra for the two models is less than 5%); 2) the repressor fluctuations in this frequency range are filtered out by the processing network of transcription and translation; as a result of this, the effect of the small difference in area under the curves of the repressor power spectra for the two models is reduced even further. The filtering properties of the processing network are illustrated in the inset of Figure 6.5, which shows the transfer function $g(\omega)$ as obtained from $g(\omega) = (S_P(\omega) - S_{\rm int}(\omega))/S_{\rm ext}(\omega)$ (see Eq. (6.20)). Clearly, the transfer function rapidly decreases as the frequency increases. This shows that the processing network of transcription and translation acts as a low-pass filter, rejecting the high frequency noise in the repressor dynamics that originates from the rapid rebindings.

The only effect of the repressor rebindings on the noise in gene expression is thus that it lowers the effective dissociation rate (and association rate), as explained in the previous Section. As compared to the well-stirred model with the unrenormalised rate constants for repressor (un)binding, this decreases the corner frequency μ in the repressor power spectrum (see Figure 6.6), but *increases* the power at low frequencies – recall that for a two-state model, which relaxes mono-exponentially, the power spectrum at zero frequency is $S(\omega = 0) = 2\sigma^2/\mu$, which thus increases as the relaxation rate $\mu = k_1 + k_2$ decreases as a result of the slower binding and unbinding of repressor (see Eq. (6.24)). The higher power in the repressor spectrum at low frequencies for the spatially resolved model and for the well-stirred model with the renormalised rate constants, as compared to that for the well-stirred model with the unrenormalised rate constants, is not filtered by the processing network of transcription and translation and thus manifests itself in the power spectrum of the protein concentration. Spatial fluctuations of gene regulatory proteins thus increase the noise in gene expression by increasing the power of the input signal at low frequencies.

6.7.2 Noise propagation

In Figure 6.7 we show how fluctuations in the input signal arising from the dynamics of repressor binding and unbinding, are propagated through the different stages of gene expression. In Figure 6.7(a) we illustrate how the noise in the repressor concentration (the extrinsic noise) is transferred to the level of transcription. The Figure shows for both the spatially resolved model and for the well-stirred model with renormalised rate constants for repressor (un)binding, the power spectrum of the repressor concentration and the spectrum of the concentration of the elongation complex, defined as $[ORp^*] + [T]$.

It is clear from Figure 6.7(a) that already at the level of the elongation complex, the high-frequency noise due to the rapid rebindings is filtered. Transcription can thus already be described by a well-stirred model with properly renormalised rate constants for repressor (un)binding to (from) the DNA.

$$OR \underset{k_1}{\overset{k_2}{\rightleftarrows}} O \underset{k_4}{\overset{k_3}{\rightleftarrows}} ORp'. \tag{6.25}$$

The power spectrum of the elongation complex exhibits two corner frequencies, one around $\omega_+ \approx 40\text{s}^{-1}$ and another one at $\omega_- \approx 0.02\text{s}^{-1}$. These two corner frequencies arise from the competition between repressor and RNAP for binding to the promoter. To elucidate this, we have plotted in the inset the power spectrum for RNAP bound to the promoter, thus the power spectrum for $[ORp] + [ORp^*]$. It is seen that this power spectrum has the same two corner frequencies as that of the elongation complex, showing that their dynamics is dominated by the same processes – repressor binding and RNAP binding to the promoter. These two corner frequencies can be estimated analytically by considering the reactions in Eqs. (6.3-6.6) as a three-state system, in which repressor and RNAP compete for binding to the promoter: Here, $ORp' = ORp + ORp^*$, where ORp denotes the RNAP bound to the promoter in the closed complex and ORp^* denotes RNAP bound to the promoter in the open complex. The rate constant k_1 denotes the rate at which a repressor binds to the promoter; it is given by $k_1 = k'_{fR}[R_\text{T}]$, where k'_{fR} is the renormalised association rate (see Eq. (6.16)). The rate constant k_2 denotes the renormalised rate for repressor unbinding, $k_2 = k'_{bR}$ (see Eq. (6.17)); $k_3 = k_{fRp}$ denotes the rate at which RNAP binds to the promoter. The rate constant k_4 is the rate at which the RNAP leaves the promoter. Since the promoter can become accessible for the binding of another RNAP or repressor by either the dissociation of RNAP from the closed complex or by forming the open complex and then clearing the promoter, this rate is given by $k_4 = k_{bRp} + (k_\text{OC}^{-1} + t_\text{clear})^{-1}$. If promoter clearance would be neglected, then, indeed, $k_4 = k_{bRp} + k_\text{OC}$.

The power spectrum of the RNAP dynamics in Eq. (6.25) can be calculated analytically and is given by a sum of two Lorentzians:

$$S_{ORp'}(\omega) = \frac{A\omega_-}{\omega_-^2 + \omega^2} + \frac{B\omega_+}{\omega_+^2 + \omega^2}, \tag{6.26}$$

where A and B are coefficients. The corner frequencies ω_- and ω_+ are given by $\omega_\pm = (k \pm \sqrt{k^2 - 4h})/2$, where $k = \sum_i k_i$ and $h = k_1 k_4 + k_2(k_3 + k_4)$. The dynamics of repressor binding and unbinding is much slower than that of RNAP binding and unbinding, meaning that $k_1, k_2 \ll k_3, k_4$. This allows us to approximate the corner frequencies as $\omega_+ = k_3 + k_4$ and $\omega_- = k_2 + k_1 k_4/(k_3 + k_4)$. This yields the following expressions for the corner frequencies:

$$\omega_+ = k_{fRp} + k_{bRp} + (k_\text{OC}^{-1} + t_\text{clear})^{-1} \tag{6.27}$$

$$\omega_- = k'_{bR} + k'_{fR}[R_\text{T}][O]'. \tag{6.28}$$

Here, $[O]' \equiv k_4/(k_3 + k_4)$ is the conditional probability that the promoter is not occupied by the RNAP, given that it is not occupied by repressor; it is given by the occupancy of the

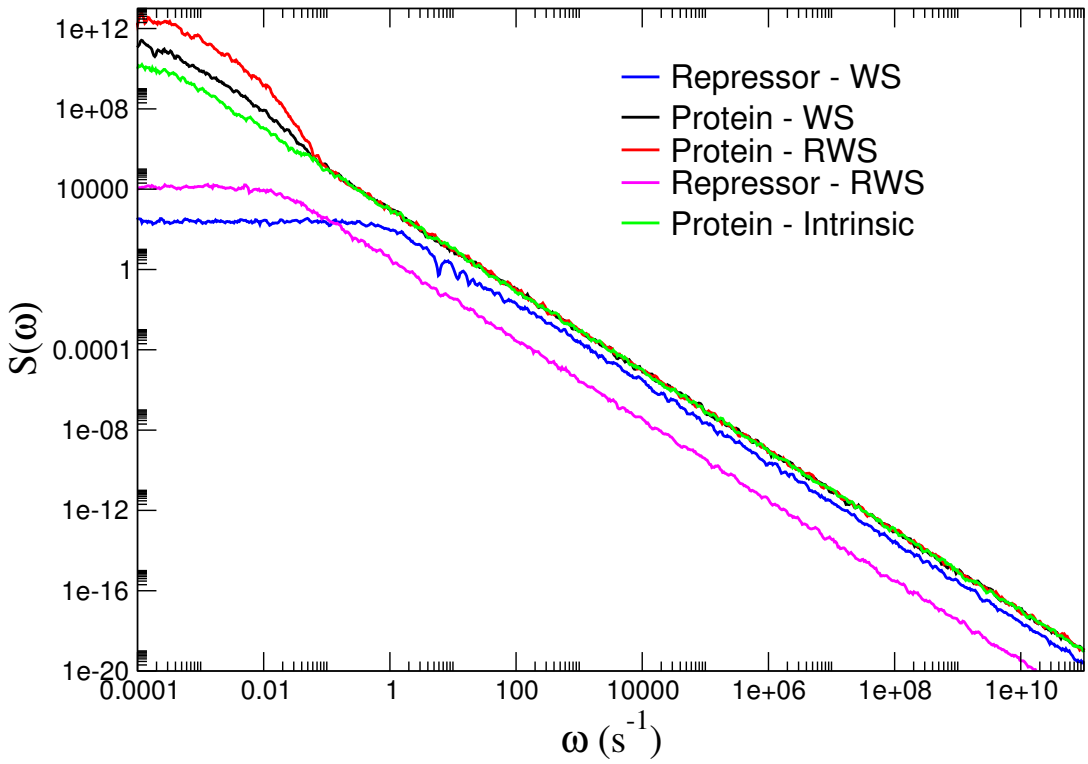


Figure 6.6: The power spectra for the well-stirred model with unrenormalised rate constants (WS) and for the well-stirred model with renormalised rate constants for repressor (un)binding to (from) the DNA (RWS). The intrinsic noise of gene expression is the same for both models. The extrinsic noise, arising from the repressor dynamics, is, however, markedly different. The repressor spectrum for the well-stirred model with renormalised rate constants has lower corner frequency, but, more importantly, also a higher power at low frequencies. The increased power at low frequencies is not filtered by the processing network and increases the noise in gene expression. For parameter values, see Figure 6.5.

promoter by RNAP in the *absence* of any repressor molecules in the system. We can now see that the highest corner frequency, ω_+, describes the *fast* dynamics of RNAP binding to, and clearing from, the promoter and that the other corner frequency, ω_-, represents the *slow* dynamics of repressor (un)binding to the DNA in the *presence* of the fast RNAP bindings to the promoter; the lower corner frequency, ω_-, is also the corner frequency in the repressor spectrum of the renormalised well-stirred model (see Figures 6.5 and 6.6). In Figure 6.7(a) we plot the power spectrum $S_{ORp'}(\omega)$ as predicted by the three-state model (Eq. (6.26); with fitted coefficients A and B) on top of the power spectrum obtained from the simulations and find excellent agreement. We also show the power spectra when we neglect the delay due to promoter clearance. As expected, in the absence of the delay due to promoter clearance, the lower corner frequency, ω_-, and, to a smaller extent, the higher corner frequency, ω_+, are shifted to higher frequencies.

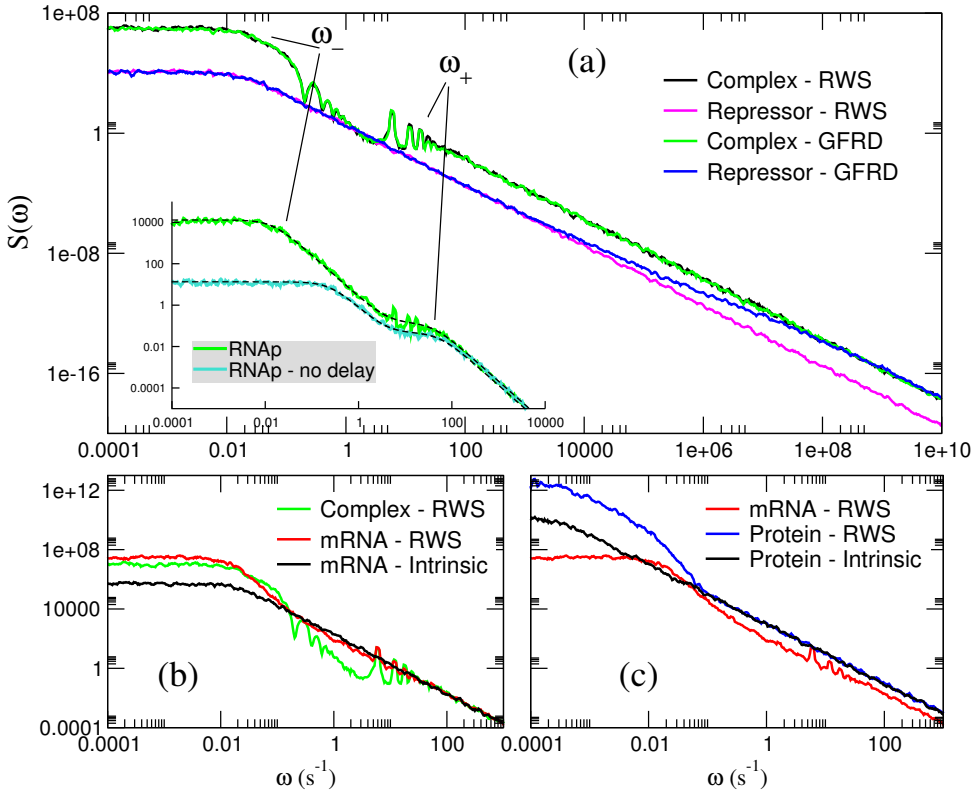


Figure 6.7: Comparison of the power spectra at different stages of gene expression. (a) Power spectrum for repressor concentration and for the elongation complex $ORp^* + T$, both for the well-stirred model with renormalised rate constants (RWS) and for GFRD. Repressor power spectra show a difference between the spatially resolved model and the well-stirred model at high frequencies, due to the diffusion of the repressor molecules. The power spectra for the elongation complexes coincide for the well-stirred and the spatially resolved model. The power spectrum of the elongation complex shows a series of peaks and valleys due to the presence of fixed delays in the dynamics of the elongation complex. (inset) Power spectrum of RNAP dynamics ($OR_p + OR_p^*$). Shown are the power spectra in the presence and absence of fixed delays in the RNAP dynamics. Due to the competition between RNAP and repressor for binding to the promoter, the power spectrum is described by a sum of two Lorentzians. (b) Power spectra of the elongation complex and mRNA. Peaks due to the delays in RNAP dynamics are still present in the mRNA dynamics. For high frequencies, the mRNA dynamics is well described by a linear birth-and-death process. (c) Spectra of mRNA and protein. The slow protein dynamics filters out all the peaks resulting from the delays in the RNAP dynamics. The only difference between the full spectrum of the output signal and that of the intrinsic noise is an increased noise at low frequencies, due to the repressor dynamics For parameter values, see Figure 6.5.

The power spectrum of the elongation complex in Figure 6.7(a) contains information that is not easily observed in the time domain and could as a result be helpful in the interpretation of the results. It is seen that there are two series of peaks. Those are associated with the two processes with fixed time delays. The first process is the promoter clearance, which takes a fixed time t_{clear}. Indeed, the first peak in the corresponding series of peaks in the power spectrum of the elongation complex, is at $\omega \approx 2\pi/(t_{\text{clear}}) = 6.3\text{s}^{-1}$; the other peaks in the series are the higher harmonics that naturally arise for processes with fixed time delays. The second process is the transcript elongation process. After the elongation complex has been formed, it takes a fixed time $t_{\text{clear}} + t_{\text{elon}}$ before the full transcript is formed and the RNAP dissociates from the DNA; the first valley of the corresponding series of peaks/valleys is, indeed, at $\omega \approx 2\pi/(t_{\text{clear}} + t_{\text{elon}}) = 0.2\text{s}^{-1}$. While the frequency $2\pi/t_{\text{clear}}$ yields, to a good approximation, the rate at which the elongation complex signal increases, the frequency $2\pi/(t_{\text{clear}} + t_{\text{elon}})$ corresponds to the frequency at which the elongation complex signal *decreases*; this explains why the shapes of the respective series of peaks and valleys are reciprocal. Lastly, the reason that both peaks and valleys are broadened is that the delay in the formation of the elongation complex is not fully deterministic: the duration of the delay is not only determined by the promoter clearance time, which, indeed, is fixed, but also by the time it takes for another RNAP to bind the DNA and then form the open complex – in the absence of repressor, the average frequency at which an elongation complex is formed is given by $2\pi/(k_{\text{fRp}}^{-1} + k_{\text{OC}}^{-1} + t_{\text{clear}})$ (see also Eqs. (6.4- 6.6)). Both RNAP binding and open complex formation are modeled as Poisson processes, and this leads to a *distribution* of delay times for the formation of the elongation complex.

For completeness, in Figure 6.7(b) and (c), we examine how the noise in the dynamics of the elongation complex propagates to the level of mRNA and protein dynamics. In Figure 6.7(b), we compare the full power spectrum of the mRNA concentration with that of the elongation complex – the input signal (extrinsic noise) for the mRNA signal – and that of the intrinsic noise of the mRNA signal; to compute the intrinsic noise, we have modeled the mRNA dynamics as a birth-and-death process (see Eq. (6.22)) with a production rate as given by the average production rate for the full system in Eqs. (6.3-6.11). As expected, for higher frequencies ($\omega > 0.1\text{s}^{-1}$), the full spectrum of mRNA overlaps almost fully with that of the intrinsic noise, although some traces of the input signal (the elongation complex) are still apparent in this high frequency regime; these are the peaks at $\omega \approx 6.3\text{s}^{-1}$ corresponding to promoter clearance. At lower frequencies ($\omega < 0.1\text{s}^{-1}$), the noise in the mRNA signal is dominated by the extrinsic noise, which is the noise in the elongation complex (the input signal). Indeed, both the spectrum of the elongation complex and that of mRNA have a corner frequency at ω_-, which, as discussed above, arises from the slow repressor (un)binding to the DNA in the presence of the fast DNA-(un)binding kinetics of RNAP.

Figure 6.7(c) shows how the noise in the mRNA concentration is propagated to that in the protein concentration. Again, at higher frequencies, the spectrum of the protein concentration coincides with that of the intrinsic noise of protein synthesis, which, as

above for mRNA, has been computed by modeling protein production as a birth-and-death process; note also that the remnants of operator clearance (the peaks in the spectrum at $\omega \approx 6.3\text{s}^{-1}$) have been filtered by the slow protein dynamics. Only for frequencies smaller than $\omega \approx 0.1\text{s}^{-1}$, does the extrinsic noise – the noise in the mRNA concentration – strongly contribute to the noise in the protein concentration. A careful inspection of the protein spectrum shows that it has a "corner" at ω_-, which arises from the repressor DNA-(un)binding dynamics (the extrinsic noise), and one, albeit much less visible, at $\omega \approx k_{\text{dp}} = 2 \times 10^{-4}\text{s}^{-1}$, which is due to the intrinsic dynamics of protein degradation.

6.8 Discussion and Outlook

Our analysis reveals that at high frequencies both mRNA and protein synthesis are well described by a linear birth-and-death model. In this frequency regime, the effect of spatial fluctuations, originating from the rapid repressor rebindings, is completely filtered by the slow dynamics of transcription and translation. These rebindings do, however, decrease the effective rate at which the repressor molecules associate with, and dissociate from, the promoter. This increases the intensity of the extrinsic (repressor) noise in the low frequency regime. Moreover, the low-frequency fluctuations in the repressor binding do propagate through the different stages of gene expression. In particular, they lead to sharp bursts in the production of mRNA and protein. These bursts increase the noise intensity at the lower frequencies in the noise spectrum of mRNA and protein. And since the noise strength σ^2 is dominated by fluctuations in the low-frequency regime, spatial fluctuations ultimately strongly increase the noise in mRNA and protein concentration.

Recently, experiments have been performed, in which the synthesis of individual mRNA transcripts [207] and individual protein molecules [208] could be detected. The systems in these studies were very similar to that studied here: a gene under the control of a (Lac) repressor. These studies unambiguously demonstrated that mRNA production [207] and protein synthesis can occur in bursts [208]. Of particular interest is the pulsatile transcription, which has been observed in experiments by Golding *et al.* [207] and in our simulations, but not in the experiments of Yu *et al.* [208]. We therefore address the question whether our analysis on transcription initiation in Section 6.7.2 can reconcile these observations. Transcription occurs in bursts if a) the operator is mostly in the repressed state, meaning that the repression strength f must be large; b) when the operator is in the derepressed state, more than one transcript is formed; this means that transcription initiation must be sufficiently fast as compared to repression-DNA association (see also Eq. (6.25)). In our simulations and in the Lac system studied by Yu *et al.* [208], the repression strength is indeed large, $f \approx 100$. With a typical *in vivo* repressor concentration of $[R_{\text{T}}] \approx 20\text{nM}$ ($N_{\text{R}} = 10$), the average repressor-DNA association rate, in the presence of RNAP, is $k'_{\text{fR}}[R_{\text{T}}][O]' \approx 0.1\text{s}^{-1}$ (see the reaction scheme in Eq. (6.25)). The rate of open complex formation of the lac promoter has been measured to be on the order of 0.1s^{-1} [187]. Hence, in the Lac system approximately one mRNA molecule is produced per gene expression event. This is consistent with the observations of Yu *et al.* [208, 209].

The observed burst-like *protein* production in these experiments is indeed due to the fact that more than one protein is formed from one mRNA transcript [208, 210, 21]. The repression strength, open complex formation rate, and repressor-DNA (un)binding rates for the system studied by Golding *et al.* are not known in similar detail [207], but, clearly, the observed pulsatile production of mRNA must mean that the repressor-DNA association rate is sufficiently slow as compared to the open complex formation rate.

The spatial fluctuations due to diffusion of the repressor molecules could have significant implications for the functioning of gene regulatory networks. Under some conditions, it might be crucial that the protein number is not only low on average, but remains low at all times. For instance, if the protein itself functions as a transcription factor, it might by accident induce the expression of another gene, when, due to a fluctuation, its concentration crosses a particular activation threshold. Thus, not all combinations of repressor copy number N_R and repressor backward rate k_{bR} that obey Eq. (6.14) and thus have the same average repression strength, are necessarily equivalent in terms of function when diffusion is taken into account. If the fluctuations in the repressed state need to be small, then the cell could increase the number of repressors and decrease the binding affinity to the operator site, such that the repressor molecules stay bound to the DNA only briefly. Alternatively, the cell could minimise the effect of fluctuations by reducing the rate at which the open complex is formed by RNAP – our analysis shows that the process of open complex formation can act as a strong low-pass filter.

The rapid rebindings observed in our simulations are a general phenomenon. However, the requirement for a specific orientation of the reactants could reduce the intrinsic association rate, and therefore limit the number of rebindings. In order to account for this effect, the GFRD scheme could be extended to explicitly include molecule orientations [41]. Alternatively, one could treat reorientations at a mean-field level, computing the the mean time the protein takes to find the right orientation and renormalise the association rate at contact k_a accordingly: we leave a quantitative study for future work. We now address the question of when the effect of spatial fluctuations due to diffusion can be described by a well-stirred model in which the association and dissociation rates are renormalised. In the current problem, the rebinding time for a dissociated repressor is exceedingly short. As a consequence, the probability that a RNAP binds to the promoter during this time, is vanishingly small. This is precisely the reason that the effective dissociation rate is simply the bare dissociation rate divided by the number of rebindings (see Eq. (6.18)); the effective association rate is renormalised accordingly, because the equilibrium constant should remain unchanged (see Eq. (6.19)). The success of the renormalised well-stirred model is thus a result of the strong separation of time scales – the time scale of repressor rebinding is well separated from that of RNAP binding. In fact, because of this strong separation of time scales, one could argue that the states, in which a repressor has just dissociated from the operator, should not be counted as unrepressed states, but rather as states that belong to the ensemble of microscopic states that together form the mesoscopic repressed state.

The separation of time scales also makes it possible to account for the effect of spatial

fluctuations by renormalising the association and dissociation rates in other cases. For instance, we have simulated a system in which repression occurs in a cooperative manner (data not shown). In this system, the repressor backward rate is smaller when two repressors are bound to the operator than when a single repressor is bound. However, when one of the two repressors dissociates, its rebinding time is so short that the probability for the other repressor to dissociate in the mean time, is negligible for reasonable values of cooperativity. As a result, the effect of spatial fluctuations can be described by a well-stirred model with properly renormalised reaction rates. We have also studied a system in which the expression of a gene is not under the control of a repressor, but rather under the control of an activator. In this system, too, diffusion of the transcription factors leads to an enhancement of noise in gene expression through a similar mechanism.

Do these observations imply that the effect of spatial fluctuations can always be described by a well-stirred model? In the system studied here, the ligand (repressor) molecules bind to a single site. We expect that the effect of spatial fluctuations becomes more intricate when the number of binding sites for a particular ligand increases – the binding of the ligand to the different sites closely located in the cell will then exhibit correlations. This could be important when the ligand binds to receptors that occur in dense clusters, as in bacterial chemotaxis [211, 212] and in the immune response [213]. In gene regulatory networks this effect could also be significant. Recently, we have shown that in *E. coli*, pairs of co-regulated genes – genes that are controlled by a common transcription factor – tend to lie exceedingly close to each other on the genome [77]: their promoter regions are often separated by a distance shorter than a few hundred base pairs. It is conceivable that spatial fluctuations of the transcription factors introduce correlations between the noise in the expression of these pairs of co-regulated genes. This study also revealed that pairs of genes that regulate each other, often lie close together, again suggesting that the diffusive motion of transcription factors could be important for the functioning of gene regulatory networks [77].

Even in the case of a single gene, the effect of spatial fluctuations is expected to be more complicated than that reported here. First and foremost, in this study we have assumed that the repressor, RNAP and ribosome molecules diffuse freely through the cytoplasm. This is likely to be a gross assumption. In fact, it has recently been observed in *Bacillus subtilis* that RNAP resides principally inside the nucleoid, while ribosomes are localised almost exclusively outside the nucleoid [214], suggesting that transcription and translation occur in separate spatial domains. Moreover, we have modeled the operator as a spherical site. However, as mentioned in Section 6.2.1, transcription factors are believed to find their operator site via a combination of free 3D diffusion and 1D sliding along the DNA. While on length scales longer than the sliding distance this process is indeed essentially 3D diffusion, on length and time scales shorter than the sliding distance and sliding time, respectively, the dynamics is more complicated. We expect that sliding could have two important effects. First, it will increase the *number* of rebindings – the probability that in 1D a random walker returns to the origin is one, while in 3D there is a finite probability that it will escape and never return. Secondly, sliding is expected to also increase

the *duration* of the rebindings, especially when diffusion along the DNA is much slower than diffusion in the cytoplasm. It is thus likely that with sliding, the non-exponential relaxation of the operator state, arising from the rebindings, shifts to lower frequencies (see Figure 6.5). Indeed, it is conceivable that with sliding, a dissociated repressor molecule can compete with RNAP for binding to the promoter. Under these conditions, the effect of spatial fluctuations might be detected experimentally in the statistics of the synthesis of the individual mRNA molecules, which could be useful for unraveling the mechanism and dynamics of transcription initiation. Importantly, we nevertheless expect that even under these conditions, the mRNA noise strength (variance) can be described by a zero-dimensional model, because the life time of the mRNA molecules, setting the time scale for time averaging, is probably still longer than the duration of the rebinding trajectories. However, the effective rate constant for repressor-DNA dissociation might no longer simply be given by the bare dissociation rate divided by the number of rebindings in the absence of RNAP. Indeed, it will depend upon the spatial fluctuations of the repressor molecules and their interplay with the RNAP-DNA association dynamics in a non-trivial manner, and deriving it would probably require a spatially resolved model. We leave this for future work.

Finally, we address the question whether spatial fluctuations, and, more in particular, the rebindings, could be studied experimentally. Interestingly, recent biochemical data on the restriction enzyme EcoRV suggests that after an initial dissociation, 10-100 rebindings occur before the enzyme escapes into the bulk solution [196, 197], in good agreement with the average number of rebindings calculated in Section 6.6. However, in our gene expression model, the rebinding times are so short that it would seem difficult to probe the repressor rebindings directly in an experiment. In fact, reaction rates measured biochemically will probably already be corrected for according to Eqs. (6.16) and (6.17). Sliding along the DNA, however, may extend the rebinding times to accessible experimental time scales. Moreover, recent experiments suggest that the motion of proteins in the nucleoid might be sub-diffusive, which would increase the importance of the rebindings [215]. Recently, magnetic tweezer experiments on a mechanically stretched, supercoiled, single DNA have made it possible to study the kinetics of the open complex formation and promoter clearance [187]. Performing these experiments *in vitro* on a promoter that is under the control of a repressor, seems a promising approach for studying the effect of spatial fluctuations due to the diffusive motion of transcription factors on the dynamics of gene expression.

Appendix **A**

Solving the dimerisation Master Equation

Following the approach described in [68], the Master Equation for the reaction $X + X \rightleftharpoons X_2$, with rate constants k_f for association and k_b for dissociation, with system volume V, and given a total number or monomers+dimers n_{X_T}, where $n_{X_T} = n_X + 2n_{X_2}$, is the following:

$$\begin{aligned}\frac{\partial}{\partial t}\,p(n_{X_2}|n_{X_T}) &= k_\mathrm{b}(n_{X_2}+1)\,p(n_{X_2}+1|n_{X_T}) - \\ &\left[\frac{k_\mathrm{f}(n_{X_T}-2n_{X_2})(n_{X_T}-2n_{X_2}-1)}{2V} + k_\mathrm{b}n_{X_2}\right]\,p(n_{X_2}|n_{X_T}) + \\ &\frac{k_\mathrm{f}(n_{X_T}-2n_{X_2}+2)(n_{X_T}-2n_{X_2}+1)}{2V}\,p(n_{X_2}-1|n_{X_T})\end{aligned} \tag{A.1}$$

Eq. (A.1) can be solved numerically in steady state (starting from an initial guess $n_{X_2} = 0$), to obtain the exact probability distribution for the number of dimers n_{X_2}, for a given total number n_{X_T} of monomers+dimers. The probability distribution for the monomer number can be trivially obtained from the dimer distribution noting that $p(n_X|n_{X_T}) = n_{X_T} - 2p(n_{X_2}|n_{X_T})$. Eq. (A.1) is solved for a range of values of n_{X_T}; results are stored in look-up tables, which are later used to compute effective propensities for the coarse-grained simulations.

Appendix **B**

Solving the operator binding Master Equation

In the EO approach, instead of solving the macroscopic rate equation for operator binding, one can solve the corresponding chemical Master Equation. However, as the operator states can be present only in copy number 0 or 1, the state space is extremely limited, and the solution of the Master Equation coincide with the solution for the rate equation.

The Master Equation for reactions (3.1b) is the following:

$$\begin{aligned}\frac{\partial}{\partial t}\,p(n_{A_2},n_{B_2}) &= \\ &- p(n_{A_2},n_{B_2})(k_{\rm on}n_O n_{\rm A_2}+k_{\rm on}n_O n_{\rm B_2}) \\ &- p(n_{A_2},n_{B_2})(k_{\rm off}n_{O\rm A_2}+k_{\rm off}n_{O\rm B_2} \\ &+ p(n_{A_2}-1,n_{B_2})k_{\rm off}(n_{O\rm A_2}+1) \\ &+ p(n_{A_2},n_{B_2}-1)k_{\rm off}(n_{O\rm B_2}+1) \\ &+ p(n_{A_2}+1,n_{B_2})k_{\rm on}(n_O+1)(n_{\rm A_2}-1) \\ &+ p(n_{A_2},n_{B_2}+1)k_{\rm on}(n_O+1)(n_{\rm B_2}-1).\end{aligned} \tag{B.1}$$

Only three states are possible: $(O{=}1,\mathrm{A}_2,\mathrm{B}_2)$, $(O\mathrm{A}_2{=}1,\mathrm{A}_2{-}1,\mathrm{B}_2)$ and $(O\mathrm{B}_2{=}1,\mathrm{A}_2,\mathrm{B}_2{-}1)$. This limited choice greatly simplifies Eq. (B.1):

$$\begin{aligned}p(n_{A_2},n_{B_2})k_{\rm on}(n_{\rm A_2}{+}n_{\rm A_2}) &= \\ p(n_{A_2}{-}1,n_{B_2})k_{\rm off}{+}p(n_{A_2},n_{B_2}{-}1)k_{\rm off}&, \\ p(n_{A_2},n_{B_2}{-}1) &= p(n_{A_2},n_{B_2})k_{\rm on}n_{\rm B_2}, \\ p(n_{A_2}{-}1,n_{B_2}) &= p(n_{A_2},n_{B_2})k_{\rm on}n_{\rm A_2}.\end{aligned} \tag{B.2}$$

The solutions of Eq. (B.2) can be easily computed:

$$\begin{aligned}
\langle n_O \rangle^{\mathrm{ME}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2} &= p(n_{A_2}, n_{B_2}) = \frac{1}{1+(K_D^b)^{-1}(n_{\mathrm{A}_2}+n_{\mathrm{B}_2})}, \\
\langle n_{O\mathrm{A}_2} \rangle^{\mathrm{ME}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2} &= p(n_{A_2}-1, n_{B_2}) = \frac{(K_D^b)^{-1} n_{\mathrm{A}_2}}{1+(K_D^b)^{-1}(n_{\mathrm{A}_2}+n_{\mathrm{B}_2})}, \\
\langle n_{O\mathrm{B}_2} \rangle^{\mathrm{RE}}_{\hat{\mathrm{A}}_2,\hat{\mathrm{B}}_2} &= p(n_{A_2}, n_{B_2}-1) = \frac{(K_D^b)^{-1} n_{\mathrm{B}_2}}{1+(K_D^b)^{-1}(n_{\mathrm{A}_2}+n_{\mathrm{B}_2})}.
\end{aligned} \tag{B.3}$$

Appendix C

Computing Power Spectra

The power spectrum of the time trace of the copy number $X(t)$ of a species X can be efficiently computed by exploiting the fact that in between the times t_k the signal $X(t)$ is constant. The Fourier Transform $S_{\mathrm{X}}(\omega)$ of $X(t)$ is

$$\tilde{X}(\omega) = \int X(t) e^{-i\omega t} dt = \sum_k \int_{t_{k-1}}^{t_k} X_k e^{-i\omega t} dt. \tag{C.1}$$

As $X(t)$ is constant within every interval $\{t_{k-1}, t_k\}$, the integration can easily be performed:

$$\tilde{X}(\omega) = \sum_k X_k \frac{1}{-i\omega} (e^{-i\omega t_k} - e^{-i\omega t_{k-1}}). \tag{C.2}$$

Shifting up by one the index j in the second part of the sum, we obtain:

$$\tilde{X}(\omega) = \frac{1}{i\omega} \sum_k (X_{k+1} - X_k)(e^{-i\omega t_k}). \tag{C.3}$$

The real and imaginary parts of the Fourier Transform are thus:

$$\Re[\tilde{X}(\omega)] = \frac{1}{\omega} \sum_k \delta_k (\sin \omega t_k) \tag{C.4}$$

$$\Im[\tilde{X}(\omega)] = \frac{1}{\omega} \sum_k \delta_k (\cos \omega t_k), \tag{C.5}$$

where we have defined $\delta_k = X_{k+1} - X_k$. The Power spectrum $S_{\mathrm{X}}(\omega) = \Re[\tilde{X}(\omega)]^2 + \Im[\tilde{X}(\omega)]^2$ is thus given by

$$S_{\mathrm{X}}(\omega) = \left(\frac{1}{\omega} \sum_k \delta_k \cos(\omega t_k) \right)^2 + \left(\frac{1}{\omega} \sum_k \delta_k \sin(\omega t_k) \right)^2. \tag{C.6}$$

The Fourier Transforms were computed at 10000 logarithmically spaced angular frequencies starting from $\omega_{\min} = 10 \cdot 2\pi/T$, where T is the total length of the signal. Power spectra obtained according to Eq. (C.6) were filtered with a box average over 20 neighboring points.

Appendix D

Derivation of $g(r)$

To compute the integral of equation (5.18):

$$g(r) = \frac{1}{(\pi\sigma^2)^{3/2}} \int_0^R r'^2 dr' \int_0^\pi \sin\theta d\theta \int_0^{2\pi} d\varphi \exp\left(-\frac{r'^2 - 2rr' + r^2}{\sigma^2}\right), \tag{D.1}$$

where $\sigma^2 = 4D\Delta t$.
Elementary methods can now be used: integration over the angular variables yields

$$g(r) = \frac{1}{\sqrt{\pi\sigma^2}} \frac{\exp\left(-r^2/\sigma^2\right)}{r} \int_0^R \left[\exp\left(-\frac{r'^2 - 2rr'}{\sigma^2}\right) - \exp\left(-\frac{r'^2 + 2rr'}{\sigma^2}\right)\right] r'\, dr'. \tag{D.2}$$

Finally, integrating over r' gives

$$\begin{aligned} g(r) \quad = \quad & \frac{\sigma}{\sqrt{\pi}} \frac{1}{2r} \left[\exp\left(-\frac{(r+R)^2}{\sigma^2}\right) - \exp\left(-\frac{(r-R)^2}{\sigma^2}\right)\right] + \\ & + \frac{1}{2}\left[\mathrm{erf}\left(\frac{-\mathrm{r}+\mathrm{R}}{\sigma}\right) - \mathrm{erf}\left(\frac{-\mathrm{r}}{\sigma}\right) + \mathrm{erf}\left(\frac{\mathrm{r}+\mathrm{R}}{\sigma}\right) - \mathrm{erf}\left(\frac{\mathrm{r}}{\sigma}\right)\right], \end{aligned} \tag{D.3}$$

where

$$\mathrm{erf(x)} = \frac{2}{\sqrt{\pi}} \int_0^{\mathrm{x}} \mathrm{e}^{\mathrm{t}^2/2} \mathrm{dt}.$$

Appendix **E**

Reaction set for the bacteriophage λ model

Here we list the reactions used in the basic model for the bacteriophage λ genetic switch in Chapter 4. O_{XYZ} represents the main operator, with molecule X bound to O_{R3}, molecule Y bound to O_{R2} and molecule Z bound to O_{R1}. {X,Y,Z}={0,C,R,Rp}, where 0 represents a free binding site, R a cI dimer, C a cro dimer, and Rp an RNA polymerase molecule. This last molecule binds either to O_{R3} or to O_{R1} and O_{R2} together. McI and Mcro stand, respectively, for cI and cro mRNAs.

$cI + cI \rightleftharpoons cI_2$
$O_{000} + cI_2 \rightleftharpoons O_{R00}$
$O_{000} + cI_2 \rightleftharpoons O_{00R}$
$O_{000} + cro_2 \rightleftharpoons O_{0C0}$
$O_{000} + Rp \rightleftharpoons O_{Rp00}$
$O_{R00} + cI_2 \rightleftharpoons O_{RR0}$
$O_{R00} + cro_2 \rightleftharpoons O_{RC0}$
$O_{R00} + Rp \rightleftharpoons O_{RRp}$
$O_{0R0} + cI_2 \rightleftharpoons O_{0RR}$
$O_{0R0} + cro_2 \rightleftharpoons O_{0RC}$
$O_{00R} + cI_2 \rightleftharpoons O_{ROR}$
$O_{00R} + cro_2 \rightleftharpoons O_{C0R}$
$O_{00R} + Rp \rightleftharpoons O_{Rp0R}$
$O_{C00} + cI_2 \rightleftharpoons O_{C0R}$
$O_{C00} + cro_2 \rightleftharpoons O_{C0C}$
$O_{0C0} + cI_2 \rightleftharpoons O_{RC0}$
$O_{0C0} + cro_2 \rightleftharpoons O_{CC0}$
$O_{0C0} + Rp \rightleftharpoons O_{RpC0}$
$cro + cro \rightleftharpoons cro_2$
$O_{000} + cI_2 \rightleftharpoons O_{0R0}$
$O_{000} + cro_2 \rightleftharpoons O_{C00}$
$O_{000} + cro_2 \rightleftharpoons O_{00C}$
$O_{000} + Rp \rightleftharpoons O_{0Rp}$
$O_{R00} + cI_2 \rightleftharpoons O_{ROR}$
$O_{R00} + cro_2 \rightleftharpoons O_{R0C}$
$O_{0R0} + cI_2 \rightleftharpoons O_{RR0}$
$O_{0R0} + cro_2 \rightleftharpoons O_{CR0}$
$O_{0R0} + Rp \rightleftharpoons O_{RpR0}$
$O_{00R} + cI_2 \rightleftharpoons O_{0RR}$
$O_{00R} + cro_2 \rightleftharpoons O_{OCR}$
$O_{C00} + cI_2 \rightleftharpoons O_{CR0}$
$O_{C00} + cro_2 \rightleftharpoons O_{CC0}$
$O_{C00} + Rp \rightleftharpoons O_{CRp}$
$O_{0C0} + cI_2 \rightleftharpoons O_{0CR}$
$O_{0C0} + cro_2 \rightleftharpoons O_{0CC}$
$O_{00C} + cI_2 \rightleftharpoons O_{R0C}$

$O_{00C} + cI_2 \rightleftharpoons O_{0RC}$
$O_{00C} + cro_2 \rightleftharpoons O_{0CC}$
$O_{Rp00} + cI_2 \rightleftharpoons O_{RpR0}$
$O_{Rp00} + cro_2 \rightleftharpoons O_{RpC0}$
$O_{Rp00} + Rp \rightleftharpoons O_{RpRp}$
$O_{0Rp} + cro_2 \rightleftharpoons O_{CRp}$
$O_{RR0} + cI_2 \rightleftharpoons O_{RRR}$
$O_{ROR} + cI_2 \rightleftharpoons O_{RRR}$
$O_{0RR} + cI_2 \rightleftharpoons O_{RRR}$
$O_{0RR} + Rp \rightleftharpoons O_{RpRR}$
$O_{RC0} + cro_2 \rightleftharpoons O_{RCC}$
$O_{R0C} + cro_2 \rightleftharpoons O_{RCC}$
$O_{CR0} + cro_2 \rightleftharpoons O_{CRC}$
$O_{C0R} + cro_2 \rightleftharpoons O_{CCR}$
$O_{0CR} + cro_2 \rightleftharpoons O_{CCR}$
$O_{0RC} + cro_2 \rightleftharpoons O_{CRC}$
$O_{CC0} + cI_2 \rightleftharpoons O_{CCR}$
$O_{C0C} + cI_2 \rightleftharpoons O_{CRC}$
$O_{0CC} + cI_2 \rightleftharpoons O_{RCC}$
$O_{0CC} + Rp \rightleftharpoons O_{RpCC}$
$O_{RpR0} + cro_2 \rightleftharpoons O_{RpRC}$
$O_{Rp0R} + cro_2 \rightleftharpoons O_{RpCR}$
$O_{RpC0} + cro_2 \rightleftharpoons O_{RpCC}$
$O_{Rp0C} + cro_2 \rightleftharpoons O_{RpCC}$
$O_{00C} + cro_2 \rightleftharpoons O_{C0C}$
$O_{00C} + Rp \rightleftharpoons O_{Rp0C}$
$O_{Rp00} + cI_2 \rightleftharpoons O_{Rp0R}$
$O_{Rp00} + cro_2 \rightleftharpoons O_{Rp0C}$
$O_{0Rp} + cI_2 \rightleftharpoons O_{RRp}$
$O_{0Rp} + Rp \rightleftharpoons O_{RpRp}$
$O_{RR0} + cro_2 \rightleftharpoons O_{RRC}$
$O_{ROR} + cro_2 \rightleftharpoons O_{RCR}$
$O_{0RR} + cro_2 \rightleftharpoons O_{CRR}$
$O_{RC0} + cI_2 \rightleftharpoons O_{RCR}$
$O_{R0C} + cI_2 \rightleftharpoons O_{RRC}$
$O_{CR0} + cI_2 \rightleftharpoons O_{CRR}$
$O_{C0R} + cI_2 \rightleftharpoons O_{CRR}$
$O_{0CR} + cI_2 \rightleftharpoons O_{RCR}$
$O_{0RC} + cI_2 \rightleftharpoons O_{RRC}$
$O_{0RC} + Rp \rightleftharpoons O_{RpRC}$
$O_{CC0} + cro_2 \rightleftharpoons O_{CCC}$
$O_{C0C} + cro_2 \rightleftharpoons O_{CCC}$
$O_{0CC} + cro_2 \rightleftharpoons O_{CCC}$
$O_{RpR0} + cI_2 \rightleftharpoons O_{RpRR}$
$O_{Rp0R} + cI_2 \rightleftharpoons O_{RpRR}$
$O_{RpC0} + cI_2 \rightleftharpoons O_{RpCR}$
$O_{Rp0C} + cI_2 \rightleftharpoons O_{RpRC}$

$O_{Rp00} \rightarrow O_{000} + Rp + McI$
$O_{Rp0R} \rightarrow O_{00R} + Rp + McI$
$O_{Rp0C} \rightarrow O_{00C} + Rp + McI$
$O_{RpCR} \rightarrow O_{0CR} + Rp + McI$
$O_{RpCC} \rightarrow O_{0CC} + Rp + McI$
$O_{0Rp} \rightarrow O_{000} + Rp + Mcro$
$O_{CRp} \rightarrow O_{C00} + Rp + Mcro$
$McI \rightarrow McI + cI$
$McI \rightarrow 0$
$cI \rightarrow 0$
$cI_2 \rightarrow 0$
$O_{RpR0} \rightarrow O_{0R0} + Rp + McI$
$O_{RpC0} \rightarrow O_{0C0} + Rp + McI$
$O_{RpRR} \rightarrow O_{0RR} + Rp + McI$
$O_{RpRC} \rightarrow O_{0RC} + Rp + McI$
$O_{RpRp} \rightarrow O_{0Rp} + Rp + McI$
$O_{RRp} \rightarrow O_{R00} + Rp + Mcro$
$O_{RpRp} \rightarrow O_{Rp00} + Rp + Mcro$
$Mcro \rightarrow Mcro + cro$
$Mcro \rightarrow 0$
$cro \rightarrow 0$
$cro_2 \rightarrow 0$

Bibliography

[1] Alberts, B. et al., *Molecular Biology of the Cell*, Gardland Publishing, New York, 1994.

[2] Barabási, A. L. and Oltvai, Z. N., Nat. Rev. Gen. **5** (2004) 101.

[3] Dobrin, R., Beg, Q. K., Barabási, A. L., and Oltvai, Z. N., BMC Bioinformatics **5** (2004) 10.

[4] Watts, D. J. and Strogatz, S. H., Nature **393** (1998) 440.

[5] Chung, F. and Lu, L., Proc. Natl. Acad. Sci. USA **99** (2002) 15879.

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

[7] Milo, R. et al., Science **298** (2002) 824.

[8] Shen-Orr, S., Milo, R., Mangan, S., and Alon, U., Nat. Gen. **34** (2002) 64.

[9] Camas, F. M., Blazquez, J., and Poyatos, J. F., Proc. Natl. Acad. Sci. USA **103** (2006) 12718.

[10] Alon, U., Nat. Rev. Gen. **8** (2007) 450.

[11] Mangan, S., Zaslaver, A., and Alon, U., J. Mol. Biol. **334** (2003) 197.

[12] Kalir, S., Mangan, S., and Alon, U., Molecular Systems Biology **1** (2005) 2005.0006.

[13] Basu, S., Mehreja, R., Thiberge, S., Chen, M. T., and Weiss, R., Proc. Natl. Acad. Sci. USA **101** (2004) 6355.

[14] Zaslaver, A. et al., Nat. Gen. **36** (2004) 486.

[15] Hengge-Aronis, R., Cell **72** (1993) 165.

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

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

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

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

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

[21] Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and van Oudenaarden, A., Nat. Gen. **31** (2002) 69 .

[22] Blake, W. J., Kaern, M., Cantor, C. R., and Collins, J. J., Nature **422** (2003) 633.

[23] Barkai, N. and Leibler, S., Nature **387** (1997) 913.

[24] Alon, U., Surette, M. G., Barkai, N., and Leibler, S., **397** (1999) 168.

[25] Dutton, K., Thompson, S., and Barraclough, B., *The Art of Control Engineering*, Addison-Wesley, Harlaw, 1997.

[26] Becksei, A. and Serrano, L., Nature **405** (2000) 590.

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

[28] Ahmad, K. and Heinkoff, S., Cell **104** (2001) 839.

[29] Thattai, M. and van Oudenaarden, A., Genetics **167** (2004) 523 .

[30] Austin, D. W. et al., Nature **439** (2006) 608.

[31] van Kampen, N. G., *Stochastic Processes in Physics and Chemistry, 2nd edition*, Elsevier, Amsterdam, 2001.

[32] Shnerb, N. M., Louzoun, Y., Bettelheim, E., and Solomon, S., Proc. Natl. Acad. Sci. USA **97** (2000) 10322.

[33] Togashi, Y. and Kaneko, K., Phys. Rev. Lett. **86** (2001) 2459.

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

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

[36] Bortz, A. B., Kalos, M. H., and Lebowitz, J. L., J. Comp. Phys. **17** (1975) 10.

[37] Gibson, M. A. and Bruck, J., J. Phys. Chem. A **104** (2000) 1876.

[38] Elf, J. and Ehrenberg, M., Sys. Biol. **2** (2004) 230.

[39] Ander, M. et al., Syst. Biol. **1** (2004) 129.

[40] van Zon, J. S. and ten Wolde, P. R., Phys. Rev. Lett. **94** (2005) 018104.

[41] van Zon, J. S. and ten Wolde, P. R., J. Chem. Phys. **123** (2005) 234910.

[42] Allen, R. J., Warren, P. B., and ten Wolde, P. R., Phys. Rev. Lett. **94** (2005) 128103.

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

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

[45] Ptashne, M. and Gann, A., *Genes and signals*, Cold Spring Harbor Laboratory Press, New York, 2002.

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

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

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

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

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

[51] Valeriani, C., Allen, R. J., Morelli, M. J., Frenkel, D., and ten Wolde, P. R., (2007), In press on J. Chem. Phys.

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

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

[54] Zhu, X. M., Yin, L., Hood, L., and Ao, P., **2** (2004) 785.

[55] Warren, P. B. and ten Wolde, P. R., Phys. Rev. Lett. **92** (2004) 128101.

[56] Lipshtat, A., Loinger, A., Balaban, N. Q., and Biham, O., Phys. Rev. Lett. **96** (2006) 188101.

[57] Ushikubo, T., Inoue, W., Yoda, M., and Sasai, M., Chem. Phys. Lett. **430** (2006) 139.

[58] Loinger, A., Lipshtat, A., Balaban, N. Q., and Biham, O., Phys. Rev. E **75** (2007) 021904.

[59] Bolhuis, P. G., Dellago, C., and Chandler, D., Proc. Natl. Acad. Sci. USA **97** (2000) 5877.

[60] Geissler, P. G., Dellago, C., and Chandler, D., J. Phys. Chem. B **103** (1999) 3706.

[61] Cherry, J. L. and Adler, F. R., J. theor. Biol. **203** (2000) 117.

[62] Faradjian, A. K. and Elber, R., J. Chem. Phys. **102** (2004) 10880.

[63] Moroni, D., Bolhuis, P. G., and van Erp, T. S., J. Chem. Phys. **120** (2004) 4055.

[64] Buchler, N. E., Gerland, U., and Hwa, T., Proc. Natl. Acad. Sci. USA **102** (2005) 559.

[65] Record, M. T., Reznikoff, W., Craig, M., McQuade, K., and Schlax, P., Escherichia coli rnapolymerase (es70), promoters and the kinetics of the steps of transcription initiation, in *Escherichia Coli and Salmonella, 2nd ed.*, edited by Neidhardt, F. C. and *et al.*, American Society for Microbiology Press, Washington, 1996.

[66] Gillespie, D. T., J. Comp. Phys. **22** (1976) 403.

[67] Haseltine, E. L. and Rawlings, J. B., J. Chem. Phys. **117** (2002) 6959.

[68] Bundschuh, R., Hayot, F., and Jayaprakash, C., Biophys. J. **84** (2003) 1606.

[69] Rao, C. V. and Arkin, A. P., J. Chem. Phys. **118** (2003) 4999.

[70] Puchałka, J. and Kierzek, A. M., Biophys. J. **86** (2004) 1357.

[71] Cao, Y., Gillespie, D. T., and Petzold, L. R., J. Chem. Phys. **122** (2005) 014116.

[72] Weinan, E., Liu, D., and Vanden-Eijnden, E., JCP **123** (2005) 194107.

[73] Salis, H. and Kaznessis, Y. N., J. Chem. Phys. **123** (2005) 214106.

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

[75] Kiehl, T. R., Mattheyses, R. M., and Simmons, M. K., Bioinformatics **20** (2004) 316 .

[76] Takahashi, K., Kaizu, K., Hu, B., and Tomita, M., Bioinformatics **20** (2004) 538 .

[77] Warren, P. B. and ten Wolde, P. R., J. Mol. Biol. **342** (2004) 1379.

[78] Ptashne, M., *A Genetic Switch: Gene Control and* λ *phage*, Cell Press and Blackwell Scientific Publications, Cambridge, MA, 1986.

[79] Darling, P. J., Holt, J. M., and Ackers, G. K., J. Mol. Biol. **302** (2000) 625.

[80] Hawley, D. K. and McClure, W. R., Cell **32** (1983) 327.

[81] Darling, P. J., Holt, J. M., and Ackers, G. K., Biochemistry **39** (2000) 11500.

[82] Burz, D. S., Beckett, D., Benson, N., and Ackers, G. K., Biochemistry **33** (1994) 8399.

[83] Press, W. H., Teukolsky, S., Vetterling, W. T., and Flannery, B. P., *Numerical Recipes in C, 2nd ed.*, Cambridge University Press, Oakleigh, 1992.

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

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

[86] Santillan, M. and Mackey, M. C., Biophys. Jour. **86** (2004) 75.

[87] Reinitz, J. and Vaisnys, J. R., J. Theor. Biol. **145** (1990) 295.

[88] Beckett, D., Koblan, K. S., and Ackers, G. K., Anal. Biochem. **196** (1991) 69.

[89] Shea, M. A. and Ackers, G. K., J. Mol. Biol. **181** (1985) 211.

[90] Dodd, I. B., Perkins, A. J., Tsemitsidis, S., and Egan, J. B., Genes and Development **15** (2001) 3013.

[91] Baek, K., Svenningsen, S., Eisen, H., Sneppen, K., and Brown, S., J. Mol. Biol. **334** (2003) 363.

[92] Dodd, I. B., Shearwin, K. E., Perkins, A. J., Burr, T., and Egan, J. B., Genes and Development **18** (2004) 344.

[93] Elowitz, M. B., Surette, M. G., Wolf, P.-E., Stock, J. B., and Leibler, S., J. Bacteriol. **181** (1999) 197.

[94] Goodsell, D. S. and Olson, A. J., Trends in Biochem. Sci. **18** (1993) 65.

[95] Atkins, P. and de Paula, J., *Atkins' Physical Chemistry, 7th ed.*, Oxford Unversity Press, New York, 2002.

[96] Folkmanis, A., Takeda, Y., Simuth, J., Gussin, G., and Echols, H., Proc. Natl. Acad. Sci. USA **73** (1976) 2249.

[97] Sauer, R. T., Pabo, C. O., Meyer, B. J., and Ptashne, M., Nature **279** (1979) 396.

[98] Jia, H., Satumba, J., Bidwell, G. L., and Mossing, M. C., J. Mol. Biol. **350** (2005) 919.

[99] McClure, W. R., Ann. Rev. Biochem. **54** (1985) 171.

[100] Hawley, D. K. and McClure, W. R., J. Mol. Biol. **157** (1982) 493.

[101] Hwang, J. J., Brown, S., and Gussin, G., J. Mol. Biol. **200** (1988) 695.

[102] Li, M., McClure, W. R., and Susskind, M. M., Proc. Natl. Acad. Sci. USA **94** (1997) 3691.

[103] Bernstein, J. A., Khodursky, A. B., Lin, P. H., Lin-Chao, S., and Cohen, S. N., Proc. Natl. Acad. Sci. USA **99** (2002) 9697.

[104] Shean, C. S. and Gottesman, M. E., Cell **70** (1992) 513.

[105] Ringquist, S. and *et al.*, Mol. Microbiol. **6** (1992) 1219.

[106] Pakula, A. A., Young, V. B., and Sauer, R. T., Proc. Natl. Acad. Sci. USA **83** (1986) 8829.

[107] Bremer, H. and Dennis, P. P., Modulation of chemical composition and other parameters of the cell by growth rate, in *Escherichia Coli and Salmonella thyphymurium: Cellular and Molecular Biology, Vol. 2*, edited by et al., F. C. N., American Society for Microbiology Press, Washington, 1996.

[108] Calef, E. et al., in *Introduction to Lambda*, edited by Hershey, A. D. and Dove, W., Cold Spring Laboratory, New York, 1971.

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

[110] Torrie, G. M. and Valleau, J. P., Chem. Phys. Lett. **28** (1974) 578.

[111] van Duijneveld, J. S. and Frenkel, D., J. Chem. Phys. **96** (1992) 4655.

[112] ten Wolde, P. R., Ruiz-Montero, M. J., and Frenkel, D., J. Chem. Phys. **104** (1996) 9932.

[113] ten Wolde, P. R., Ruiz-Montero, M. J., and Frenkel, D., Faraday Discuss. **104** (1996) 93.

[114] Bennet, C. H., in *Algorithms for Chemical Computations, ACS Symposium, Series No.46*, edited by R.Christofferson, Washington, D.C., 1977, Americal Chemical Society.

[115] Chandler, D., J. Chem. Phys. **68** (1978) 2959.

[116] Dellago, C., Bolhuis, P. G., Csajka, F., and Chandler, D., J. Chem. Phys. **108** (1998) 1964.

[117] Dellago, C., Bolhuis, P. G., and Geissler, P. L., Adv. Chem. Phys. **123** (2002) 1.

[118] van Erp, T. S., Moroni, D., and Bolhuis, P. G., J. Chem. Phys. **118** (2003) 7762.

[119] van Erp, T. S. and Bolhuis, P. G., J. Comp. Phys. **205** (2005) 157.

[120] Shibata, T. and Fujimoto, K., Proc. Natl. Acad. Sci. USA **102** (2005) 331 .

[121] Svenningsen, S. L., Costantino, N., Court, D. L., and Adhya, S., Proc. Natl. Acad. Sci. USA **102** (2005) 4465.

[122] Neubauer, Z. and Calef, E., J. Mol. Biol. **51** (1970) 1.

[123] Eisen, H., Brachet, P., Pereira da Silva, L., and Jacob, F., Proc. Natl. Acad. Sci. USA **66** (1970) 855.

[124] Ellis, R. J., TRENDS in Biochem. Sci. **26** (2001) 597.

[125] Minton, A. P., J. Biol. Chem. **276** (2001) 10577.

[126] Minton, A. P., J. Pharm. Sci. **94** (2005) 1668.

[127] Jarvis, T. C., Ring, D. M., Daube, S. S., and von Hippel, P. H., J. Biol. Chem. **25** (1990) 15160.

[128] Muramatsu, N. and Minton, A. P., Proc. Natl. Acad. Sci. USA **85** (1988) 2984.

[129] Konopka, M. C., Shkel, I. A., Cayley, S., Record, M. T., and Weisshaar, J. C., J. Bacteriol. **188** (2006) 6115.

[130] van den Bogaart, G., Hermans, N., Krasnikov, V., and Poolman, B., Mol. Microbiol. **64** (2007) 858.

[131] Reits, E. A. J. and Neefjes, J. J., Nat. Cell. Biol. **3** (2001) E145.

[132] Révet, B., von Wilcken-Bergmann, B., Bessert, H., Barker, A., and Muller-Hill, B., Curr. Biol. **9** (1999) 151.

[133] Senear, D. F. et al., Biochemistry **32** (1993) 6179.

[134] Meiners, J. C. and Quake, S. R., Phys. Rev. Lett. **84** (2000) 5014.

[135] Dekel, E. and Alon, U., Nature **436** (2005) 588 .

[136] Atsumi, S. and Little, J. W., Proc. Natl. Acad. Sci. USA **103** (2006) 4558.

[137] Gottesman, M., J. Mol. Biol. **293** (1999) 177.

[138] Dodd, I. B., Shearwin, K. E., and Egan, J. B., Curr. Op. in Gen. and Dev. **15** (2005) 145.

[139] Court, D. L., Oppenheim, A. B., and Adhya, S. L., J. Bacteriol. **189** (2007) 298.

[140] Bakk, A. and Metzler, R., J. Theor. Biol. **231** (2004) 525.

[141] Bakk, A. and Metzler, R., FEBS Letters **563** (2004) 66.

[142] Dorman, C. J. and Deighan, P., Curr. Op. in Gen. and Dev. **13** (2003) 179.

[143] Rimsky, S., Curr. Op. in Microb. **7** (2004) 109.

[144] Cunha, S., Woldringh, C. L., and Odijk, T., J. Struct. Biol. **150** (2001) 53.

[145] Odijk, T., Biophys. Chem. **73** (1998) 23.

[146] Vilar, J. M. G. and Leibler, S., J. Mol. Biol. **331** (2003) 981.

[147] Vilar, J. M. G. and Leibler, S., Curr. Op. in Gen. and Dev. **15** (2005) 136.

[148] Ermak, D. L. and McCammon, J. A., J. Chem. Phys. **69** (1978) 1352.

[149] Wade, R. C., Biochem. Soc. Trans. **24** (1996) 254.

[150] Huber, G. A. and Kim, S., Biophys. J. **70** (1996) 97.

[151] Gabdoulline, R. R. and Wade, R. C., Biophys. J. **72** (1997) 1917.

[152] Zou, G., Skeel, R. D., and Subramaniam, S., Biophys. J. **79** (2000) 638.

[153] Gabdoulline, R. R. and Wade, R. C., J. Mol. Biol. **306** (2001) 1139.

[154] Gabdoulline, R. R. and Wade, R. C., Curr. Opin. Struct. Biol. **12** (2002) 204.

[155] Northrup, S. H. and Erickson, H. P., Proc. Natl. Acad. Sci. USA **89** (1992) 3338.

[156] Elcock, A. H., Proc. Natl. Acad. Sci. USA **100** (2003) 2340.

[157] Elcock, A. H., Biophys. J. **82** (2002) 2326.

[158] Schaff, J., Fink, C. C., Slepchenko, B., Carson, J. H., and Loew, L. M., Biophys. J. **73** (1997) 1135.

[159] Andrews, S. S. and Bray, D., Physical Biology **1** (2004) 137.

[160] Stiles, J. R., in *Computational Neuroscience: Realistic Modeling for Experimentalists*, edited by Schutter, E. D., CRC Press, Boca Raton, 2000.

[161] Elf, J., Doncic, A., and Ehrenberg, M., 2003, Paper presented at the SPIE's First International Symposium on fluctuations and noise.

[162] Le Novere, N. and Shimizu, T. S., Bioinformatics **17** (2001) 575.

[163] Lemerle, C., Ventura, B. D., and Serrano, L., FEBS Letters **579** (2005) 1789.

[164] Goldbeter, A. and D. E. Koshland, J., Proc. Natl. Acad. Sci. USA **78** (1981) 6840.

[165] Frenkel, D. and Smit, B., *Understanding Molecular Simulations: From Algorithms to Applications, 2nd ed.*, Academic, Boston, 2002.

[166] Agmon, N. and Szabo, A., J. Chem. Phys. **92** (1990) 5270.

[167] Kim, H. and Shin, K. J., Phys. Rev. Lett. **82** (1998) 1578 .

[168] van Zon, J. S., Morelli, M. J., Tanase-Nicola, S., and ten Wolde, P. R., Biophys. J. **91** (2006) 4350.

[169] Berg, O. G., Paulsson, J., and Ehrenberg, M., Biophys. J. **79** (2000) 1228.

[170] Delbruck, M., J. Bacteriol. **50** (1945) 131.

[171] Novick, A. and Weiner, M., PNAS **43** (1957) 553.

[172] McAdams, H. H. and Arkin, A., Proc. Natl. Acad. Sci. USA **94** (1997) 814 .

[173] Rigney, D. R. and Schieve, W. C., Journal of Theoretical Biology **69** (1977) 761.

[174] Berg, O. G., Journal of Theoretical Biology **71** (1978) 587.

[175] Rao, C. V., Wolf, D. M., and Arkin, A. P., Nature **420** (2002) 231 .

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

[177] Paulsson, J., Physics of Life Reviews **2** (2005) 157.

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

[179] Ko, M. S. H., J. Theor. Biol. **153** (1991) 181.

[180] Peccoud, J. and Ycart, B., Theor. Popul. Biol. **48** (1995) 222.

[181] Metzler, R., Phys. Rev. Lett. **87** (2001) 068103.

[182] Karmakar, R. and Bose, I., Physical Biology **1** (2004) 197 .

[183] Pirone, J. R. and Elston, T. C., J. Theor. Biol. **226** (2004) 111 .

[184] Simpson, M. L., Cox, C. D., and Sayler, G. S., J. Theor. Biol. **229** (2004) 383 .

[185] Bialek, W. and Setayeshgar, S., Proc. Natl. Acad. Sci. USA **102** (2005) 10040 .

[186] Hornos, J. E. M. et al., Phys. Rev. E **72** (2005) 051907.

[187] Revyakin, A., Ebright, R. H., and Strick, T. R., Proc. Natl. Acad. Sci. USA **101** (2004) 4776.

[188] Riggs, A. D., Bourgeois, S., and Cohn, M., J. Mol. Biol. **53** (1970) 401 .

[189] Berg, O. G., Winter, R. B., and von Hippel, P. H., Biochemistry **20** (1981) 6929.

[190] Halford, S. E. and Marko, J. F., Nucl. Acids Res. **32** (2004) 3040 .

[191] Gerland, U., Moroz, J., and Hwa, T., Proc. Natl. Acad. Sci. USA **99** (2002) 12015.

[192] Coppey, M., Bénichou, O., Voituriez, R., and Moreau, M., Biophys. J. **87** (2004) 1640 .

[193] Slutsky, M. and Mirny, L. A., Biophys. J. **87** (2004) 4021 .

[194] Klenin, K. V., Merlitz, H., Langowski, J., and Wu, C.-X., Phys. Rev. Lett. **96** (2006) 018104.

[195] Hu, T., Grosberg, A., and Shklovskii, B. I., Biophys. J. **90** (2006) 2731 .

[196] Stanford, N. P., Szczelkun, M. D., Marko, J. F., and Halford, S. E., EMBO J. **19** (2000) 6546 .

[197] Gowers, D. M., Wilson, G. G., and Halford, S. E., Proc. Natl. Acad. Sci. USA **102** (2005) 15883 .

[198] Bremer, H., Dennis, P., and Ehrenberg, M., Biochimie **84** (2003) 597.

[199] Manor, H., Goodman, D., and Stent, G. S., J. Mol. Biol. **39** (1969) 1.

[200] Kushner, S., mrna decay, in *Escherichia coli and Salmonella*, edited by Neidhardt *et al.*, F. C., pages 849 – 860, ASM Press, Washington D.C., second edition, 1996.

[201] Smoluchowski, M., Z. Phys. Chem. **92** (1917) 129.

[202] Eigen, M., Stud. Nat. Sci. **4** (1974) 37.

[203] Shoup, D. and Szabo, A., Biophys. J. **40** (1982) 33.

[204] Kim, H. and Shin, K. J., Phys. Rev. Lett. **82** (1999) 1578.

[205] Detwiler, P. B., Ramanathan, S., Sengupta, A., and Shraimann, B. I., Biophys. J. **79** (2000) 2801 .

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

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

[208] Yu, J., Xiaojia, J., Lao, K., and Xie, X. S., Science **311** (2006) 1600.

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

[210] Thattai, M. and van Oudenaarden, A., Proc. Natl. Acad. Sci. USA **98** (2001) 8614.

[211] Bray, D., Levin, M. D., and Morton-Firth, C. J., Nature **393** (1998) 85.

[212] Andrews, S. S., Physical Biology **2** (2005) 111.

[213] Valitutti, S., Muller, S., Cella., M., Padovan, E., and Lanzavecchia, A., Nature **375** (1995) 148.

[214] Lewis, P. J., Thaker, S. D., and Errington, J., EMBO J. **19** (2000) 710.

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

Summary

Si el Señor Todopoderoso me hubiera consultado
ante de embarcarse en la Creación,
le habria recomendado algo más simple.
(Attributed to) Alfonso X (Alfonso el sabio)

Even the simplest forms of life must able to detect changes in the environment, adapt to new conditions and take decisions to optimize their survival. Bacteria, small organisms typically composed of a single prokaryotic cell, carry out these tasks by means of networks of biomolecules that interact chemically and physically. Some genes are used by a cell only under some particular conditions. The networks of protein-protein and protein-DNA interactions regulating the expression of genes in a cell called genetic networks. Genetic networks are found in all living cells. However, bacterial genetic networks are often much less complex than their equivalent in eukaryotic cells, and therefore stand out as an ideal starting point for investigating their behavior. The functioning of the networks must in general be very precise to avoid potentially lethal mistakes. Nevertheless, some bacterial genetic networks operate with very low concentrations of reactants, and are thus exposed to a strong molecular noise, which can in principle hamper the functioning of the network, and lead to less precise responses.

This Thesis investigates the effect of fluctuations on small genetic networks, by means of numerical methods. The analysis aims to highlight general properties of these networks, stemming from simple physicochemical assumptions (*i.e.* proteins move in a bacterial cell primarily by diffusion). In order to pursue this goal, coarse-grained stochastic models are simulated up to physiological time scales, neglecting the molecular details of the reactants. It has to be noted that most commonly-used soft matter numerical techniques are either not able to properly assess the effect of all the sources of fluctuations, or they are not efficient enough to guarantee a proper sampling of the interesting events. Therefore, in this Thesis, novel numerical techniques, partially developed by the author, are exploited.

In Chapter 1, a history of genetics is presented, with particular emphasis on the historical evolution of the relevant questions that pushed the scientific research ahead. To introduce the rest of the thesis, emphasis is put on the concept and mechanisms of gene regulation: some proteins can help the cell to modulate the expression of genes and turn them off when they are not needed. Recently, the regulatory relations between proteins and genes have been schematised in the form of networks, made of nodes and links. A statistical analysis of the genetic regulatory network for the bacterium *E. coli* reveals that some particular subgraphs are greatly overrepresented with respect to a randomised version of the same network. These subgraphs, indicated as motifs, have been found to carry specific functional roles. The molecules arranged in some of the motifs are present in concentrations as low as in the nanomolar range, which corresponds to a handful of molecules per bacterial cell. The biochemical reactions undergone by these molecules display then a stochastic behavior, due to the erratic behavior of the reactants and from the intrinsic randomness in the reactive events. Standard numerical techniques are not able to efficiently deal with both sources of fluctuations at the same time. In this Thesis, the Green's Function Reaction Dynamics (GFRD) technique is used to assess effects of spatial fluctuations on the dynamics of a gene under the control of a repressor, which can be considered the simplest possible motif. Moreover, if one is interested in characterising some rare event, the sampling cannot be done efficiently without some special algorithm, which in our case needs to apply to systems out of equilibrium. In this work, the novel Forward Flux Sampling (FFS) method is used to sample rare events in a model genetic switch and in a model of the bacteriophage λ genetic switch.

In Chapter 2, the dynamics of a model exclusive genetic switch is investigated. This simple genetic network has two deterministic stable states, and it can spontaneously flip from one to the other by means of stochastic fluctuations. The switch is formed by two divergently-transcribed genes under the control of the same DNA sequence (the operator). Each gene product can dimerise and bind to the operator to shut the other gene down. The switch is then characterised by three sets of reactions: birth/death of proteins, protein-protein interactions (dimerisations) and protein-DNA interactions (operator binding). The switching events are both rare and fast compared to the mean residence times in the stable states, and FFS is used to sample them efficiently. The switching rate is computed, and its variations as a function of the rate of protein-protein and protein-DNA reactions are characterised. Interestingly, the switching rate decreases if the time scales of fluctuations in DNA binding reactions become shorter, whereas it increases if the time scales of fluctuations in protein-protein interactions become shorter. This difference can be understood if the switching rate is factorised by a kinetic prefactor times the probability of being on the surface dividing the two basins of attraction, in analogy with equilibrium systems. It is shown that varying the dimerisation rates changes only the first contribution, while varying the operator binding rates change both factors. This result is elucidated by a sampling of the switching pathways as a function of several order parameters: changing the rate of DNA binding can drastically change the location of the switching paths, while varying the rate of dimerisation does not change the path locations, but only affects the

speed these paths are travelled.

The model switch displays fluctuations on several time scales. When some of the reactions become very fast, the simulation of the system becomes increasingly inefficient. It would therefore be desirable to integrate out some fast degrees of freedom of the switch, while preserving its equilibrium and dynamical properties. In Chapter 3, several dynamical coarse-graining techniques are applied to the model genetic switch described in Chapter 2. Protein-protein and protein-DNA reactions are integrated out either singularly or together, with techniques solving either the macroscopic rate equation or the chemical master equation of the system. All these techniques work well when reproducing the steady-state probability distribution of the switch. However, the macroscopic approach always leads to the wrong result when computing the fluctuation-driven switching rate (obtained with FFS): the switch works with a low number of molecules, and a mean-field treatment leads to large errors. Therefore, only protein-protein interaction can be safely integrated out, using a master equation-based approach; fluctuations in protein-DNA reactions are essential to achieve the transition and can not be removed.

In Chapter 4, we use the results of the previous Chapters to compute the spontaneous switching rate of a real biological system: the bacteriophage λ. This virus infects the bacterium *E. coli* and can enter one of two alternative stable states: it either integrates its genome on the host chromosome and stays dormant for many host generations (lysogenic pathway), or it replicates as much as possible, killing the host and releasing the progeny (lytic pathway). Since the system is subject to molecular noise, spontaneous transitions between the states should be expected. However, the spontaneous switching rate from the lysogenic state is extremely low, for reasons that have not been yet understood. Several modelling attempts, based on equilibrium assumptions, have computed this rate and overestimated the measured value by several order of magnitudes. Phage λ being one of the best characterised systems in molecular biology, the large quantity of available data is exploited to build a fully stochastic model of the core genetic network keeping the system in either of the two mutually excluding pathways. With FFS, a spontaneous switching rate of $\sim 10^{-15}\mathrm{s}^{-1}$, that is $\sim 10^{-12}$ per generation per cell is measured. This result is compatible with experimental measurements, which provide an upper bound of $\sim 10^{-9}$ per generation per cell. Furthermore, the effects of macromolecular crowding within the cell are investigated: the caging effect around bound complexes could significantly shift chemical equilibria as measured *in vitro*, and increase the stability of the lysogenic state. Finally, a DNA loop, able to mechanically "lock" the lysogenic state has recently been observed. The model is extended to incorporate these additional features, to show that the spontaneous switching rate decreases by several orders of magnitude.

Brownian Dynamics algorithms are used to simulate chemical and biological systems in time and space. However, when reactions are included, a violation of the detailed balance rule could introduce systematic errors in the simulation. Chapter 5 describes a Brownian Dynamics algorithm which rigorously obeys detailed balance. Stringent tests reveal that the algorithm is able to reproduce the equilibrium properties of a simple reaction-diffusion system, and its dynamics for small enough time steps. The algorithm is applied

to a “push-pull” network where two antagonistic enzymes covalently modify a substrate. The diffusion of the reactants can strongly reduce the gain of the response curve for this network.

The diffusive behavior of regulating molecules can strongly increase the noise in protein production. In Chapter 6, the effects of spatial fluctuations of a repressor molecule on the regulation of a gene are investigated with the GFRD scheme. This method is able to highlight a purely spatial effect: the possibility of multiple fast rebindings of the repressor molecule due to the proximity to its reaction site after dissociation. Therefore, the repressor turns the gene on and off on time scales much longer than its reaction rates would suggest. As a result, the noise in gene expression is substantially increased when the repressor is found at low concentrations. However, the time scales of these rebindings are so short that an RNA polymerase cannot effectively initiate the transcription of the gene during the time between two consecutive rebindings. This time scale separation between the repressor rebindings and RNA polymerase associations can be exploited to account for the effects of spatial fluctuations via a simulation in which space is absent, but reaction rates are suitably renormalised. Finally, a frequency analysis of the system highlights how the slow dynamics of proteins makes the network behave like a low-pass filter.

Sintesi

È una Via Crucis.

Questo libretto illustra e certifica il lavoro da me svolto durante il mio dottorato all'Istituto di Fisica Atomica e Molecolare (AMOLF) di Amsterdam. Essendo la comunità scientifica (e ormai il mondo tutto intero) obbligata ad utilizzare la lingua di Albione, o qualche sua grezza storpiatura, per comprendersi senza ambiguità, esso è quasi interamente redatto in lingua inglese. Fortunatamente, la magnanimità dell'Università di Amsterdam permette ai suoi pupilli di dedicare qualche pagina nella lingua materna del candidato dottore ad un riassunto dei contenuti di tale libretto.

Tutti noi conosciamo qualche coppia di fratelli gemelli. Qualcuno di noi si sarà imbattuto in alcune di tali coppie i cui elementi presentano un'impressionante somiglianza fisica. In questo caso i gemelli sono detti *omozigoti*: nella parte più intima di ogni singola cellula di ambedue le persone è conservata e custodita gelosamente una copia, identica, di una lunga molecola a doppia spirale, stivata con cura come un microscopico gomitolo: l'acido deossiribonucleico, comunemente noto come DNA. Eppure, vi sarà capitato di parlare con le mamme di tali fratelli gemelli, le quali invariabilmente ripeteranno: "di carattere, non potrebbero essere più diversi!". Ad un'analisi minuziosa, inoltre, piccole differenze fra i due germani risulteranno presto evidenti, tanto da renderli perfettamente distinguibili ad un occhio allenato. Come sono possibili queste differenze, morfologiche o più sottili, se l'informazione contenuta nei due ovuli fecondati era esattamente la stessa al momento del concepimento? C'è quindi qualcos'altro che regola e dirige la formazione del feto e successivamente della persona umana oltre al DNA? Questo è uno dei tipici problemi della genetica moderna, e non richiede l'invocazione della presenza di Dio per essere risolto, ma solo molta pazienza, creatività, tempo, e un laboratorio ben finanziato, dotato di apparecchiature moderne.

Molti lettori si ricorderanno che mi sono laureato in Fisica; qualcuno si ricorderà perfino di avermi sentito parlare di mercati finanziari nel periodo della mia tesi di laurea. Ebbene, cosa c'entrano adesso i gemelli, il DNA, la genetica? C'entrano, c'entrano. Ancor prima di iniziare a studiare Fisica, mi venne ricordato come una delle qualità peculiari

del fisico fosse la sua duttilità. Si impara la struttura degli atomi e dei cristalli, la Meccanica Quantistica e quella Statistica, ma quello che più resta è la capacità di affrontare e risolvere con mentalità analitica e quantitativa problemi complessi, misurarsi ed esplorare realtà ignote, gestire rapidamente grandi quantità di dati e conoscenze. Per questo, molti bravi fisici fanno una brillante carriera in campi quali la consulenza, la finanza o la gestione aziendale. Nel mio caso, però, l'analogia è più profonda. Da sempre mi sono interessato agli effetti che le *fluttuazioni* possono avere sul comportamento di un sistema, sia esso appartenente alla tradizionale termodinamica, oppure ad altre diramazioni della scienza: l'andamento di un mercato finanziario, per esempio, oppure l'interno di una cellula. Il movimento della polvere di farina sull'acqua, l'andamento irregolare di un'azione in Borsa, il rumore di fondo di una radio, sono tutti effetti di fenomeni interessati da fluttuazioni: la parola "rumore" ne è spesso considerata un sinonimo dagli addetti ai lavori. Tali fluttuazioni sono parzialmente responsabili della diversità di organismi dotati di un identico corredo genetico, come i gemelli omozigoti. Nonostante il loro carattere intrinsecamente aleatorio, tali fluttuazioni possono essere caratterizzate *statisticamente*, cioè calcolando le loro proprietà medie e le loro deviazioni da tali medie. Questa intera Tesi rigurgita di probablità, loro distribuzioni e proprietà. Le tecniche utilizzate per analizzare queste fluttuazioni sono molteplici, e si avvalgono di calcoli analitici (quelli che si fanno con carta e penna), oppure di simulazioni al calcolatore, dove un codice di programmazione viene "fatto girare" per ore ed ore (spesso per giorni o settimane) su una o più macchine, simulando così il comportamento di un sistema complesso le cui equazioni sono troppo difficili da risolvere "a mano". In questa Tesi, ho prevalentemente utilizzato tecniche del secondo tipo.

Per entrare ora più propriamente nel vivo della mia Tesi, sono costretto ad introdurre qualche essenziale concetto di biologia molecolare. Le tecniche della mia ricerca, infatti, sono sì derivate dalla fisica, ma l'oggetto è biologico; la intima fusione di questi due punti di vista è alla base della disciplina all'interno della quale la mia ricerca si colloca: la biofisica. Tutti sappiamo che alcune regioni del DNA sono chiamate *geni*, i portatori dell'informazione genetica. In generale, non si sa con precisione dove tali geni siano localizzati sulla doppia elica; di certo si sa che tali pezzi di DNA contengono l'informazione che serve alle cellule per sintetizzare le nostre *proteine*. Tali proteine ci permettono di costruire il nostro corpo, organizzarlo in strutture più complesse come cellule, tessuti e organi, ed espletare tutte le funzioni vitali. È interessante notare come nella specie umana i geni, e quindi le proteine, siano circa 30000 e rappresentino solo l'1% del DNA. La funzione dell'altro 99% non è al momento conosciuta. Ognuno di noi produce una versione "personalizzata" delle proprie 30000 proteine, riflettendo la nostra diversità di persone umane. Tuttavia, esse risultano estremamente simili da individuo ad individuo, allo stesso modo in cui ogni essere umano ha due braccia, due gambe, una testa, e così via. Il prossimo concetto fondamentale è che non tutte le cellule hanno sempre bisogno di tutte le proteine possibili. Visto che la loro produzione ha un costo, una cellula si sforzerà di evitare di produrre le proteine che non servono. Chiaramente, alcune proteine legate ad attività "di base" come la produzione di energia o la duplicazione del DNA verranno

espresse in tutte le cellule; tuttavia, è altresì evidente che le proteine di cui ha bisogno una cellula neuronale saranno in gran parte diverse da quelle necessarie per il corretto sviluppo di una cellula di un muscolo, o della pelle. Si stima che nella specie umana circa la metà delle nostre proteine siano utilizzate solo durante lo sviluppo embrionale, e mai più utilizzate durante la nostra vita extrauterina.

Si evidenzia quindi la necessità di *regolare* l'espressione dei geni e dunque la produzione delle proteine. Ogni cellula utilizza diverse strategie, tutte volte al conseguimento di tale obiettivo. E qui le cose iniziano a farsi complicate. La regolazione genetica negli organismi superiori, come i mammiferi, si avvale di meccanismi estremamente complessi e interdipendenti, il cui studio non può essere condotto prescindendo dal contesto in cui la cellula vive, e sui quali si conosce molto poco (è questa infatti una delle ragioni della nostra relativa ignoranza e impotenza di fronte ai casi di tumore). Sfoderando un affilato rasoio di Ockham, operiamo dunque un radicale taglio alla complessità dell'organismo la cui regolazione genetica vogliamo studiare. Dagli organismi superiori scendiamo ad esseri più semplici come i comuni lombrichi, giù giù fino a forme di vita microscopiche composte da una sola cellula, come il lievito di birra, fino ad arrestarci ai più semplici organismi autonomi conosciuti: i batteri. Anch'essi formati da un'unica cellula, essa è però di un tipo particolarmente semplice, detto *procariote*, mancante cioè di molti degli apparati e compartimenti presenti nelle cellule superiori, o *eucarioti*, presenti nel nostro corpo. I batteri sono piccoli (nel punto alla fine di questa frase se ne possono contare fino a mezzo milione) e relativamente rudimentali, nondimeno costituiscono una delle più robuste forme di vita, in grado di sopravvivere in qualsiasi condizione ambientale. È sorprendente apprendere che nel nostro organismo, ci sono all'incirca 10 volte più batteri, appartenenti a circa 200 specie diverse, che cellule umane; tuttavia, essendo i primi così piccoli e leggeri, non riescono a far pesare eccessivamente la loro preponderanza. In questa Tesi, quando parleremo di batteri, ci riferiremo costantemente alla specie di gran lunga più popolare nella comunità scientifica, dato il suo prestarsi paziente ad ogni tipo di tortura in laboratorio: il batterio *Escherichia coli*, che vive abbondante nel nostro intestino, produce alcune vitamine, e ci aiuta a completare la digestione.

Anche i batteri hanno bisogno di regolare i loro geni; il modo in cui li regolano è però molto più semplice che nelle cellule eucarioti: generalmente il batterio produce una proteina che si lega al DNA in corrispondenza dell'inizio del gene, e lo blocca. L'azione di bloccaggio è puramente meccanica: la produzione di un gene inizia con l'arrivo di una grossa molecola, chiamata RNA polimerasi, che apre la doppia elica e scorre come una cerniera, creando una copia del gene. Se una proteina blocca fisicamente l'inizio del gene, questo processo non può avvenire, e la proteina codificata nel gene non viene espressa. Oltre a questo semplice meccanismo, molti altri sono possibili. Quasi tutti coinvolgono l'utilizzo di proteine (codificate quindi anch'esse da qualche parte sul DNA) per regolare l'espressione di altre proteine. Questo insieme di relazioni di regolazione tra geni e proteine può essere visualizzato come una rete, un grafico dotato di punti e frecce che li collegano. Nella mia Tesi, ho analizzato in grande dettaglio alcuni componenti essenziali di queste reti: è infatti dimostrato che, almeno per i batteri, l'elevato livello di

complessità delle reti è ottenuto con la ripetizione di alcune sottoreti semplici dette *motivi*, collegati fra di loro in architetture via via più complicate, esattamente come succede in un circuito elettrico, dove elementi come resistenze, condensatori, induttanze e amplificatori sono combinati per ottenere radio, motori, o altri dispositivi elettrici o elettronici.

Chiudendo il circolo aperto qualche pagina fa, durante il mio Dottorato, mi sono interessato in particolare di come le fluttuazioni possano interferire con il funzionamento di queste "reti genetiche", che deve essere il più possibile preciso, pena lo sviluppo di comportamenti aberranti e spesso la morte del batterio. È importante capire che queste fluttuazioni possono avere origini diverse: gran parte della Tesi è dedicata a capire quali di queste fluttuazioni siano importanti e quali no per il funzionamento di alcuni motivi. Le proteine a cui siamo interessati si muovono nella cellula per *diffusione*, esattamente come fanno le particelle colorate di una goccia di vino rosso in un bicchiere d'acqua. Il moto disordinato delle proteine, è una di tali sorgenti di rumore; la loro presenza, spesso, in un numero infimo, inasprisce questo ed altri effetti, amplificando le deviazioni del sistema rispetto al suo comportamento medio. Nel Capitolo 6 di questa Tesi, l'importanza del comportamento diffusivo di proteine regolatrici è accuratamente quantificato, rivelandosi la principale fonte di fluttuazioni nel sistema. Per ottenere questo risultato, ho utilizzato un algoritmo nuovo, che ho contribuito a sviluppare, chiamato Green's Function Reaction Dynamics. Tale metodo, estremamente complicato, permette di risolvere tutta una serie di problemi tecnici utilizzando la soluzione analitica del problema di reazione-diffusione per una coppia di particelle, e assicurandosi che un problema di qualsiasi grado di complicazione venga ridotto ad una somma di problemi a uno e due corpi. Un algoritmo più tradizionale, detto di Dinamica Browniana, viene spesso usato per risolvere problemi di reazione-diffusione. Tuttavia, come è dimostrato nel Capitolo 5, una frettolosa implementazione di tale schema può condurre a errori sistematici, inficiando il risultato delle simulazioni. Per ottenere risultati corretti, bisogna prestare molta attenzione a non violare la regola del bilancio dettagliato nei casi di reazioni reversibili. Nel Capitolo 5 ho indicato una semplice ricetta per evitare tali errori, ottenendo così un algoritmo rigorosamente corretto anche se non particolarmente efficiente.

Il Capitolo 1 ha la funzione di introdurre, spero godibilmente, la storia della genetica, l'evoluzione delle questioni scientifiche che hanno diretto lo sviluppo di questo campo a partire da Darwin fino a giorni nostri, il concetto di regolazione genetica che ho cercato rudimentalmente di rivisitare in queste pagine, l'oggetto della Tesi (le reti genetiche) e una rassegna dei principali metodi, vecchi e nuovi, utilizzati da me e dagli altri addetti ai lavori per studiare i sistemi biochimici soggetti a fluttuazioni. I Capitoli 2 e 3 sono invece un crescendo di risultati su un altro importante elemento costituente le reti di regolazione genetica nei batteri: l'interruttore genetico. Esso funziona in maniera esattamente analoga ad un interruttore elettrico: commuta stabilmente tra due stati alternativi (acceso/spento). Gli stati di cui parliamo in questo contesto si riferiscono ai livelli di due diverse proteine, che possono essere, rispettivamente, alto per la prima e basso per la seconda, o viceversa, ma mai entrambi alti o bassi. Fisicamente, tale meccanismo è ottenuto giustapponendo due geni sul DNA e inserendo una regione di regolazione in comune fra di loro. Le

fluttuazioni presenti nel sistema inducono transizioni spontanee fra i due stati, esattamente come una pallina posta in una terrina da insalata che vibri vigorosamente, prima o poi ne esce. Il mio lavoro in questo caso è stato quello di caratterizzare esattamente quali fluttuazioni sono responsabili per queste transizioni spontanee, e in quale misura; come esattamente avvengano i processi di transizione; quale sia la frequenza delle transizioni e come essa cambi al variare delle condizioni del sistema. Siccome tali transizione sono sia rare sia molto rapide, per caratterizzarle adeguatamente, ho utilizzato un altro algoritmo originale, che ho contribuito a sviluppare, chiamato Forward Flux Sampling.

Tutto questo corpo di conoscenze che ho accumulato su un semplice modello di interruttore genetico sono state sfruttate, nel Capitolo 4, per dare l'assalto alla modellizzazione di un famigerato sistema reale: il batteriofago λ. Tale organismo è un virus infettante, tanto per cambiare, il batterio *E. coli*, ed utilizza un interruttore genetico per scegliere quale comportamento tenere una volta penetrato all'interno di una cellula batterica. Il virus può infatti moltiplicarsi enormemente e distruggere l'organismo ospite, oppure può inserire il DNA virale in quello dell'ospite, e rimanere quiescente per molte generazioni, diffondendosi nella popolazione batterica, pronto a distruggere il proprio ospite quando esso sperimenti una situazione di stress. Il fago λ è ben noto alla comunità dei biologi molecolari, poiché proprio studiandolo sono nati il concetto di regolazione genetica e una buona parte della biologia molecolare moderna. La grande quantità di esperimenti condotti sul fago hanno portato ad una sua caratterizzazione estremamente accurata, ben oltre gli standard di un generico sistema biologico. Tuttavia rimane ancora misteriosa una delle proprietà di tale semplice organismo: l'estrema stabilità del suo stato quiescente, altrimenti detto lisogenico. Diversi modelli hanno cercato di stimare la frequenza di transizione spontanea dell'interruttore del fago, ma sono tutti giunti ad una sovrastima, errata per diversi ordini di grandezza. Nondimeno, tali modelli si reggono tutti su qualche ipotesi di equilibrio chimico per alcune reazioni coinvolte nella regolazione dell'interruttore: assumere equilibrio significa rimuovere fluttuazioni, e il nostro lavoro preliminare ci ha portati a concludere che alcuni tipi di fluttuazioni non possono essere rimosse dal sistema senza snaturarne il comportamento dinamico. Abbiamo quindi creato un modello completamente stocastico, che ci ha permesso di ottenere una frequenza di transizioni spontanee finalmente nei limiti dell'evidenza sperimentale (anch'essa, in verità, molto difficile da ottenere e riprodurre). Infine, abbiamo cercato di quantificare quali possano essere gli effetti di altre due condizioni che potrebbero ulteriormente stabilizzare lo stato quiescente del virus: l'affollamento dello spazio intracellulare e un anello del DNA. Il primo contributo si riferisce alla grande quantità di macromolecole racchiuse dalla membrana cellulare, che rendono lo spazio intracellulare molto diverso da una soluzione diluita. La presenza di tutti questi agenti affollanti aspecifici creerebbe una pressione efficace che favorirebbe il mantenersi di stati legati attraverso un "ingabbiamento" dei loro componenti, prevenendone la possibilità di una "fuga". Nel nostro modello, tale effetto è incluso attraverso un adeguato spostamento degli equilibri chimici dovuto a questa interazione da volume escluso. L'altra condizione deriva dalla presenza di un altro sito di regolazione del sistema, situato a grande distanza sul DNA dai geni dell'interruttore: ad esso, proteine regolatrici si pos-

sono legare, e interagire con il sito principale di regolazione piegando il DNA ad anello (in realtà tale analogia visiva non è del tutto appropriata, dato che alle scale di lunghezza che stiamo considerando, il DNA è completamente flessibile e assomiglia più as un piatto di spaghetti che ad un anello di gomma). L'anello di DNA sigillerebbe lo stato quiescente e contribuirebbe quindi alla sua grande stabilità. Nonostante la scarsezza di dati sul sito ausiliario di regolazione, abbiamo incorporato l'anello nel nostro modello mediante una serie di stime fisiche, e verificato che esso contribuisce a diminuire la frequenza di transizione spontanea di svariati ordini di grandezza.

Alcuni motivi presenti nelle reti genetiche batteriche sono dunque soggetti a forti fluttuazioni, radicate nell interazioni fisiche e chimiche fra le molecole biologiche che le costituiscono. La mia analisi ha svelato che alcuni di tali motivi adottano strategie per filtrare questo rumore, in particolare quello ad alta frequenza, sfruttando la vita media delle proteine (che è nell'ordine delle ore) come tempo di integrazione. Tuttavia, nel caso degli interruttori genetici, le fluttuazioni, e in particolare quelle presenti nell'interazione tra proteine regolatrici e DNA, possono, controintuitivamente, conferire *maggiore stabilità* al sistema, e prevenire transizioni spontanee tra stati stazionari.

Samenvatting

Zelfs de meest eenvoudige levensvormen moeten veranderingen in hun omgeving kunnen waarnemen, zich kunnen aanpassen aan nieuwe omstandigheden en beslissingen nemen die hun overlevingskansen maximaliseren. Bacteriën, kleine organismen die typisch bestaan uit een enkele prokaryotische cel, doen dit door middel van netwerken van biomoleculen die onderling zowel chemische als fysische wisselwerkingen hebben. Sommige genen worden door de cel alleen onder bepaalde omstandigheden gebruikt. De netwerken van eiwit-eiwit- en eiwit-DNA-interacties die de expressie van genen in een cel reguleren heten genetische netwerken. Genetische netwerken komen voor in alle levende cellen, maar bacteriële genetische netwerken zijn vaak veel minder complex dan hun tegenhangers in eukaryotische cellen, en vormen daarom een ideaal startpunt om het gedrag van deze netwerken te onderzoeken. De netwerken moeten zeer nauwkeurig functioneren om mogelijk fatale vergissingen te voorkomen. Desondanks opereren sommige bacteriële netwerken met zeer lage concentraties van reactanten waardoor zij worden blootgesteld aan sterke moleculaire ruis, wat in principe kan interfereren met het functioneren van het netwerk.

Dit proefschrift onderzoekt het effect van fluctuaties op kleine genetische netwerken, gebruik makend van numerieke methoden. De analyse tracht algemene eigenschappen van deze netwerken te benadrukken, uitgaande van eenvoudige fysisch-chemische aannames (bijvoorbeeld dat eiwitten in een bacteriële cel voornamelijk bewegen door middel van diffusie). Met dit doel worden vereenvoudigde stochastische modellen gesimuleerd tot op fysiologische tijdsschalen, onder verwaarlozing van de moleculaire details van de reactanten. Er moet opgemerkt worden dat de gebruikelijke numerieke methoden uit het veld van de zachte materie ofwel niet in staat zijn om de effecten van alle bronnen van fluctuaties te bepalen, ofwel niet efficiënt genoeg zijn om een goed sample van de interessante gebeurtenissen te produceren. Om deze reden worden in dit proefschrift nieuwe numerieke methoden toegepast, die mede door de auteur zijn ontwikkeld.

In hoofdstuk 1 wordt een korte geschiedenis van de genetica gepresenteerd, met een bijzondere nadruk op de historische evolutie van de relevante vragen die hebben bijgedragen aan de vooruitgang van het wetenschappelijke onderzoek. Ter introductie van de rest van dit proefschrift wordt de nadruk gelegd op het concept en de mechanismes van gen-

regulatie: sommige eiwitten kunnen de cel helpen bij het reguleren van de expressie van genen en deze uitschakelen wanneer ze niet nodig zijn. Recentelijk zijn de regulerende interacties tussen eiwitten en genen geschematiseerd in de vorm van netwerken, opgebouwd uit knooppunten en verbindingen. Een statistische analyse van het genetische regulatienetwerk van de bacterie *E. coli* heeft aan het licht gebracht dat bepaalde subgrafen in hoge mate oververtegenwoordigd zijn in vergelijking met een willekeurig verbonden versie van hetzelfde netwerk. Deze subgrafen, die worden aangeduid als motieven, blijken specifieke functies te hebben. De concentraties van de moleculen die in sommige van deze motieven voorkomen, zijn van de orde van een nanomolair, wat overeenkomt met een handvol moleculen per bacteriële cel. De biochemische interacties tussen deze moleculen zijn dus stochastisch van aard, vanwege de onvoorspelbare bewegingen van de reactanten en de intrinsieke willekeurigheid van de reacties. De gangbare numerieke methoden zijn niet in staat om beide soorten fluctuaties tegelijkertijd op een efficiënte wijze te simuleren. In dit proefschrift wordt de Greense-functiereactiedynamicamethode (GFRD) gebruikt om de effecten te meten die ruimtelijke fluctuaties hebben op de dynamica van een gen onder controle van een repressor, wat als het meest eenvoudige motief kan worden beschouwd. Bovendien, wanneer men geïnteresseerd is in het karakteriseren van een zeldzame gebeurtenis, kan het samplen niet worden gedaan zonder een speciaal algoritme, dat in dit geval ook geldig moet zijn voor systemen buiten evenwicht. In dit werk wordt een nieuwe methode met de naam *Forward Flux Sampling* (FFS) gebruikt om zeldzame gebeurtenissen te samplen in een typische genetische schakelaar en een model van de genetische schakelaar van bacteriofaag λ.

In hoofdstuk 2 wordt de dynamica van een typische genetische schakelaar onderzocht. Dit eenvoudige genetische netwerk heeft twee deterministisch stabiele toestanden en het kan spontaan schakelen tussen deze toestanden door middel van stochastische fluctuaties. De schakelaar wordt gevormd door twee genen die divergent getranscribeerd worden en via dezelfde DNA-sequentie worden aangestuurd (de operator). Elk van beide genproducten kan een dimeer vormen en in die vorm aan de operator binden om het andere gen te onderdrukken. De schakelaar wordt dus gekarakteriseerd door drie typen reacties: de geboorte en sterfte van eiwitten, eiwit-eiwit-interacties (dimerisatie) en eiwit-DNA-interacties (het binden aan de operator). De schakelaar schakelt zelden, en doet dat snel in vergelijking met de gemiddelde verblijfstijd in een van de stabiele toestanden, dus wordt FFS gebruikt om ze efficiënt te samplen. De schakelfrequentie wordt berekend, en de variatie daarvan als functie van de frequenties van de eiwit-eiwit- en eitwit-DNA-reacties wordt gekarakteriseerd. Opvallend is dat de schakelfrequentie afneemt als de tijdschaal van de fluctuaties in DNA-bindingsreacties afneemt, terwijl de schakelfrequentie juist toeneemt als de tijdschaal van de fluctuaties van eiwit-eiwitreacties korter wordt. Dit verschil kan worden verklaard als de schakelfrequentie wordt gefactoriseerd in een kinetische voorfactor maal de waarschijnlijkheid om zich bovenop de “barrière” tussen de twee aantrekkende bassins te bevinden, analoog aan systemen in evenwicht. Er wordt aangetoond dat het variëren van de dimerisatiefrequenties alleen de eerste bijdrage verandert, terwijl het variëren van de operatorbindingsfrequentie beide factoren beïnvloedt.

Dit resultaat wordt verhelderd door middel van het samplen van de transitiepaden als een functie van verscheidene ordeparameters: het veranderen van de DNA-bindingsfrequentie kan de ligging van deze schakelpaden drastisch beïnvloeden, terwijl het variëren van de dimerisatiefrequentie geen effect heeft op de ligging van de paden, maar slechts op de snelheid waarmee deze paden worden afgelegd.

De modelschakelaar vertoont fluctuaties op verschillende tijdschalen. Naarmate de reactiesnelheden toenemen, neemt de simulatie-efficiëntie af. Het is daarom aantrekkelijk om enkele vrijheidsgraden van de schakelaar uit te integreren, onder behoud van de evenwichts- en dynamische eigenschappen. In hoofdstuk 3 worden verscheidene dynamische simplificatietechnieken toegepast op de genetische modelschakelaar zoals beschreven in hoofdstuk 2. Eiwit-eiwit and eiwit-DNA-reacties zijn ofwel afzonderlijk ofwel gezamenlijk uitgeïntegreerd, door middel van technieken die respectievelijk de macroscopische frequentievergelijkingen en de chemische *master equation* oplossen. Al deze technieken werken goed om de steady-state waarschijnlijkheidsverdeling van de schakelaar te reproduceren. Helaas leidt de macroscopische aanpak altijd tot een verkeerd resultaat bij het berekenen van de fluctuatie-gedreven schakelfrequentie (bepaald met FFS): de schakelaar werkt met een zeer klein aantal moleculen, waardoor een gemiddeldeveldbenadering tot grote fouten leidt. Om deze reden kan alleen de eiwit-eiwitinteractie veilig uitgeïntegreerd worden, met een aanpak die gebaseerd is op de *master equation*; schommelingen in eiwit-DNA-reacties zijn essentieel voor het bereiken van de transitie en kunnen niet verwijderd worden.

In hoofstuk 4 gebruiken we de resultaten van de voorgaande hoofdstukken om de spontane schakelfrequentie van een echt biologisch systeem te berekenen: bacteriofaag λ. Dit is een virus dat de bacterie *E. coli* infecteert en zich vervolgens in een van twee stabiele toestanden kan bevinden: ofwel het integreert zijn genoom in het chromosoom van de gastcel en blijft gedurende vele generaties in deze slapende toestand (lysogeen), of het reproduceert zich zo vaak mogelijk, waardoor de gastcel sterft en het nageslacht vrijkomt (lytisch). Omdat het systeem onderhevig is aan moleculaire ruis, verwacht men spontane transities tussen de beide toestanden. In de praktijk is de spontane schakelfrequentie vanuit de lysogene toestand echter zeer laag, om nog onbekende redenen. Verschillende modelleerpogingen, gebaseerd op evenwichtsaannames, hebben deze frequentie berekend en de gemeten waarden met enkele ordes van grootte overschat. Omdat bacteriofaag λ een van de best gekarakteriseerde systemen in de biologie is, hebben we de grote hoeveelheid beschikbare data gebruikt om een volledig stochastisch model te construeren van het basale genetische netwerk dat het systeem in een van de twee toestanden houdt. Met behulp van FFS is een spontane schakelfrequentie van $\sim 10^{-15}\mathrm{s}^{-1}$, dat wil zeggen $\sim 10^{-12}$ per generatie per cel, gemeten, in overeenstemming met experimentele waarnemingen. Bovendien zijn de effecten van macromoleculaire *crowding* binnen de cel onderzocht: het effect van kooivorming rondom gebonden complexen kan belangrijke verschuivingen veroorzaken in *in vitro* metingen van chemische evenwichten en kan de stabiliteit van de lysogene toestand vergroten. Tenslotte is onlangs een DNA-ring waargenomen, die in staat is om op mechanische wijze de lysogene toestand te “bevriezen”. Het model is

uitgebreid met deze additionele eigenschappen, waardoor de spontane schakelfrequentie wederom met enkele ordes van grootte afneemt, maar nog steeds niet strijdig is met de experimentele waarnemingen.

Brownse-Dynamica-algoritmes worden gebruikt om chemische en biologische systemen te simuleren in de ruimte en tijd. Zodra er echter reacties in worden opgenomen, kan een schending van de detailed-balanceregel ertoe leiden dat er systematische fouten ontstaan in de simulatie. In hoofdstuk 5 wordt een Brownse-Dynamica-algoritme beschreven dat wel strikt voldoet aan de detailed-balanceregel. Rigoreuze tests tonen aan dat het algoritme in staat is om de evenwichtseigenschappen van een simpel reactie-diffusiesysteem te reproduceren en, voor tijdstappen die klein genoeg zijn, ook de dynamica. Het algoritme wordt toegepast op een "push-pull"-netwerk waarin twee antagonistische enzymen een substraat op covalente wijze modificeren. De diffusie van de reactanten kan de steilheid van de responscurve van dit netwerk sterk verminderen.

Het diffusieve gedrag van regulerende moleculen kan de ruis in eiwitproductie in belangrijke mate versterken. In hoofdstuk 6 worden de effecten van een repressormolecuul op de regulatie van een gen onderzocht met behulp van GFRD. Deze methode is in staat om een zuiver ruimtelijk effect aan te tonen: de mogelijkheid van een repressormolecuul om meermaals snel achter elkaar te binden vanwege zijn nabijheid tot de reactielocatie na dissociatie. Door dit effect schakelt de repressor het gen aan en uit op tijdsschalen die veel langer zijn dan verwacht op basis van de reactiefrequenties. Dit heeft tot gevolg dat de ruis in genexpressie substantieel verhoogd wordt als de repressor in lage concentraties aanwezig is. De tijd tussen twee van dergelijke opeenvolgende bindingsmomenten is echter zo kort dat een RNA-polymerase effectief niet in staat is om de transcriptie te initiëren. Het verschil in tijdschalen tussen de repressor-herbindingen en de RNA-polymerase-associatie verklaart dat de ruimtelijke effecten kunnen worden verdisconteerd in een simulatie waarin de ruimte afwezig is, maar de reactieconstanten op passende wijzen zijn gerenormaliseerd. Tenslotte laat een frequentieanalyse van het systeem zien dat de langzame eiwit-dynamica ervoor zorgt dat het netwerk zich gedraagt als een laagfrequente filter.

In viaggio

Caelum, non animum mutant
qui trans mare currunt
Quintus Horatius Flaccus

Un Dottorato è un lungo viaggio. Si parte leggeri, pieni di entusiasmo e curiosità. Si visitano luoghi nuovi, si incontrano persone intelligenti e arroganti, umili e ambiziose. Si inizia a lavorare e si scopre che l'attività di ricerca ha bisogno di almeno dieci volte più tempo di quanto avessimo pensato. A metà strada, come in ogni cammino, il traguardo sembra lontanissimo, l'entusiasmo si spegne, i progressi fatti sembrano quasi inesistenti, e una domanda si ripresenta ogni notte prima di addormentarsi, come dipinta sul soffitto: "chi me l'ha fatto fare?". Poi, d'improvviso, la meta appare all'orizzonte, la prima scintilla di luce si accende in fondo ad un tunnel che sembrava infinito, finalmente gli sforzi producono risultati, e allora è tutto un affrettarsi a finire, controllare, scrivere, fare i conti con il tempo che resta; l'entusiasmo, più maturo, ritorna, le ore di lavoro aumentano e, alla fine, si concretizzano in un libretto come questo. Nel mio caso, il viaggio è stato anche fisico, mi ha portato a convivere e sopravvivere in una realtà molto più aliena di quanto avessi immaginato, a conoscere culture e tradizioni diverse, e a stemperare le mie robuste radici italiche in abitudini e modi di fare stranieri. Tutto questo non sarebbe stato possibile senza la presenza di moltissime persone, con cui ho vissuto innumerevoli momenti indimenticabili, che mi hanno aiutato e sostenuto in questi ultimi anni. Cercherò qui di ringraziarle, e mi scuso fin d'ora se, dato il loro numero, non potrò citare tutti coloro che hanno contribuito a rasserenare una mia giornata con un sorriso inaspettato.

First of all, I should thank my supervisor, Pieter Rein. The impression he made on me the very first time I met him at the Vrije Universiteit (the "spilungone simpatico", for those who have my first emails) was one of the elements who brought me on this side of the Alps. In four years we learned to know each other, acknowledge our differences, and find a pretty way to work together. His legendary enthousiasm and stamina conveyed to me a profound, genuine passion for science, and his humanity and comprehension during the disease of my father alleviated my troubles relieving me from work pressure when I could not have born it.

The second word goes to my promoter, Daan. After having developped a rough expander for his unbelievably sharp remarks, I started enjoying the scientific part of our conversations, besides

the witty, interesting chats goading the curiosity of the whole corridor. I will value and remember his availability, together with his support and help with the final part of my PhD.

This Thesis would not be there without the help of three more persons who encouraged me during the most difficult moments, and filled me with new energies and enthousiasm, in different ways. Ass. Prof. Angelo Cacciuto, che mi ha accolto ad Amsterdam e mi ha insegnato quanto la comunità scientifica possa essere variegata. La sua limpida visione scientifica, e i suoi innumerevoli consigli pratici sono stati preziosi per il mio dottorato. Le serate passate a cucinare, suonare la chitarra, vedere film e celebrare Bacco insieme, oltre ad aver creato e cementato un'amicizia che dura inalterata, sono fra i momenti che ricordo con più felicità durante la mia permanenza olandese. Grazie. The second person is Dr. Rosalind Allen: our collaboration started almost as a joke, from a student frustrated by the lack of results in his main project; the fruits of our joint efforts now span more than half of this Thesis. Our empathy created, at least from my part, a perfect work relationship, where targets were set and reached together, paying attention to the scientific quality of research, but also to needs of the person. I can not thank her enough for the possibility she gave me to visit her group in Edinburgh, thus introducing me to a country I immediately loved, and for the time and patience she always had with me. The third is Dr. Sorin Tănase-Nicola, who solved a countless number of analytical problems for me, flooded me with interesting papers, and helped me understanding basic concepts of Physics I should have already known. His enthousiasm and conviviality during summerschools contributed to add fun and laughter to the science part of our friendship. Mult'umesc foarte mult.

AMOLF is a big family: I want to thank all the overloop dwellers, whose offices were always open to my bugging: science, procrastination (is there a neat boundary between the two?), organisation of events, and funny stories to cheer each other up when things did not work were the subjects of our conversations. Everyone played a role in the AMOLF comedy, and it would be too long to fill the wall of fame with all those that have left in last years. I will quickly mention Iorgos "Gino" Boulougouris and his generosity and humanity, Josep "Rocco" Pamies Corominas and his shoes, Fabrizio Capuani and his family, Ivan Coluzza and his hospitality, Marco Cosentino Lagomarsino and his style, Donna Beatrice Marino and her good heart, Fabiana Diotallevai and her laziness (grazie per la bella esperienza di semiconvivenza, che mi porterò sempre dentro), Olga Katsibiri and her smiles, Behnaz Bozorgui and her (and her mother's) food, Simon Tindemans and his spirit (bedankt voor het samenvatting!), Christian Tischer and his face, Tatiana Schmatko and her frenchness, Franca Fraternali and her sunny moods, Frank Poelwijk and Laura Munteanu who prevented me feeling alone when I was staying late, Bianca Mladek and her wordiness and gentleness, Yuri and Nadya and their generosity, and all the others. A special mention to the persons who were so unlucky to share the office with me, being therefore exposed to all my mood swings, italian conversations, allergies to drafts and temperatures below 25 degrees: Rhoda Hawkins and Chantal Valeriani. Quasi cinque anni di convivenza negli stessi dieci metri quadri non si scordano facilmente. Se siamo arrivati insieme a questo traguardo, lo dobbiamo anche alle nostre chiacchierate, al nostro insultarci gentile e ai nostri scherzi. Visto che ancora parliamo e siamo in buoni rapporti sinceri, deduco che, alla fine, abbiamo fatto un buon lavoro.

Outside AMOLF, I had the luck to get to know and collaborate with the VU gang, who always seemed to me an energetic and outgoing bunch of smart guys. In particular, I had an extremely fruitful and enjoyable collaboration with Dr. Jeroen van Zon, who extended far beyond office hours and eventually ended in a stable friendship, which I hope it can be the base for future scientific collaborations. Among the others, I should mention Dr. Bram van den Broek, who welcomed me

during my interview five years ago, brought me for the first time in a café in town, and kept in touch during my whole PhD, gratifying me with his appreciation of my italian cuisine.

Vivere in terra straniera è interessante e stimolante, ma talvolta non facile. Poter contare su un gruppo di persone con cui poter ricreare l'aria di casa è un grosso regalo per un emigrante. Ringrazio quindi la comunità italiana di Amsterdam che mi ha accolto e con cui ho passato serate piacevoli e divertenti. Vorrei ringraziare in particolare i miei colleghi vecchi e nuovi, Enrico Conti e Andrea Baldi e la sua famiglia.

Living in Amsterdam is a unique opportunity to get to know people from all over the world, and few times random encounters tenaciously evolve towards friendships. I want to thank all the people that decided to go beyond the standard formal conversations I quickly get annoyed about, took their risk by uncovering their hidden sides, and invested time into building a high-quality relationship with me: Salina, het begon allemaal met een pastacompetitie in Groningen en we hebben later in Amsterdam samen een hoop lol gehad, Lynn, for her oper heart and mind, and her invitation to Kolkata, Miriam, una chica pequeña y mediterránea con quién compartí una hermosa vacación y muchas otras emociones, Eve and Magdalene, that taught me once more one can be an intelligent and sensitive person without being too serious, and happy without any money, Maryam, for her uncontenible enthousiasm towards any new challenge in Life, and for her compassion during the disease of my father, Eduardo, para su amistad y su actitud fiestera, Flora pour sa patience avec mes retards et toutes les soirées passées au cinéma, regardant des films que parfois aucun autre aurait supporté.

The most memorable experience during my PhD was surely spending a total of three months in Scotland, working with Dr. Rosalind Allen at the University of Edinburgh, within the High Performance Computing european framework. I always loved to bring fresh new air into my lungs after having breathed the same one for long tim: when I realized it came from wild mountains crashing into the sea, my spirit had a jump, and I started loving that land. My stay became even more pleasant when I got to know the people in room 4305, and those who were living in Warrender Park Crescent: with them, I had the occasion to explore a new world and have interesting conversations during meals. I want to thank Gladys and her enthousiasm for nightlife: I am happy the exchange of music and films led to a stable friendship–thanks for the hospitality in Barcelona! The flawless organisation of the stay was possible only because of the work of Catherine Inglis, who, besides her brilliant administrative work, gave me tips and tricks to visit her country. Davide Marenduzzo was extremely helpful in the construction of the model of DNA looping in the phage λ system. I have great memories of music fridays held at Kate's place (thanks thanks!), of italian cooking sessions with Lucio, Federica, Maria, Giacomo and Simone, and of fun time with Rowan, Daniëlle and the others. Finally, this Thesis would probably have never seen an end without the help of Susie, who, after having given me a warm welcome in her land and fascinated me with stories of wilderness, made me think a lot of her views on life and human relationship, and listened to my endless, repeated rants during my writing time. Móran taing, leannan!

One of the reasons I chose this job was the opportunity to couple it with my passion for travelling and getting to know new people. During my PhD, I had the occasion to attend several conferences and summerschools, and I was so lucky I managed to bring back home from each of them, besides the scientific enrichment, some firm friendships. I wish to thank the people who are still in touch with me, and have followed the evolution of my PhD: Olienka, gracias por los momentos que pasamos juntos en Montreal y por la hospitaliad en Alicante, Pia, for her restless enthousiasm, her patience and her true feelings, Maria for her company in Denmark, and

the solar spirit she irradiated, Matthias for his sympathy and gentleness, and for hooking me up with Giovanna. Thanks to the fellows who were with me in Boulder and allowed me to have so many fantastic warm nights and wonderful hikes in the Rocky Mountains: tough and sweet Corinne, Peter, Megan (especially when drunk), Martin, Thierry, Aleksandra ("I don't need to be cool anymore"–I really wish we will work together once). A very special thank to Katarina, I don't know what I would have done without you when my father passed; I will treasure your sympathy forever: hvala lepa, ljubim te puno.

Prima di partire, i miei primi 25 anni mi hanno regalato persone che hanno saputo essermi così fedeli da non perdermi di vista quando sono emigrato. Non mi stancherò mai di ringraziarli per il loro sforzo volto ad incontrarmi nelle mie (sempre troppo) brevi permanenze sul suolo italico. Crescere insieme cementa le amicizie oltre ogni immaginazione: con Marcello, Valerio, Piermauro (e la sua famiglia), oltre alle sincere chiacchierate, siamo persino riusciti a creare il progetto Orobiehiking e il suo sito web: speriamo che siano rose, e che fioriscano. Dalla sella della sua bici, Ombretta non mi ha mai lasciato andare in fuga, ma mi ha sempre ripreso quando mi allontanavo. Altre volte, è il destino ad intervernire: dopo l'incontro di Bologna, so che non mi potrò mai liberare della famiglia Bartolucci; e meno male.

Anche la mia avventura universitaria e collegiale a Pavia mi ha portato a conoscere persone con le quali un sorriso e una pacca sulla spalla basta ad aggiornarsi dopo anni di silenzio: Ciccio, che mi ha ospitato in Zambia e mi viene a trovare fedelmente in ogni angolo del mondo la scienza mi porti, gli amicidelparadiso e tutte le loro amorose, che Dio li benedica tutti! Danila e Paola, la vicinanza del cuore e la poesia nelle parole di Chiara, la sorpresa di riincontrare persone come Anna, rivangare un'amicizia e scoprire che lavorare ad un progetto comune può essere incredibilmente stimolante. Il fatto che io e Linda ancora ci parliamo e viviamo esperienze ed emozioni insieme è poi qualcosa di completamente inspiegabile per qualsiasi mente razionale. Forse siamo semplicemente "strani", come tutti ci hanno sempre considerati. Comunque, a lei il mio grazie più grande perché c'è sempre stata, e ho sempre saputo di poter contare su di lei.

Infine, un grazie alla mia Famiglia, che, diversa, composita, avvolgente e bergamasca, non si può scrivere senza la lettera maiuscola. La nuvola di affetto nella quale mi sono sempre sentito circondato al mio ritorno, il rispetto per scelte e capigliature che non hanno riscontri nel resto del parentado, e la commovente vicinanza e disponibilità nei giorni più difficili degli ultimi tempi, mi hanno fatto rendere conto di quanto sia fortunato, e di quanto il nostro clan sia diverso dalle famiglie nucleari moderne. Un grazie quindi a mamma Franca e a papà Luigi, e a tutta la massa di zii, cugini e parenti.

Alla soglia dei trent'anni, mi sento di rinnovare il messaggio scritto cinque anni fa alla fine della mia tesi di Laurea: un grazie a coloro che ancora credono nei sogni, perché mi hanno sempre convinto che non esistono scelte obbligate nella nostre vite, se abbiamo la forza di buttarci e cambiare; e alle montagne, che con la loro severità continuano ad insegnarci il significato dei sacrifici e delle gioie vere.

Curriculum Vitae

Marco Jacopo Morelli was born in the beautiful town of Bergamo, in Northern Italy, on the border of the Alps, on the 14th of February, 1978. He lived for the first 19 years of his life in his hometown, where he attended the vibrant Liceo Scientifico Filippo Lussana, and he got his high school Diploma with top grade. The relative absence of important travelling during his childhood endowed him with firm roots, the love for mountains, and the hard bergamask head.

In 1997 he moved to the university town of Pavia, in the middle of the Po Valley, to study Physics. He scored first at the competition to enter the ancient Almo Collegio Borromeo (a center of excellence established in 1564), where he stayed for 4 years, with other 90 fellow students. The college life profoundly influenced his views on friendship and human relationships. In 1998 he joined the Istituto Universitario di Studi Superiori (IUSS), being selected together with 40 other pupils of his age to attend extra classes, aimed to broaden his views on science beyond physics. It was during the IUSS course held by Prof. Piero Cammarano in 2000 that he first got in touch with biology. He graduated in Physics in 2002, *summa cum laude*, with a Masters thesis on the use of neural networks to solve the Black-Scholes-Merton model for financial derivatives. He received his IUSS Diploma the same year.

After a brief experience as a PhD student at the University of Pavia, in 2003 he decided to become an expatriat and dedicate all his scientific efforts to biophysics, joining the group of Dr. Pieter Rein ten Wolde at the Institute for Atomic and Molecular Physics (AMOLF) in Amsterdam. As a PhD student there, he enjoyed the international atmosphere of the city and the research on genetic networks. He attended summerschools in Canada, Denmark and USA, where he had the occasion of getting in touch with smart and nice people. In 2006 and 2007, he was awarded twice an High Performance Computing travel grant to join the group of Dr. Rosalind Allen in Edinburgh, where he had a fantastic scientific, social and outdoor experience. In 2006, he entered the Board of Directors of the IUSS alumni association in Pavia, that he co-founded, and in 2007 he started the Orobiehiking project, aimed to increase the presence of foreign tourists in the Orobian Alps near Bergamo.