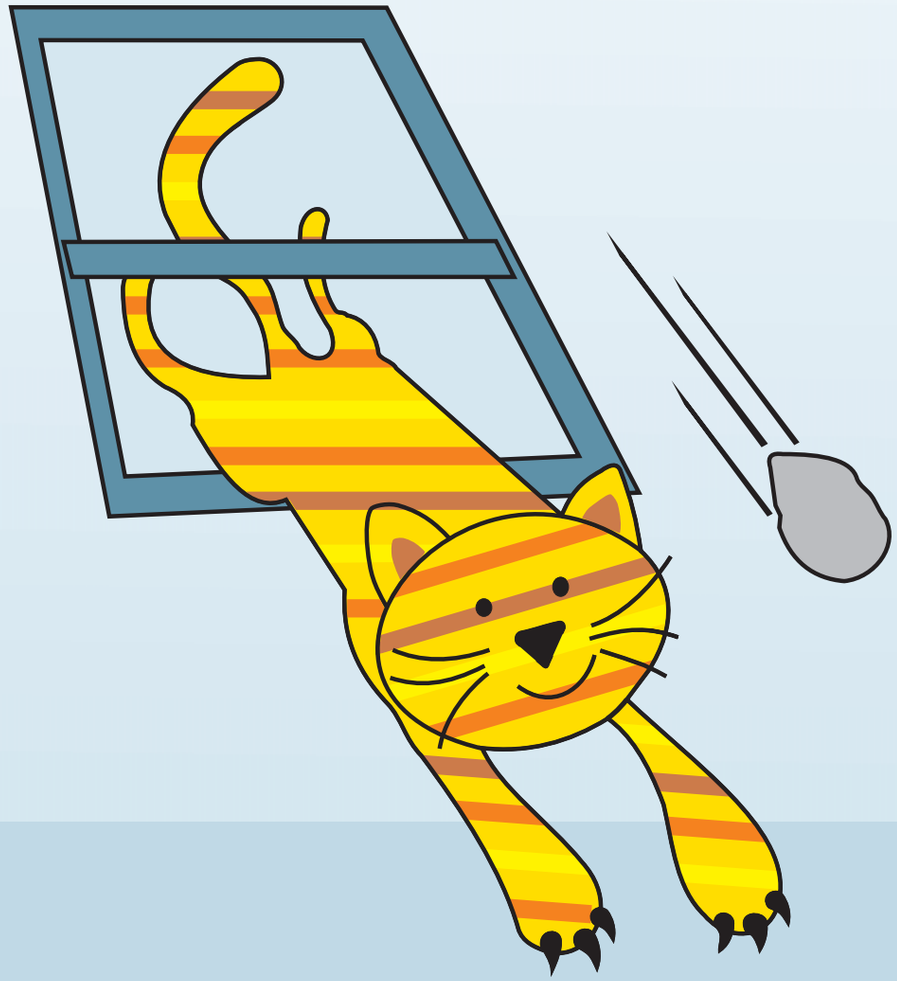
Molecules in Motion

a theoretical study of noise in gene expression and cell signaling



Maciej Dobrzyński

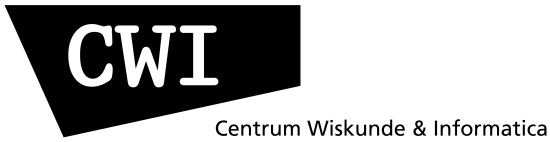
Molecules in Motion Maciej Dobrzyński

Molecules in Motion

a theoretical study of noise in gene expression and cell signaling

Maciej Dobrzyński

The research described in this thesis has been supported by the Netherlands Organization for Scientific Research (NWO) under project number 635.100.007. It was carried out at the Centrum Wiskunde & Informatica (CWI), the Dutch national research institute for mathematics and computer science, in Amsterdam.



Thomas Stieltjes Institute
for Mathematics

Molecules in Motion: a theoretical study
of noise in gene expression and cell signaling
Thesis, University of Amsterdam

Typeset in LaTeX
Layout template by Olivier Commowick
http://olivier.commowick.org/

Printing by ARTWORK, Grafika i Reklama
http://www.artwork.pl

Cover design by Alina Orchowicz

Molecules in Motion

a theoretical study of noise in gene expression and cell signaling

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. Dymph van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op donderdag 13 januari 2011, te 10.00 uur

door

Maciej Dobrzyński

geboren te Warschau, Polen

Promotiecommissie:

Promotor:	prof. dr. J. G. Verwer
Promotor:	prof. dr. H. V. Westerhoff
Overige leden:	Prof. dr. R. van Driel
	Prof. dr. M. R. H. Mandjes
	Prof. dr. P. M. A. Sloot
	Prof. dr. P. R. ten Wolde
	Dr. N. Blüthgen
	Dr. F. J. Bruggeman
	Dr. J. A. Kaandorp
Faculteit:	Natuurwetenschappen, Wiskunde en Informatica

Acknowledgments

I would like to express my deepest gratitude to Joke Blom, Frank Bruggeman and Jan Verwer for the inspiration they gave me, their time spent on long discussions, patience and always constructive critique. I also would like to thank Witold Rudnicki and Pieter Rein ten Wolde for introducing me to the field of computational biology and stochastic effects in biochemical networks.

Further, I would like to thank all my family and friends for support and encouragement, Jordi for our endless discussions, my wife Justyna for feeding me during the lengthy process of writing, Alina for designing the "falling-cat" cover, Yves for keeping me fit, and my office-mates Svetlana, Peter, Antonios, Danijela, Bernard and all other MAS group members for a great time we shared at CWI. Above all, I thank my parents Jadwiga & Sylwester, and my aunt Maryla for being there.

Thank you all!

Contents

1 Introduction **1**
- 1.1 Stochasticity fundamentals . . . 4
- 1.2 First example . . . 11
  - 1.2.1 Physiological implications of bursts . . . 13
- 1.3 Timing of biochemical reactions . . . 17
  - 1.3.1 Sequential processes . . . 17
  - 1.3.2 Diffusion-limited reactions . . . 20
- 1.4 Organization of the thesis . . . 21

2 Bursty transcription and translation **25**
- 2.1 Abstract . . . 25
- 2.2 Introduction . . . 26
- 2.3 Results . . . 27
  - 2.3.1 Analytical expression of the waiting time distribution . . . 27
  - 2.3.2 Measures for characterization of bursts . . . 29
  - 2.3.3 Motor-protein traffic jams along biopolymer chains . . . 31
  - 2.3.4 Pausing of motor proteins can generate bursts . . . 33
  - 2.3.5 Aggregative behavior of multiple burst-generators . . . 35
- 2.4 Discussion . . . 36
- 2.5 Materials and methods . . . 38
  - 2.5.1 Statistics of the arrival process . . . 38
  - 2.5.2 The limit of a large time scale separation . . . 40
  - 2.5.3 Moments of the first-passage time *pdf* . . . 40
  - 2.5.4 Quantitative characterization of bursts . . . 41
  - 2.5.5 Non-exponential waiting time distribution for the switch . . . 46
  - 2.5.6 Interarrival time CDF in a pool of unsynchronized IPPs . . . 48
  - 2.5.7 Progression of motor proteins along the polymer . . . 48

3 Swift and robust response in two-component signaling **53**
- 3.1 Abstract . . . 54
- 3.2 Introduction . . . 54
- 3.3 Results . . . 55
  - 3.3.1 Approximating the search of a single sensor by a first-order process 55
  - 3.3.2 The promoter search can be approximated by a first-order process too . . . 57
  - 3.3.3 Clustering of sensors affects the search time for sensors . . . 58
  - 3.3.4 Two-component signaling-induced gene activation proves swift, robust and efficient . . . 59
  - 3.3.5 Contribution of diffusion and molecule copy number noise to noise in response time . . . 60
  - 3.3.6 Demand for fast and robust signaling can constrain operon organization . . . 62

3.4 Discussion . . . 65
3.5 Materials and Methods . . . 67
3.5.1 Evaluation for 3D: searching the inner sphere . . . 68
3.5.2 Evaluation for 3D: searching the target on the membrane . . . 69
3.5.3 Round-trip: convolution . . . 71
3.5.4 Moments of the first order statistics of a convoluted *pdf* . . . 72
3.5.5 Bias of the optimal search time . . . 72
3.5.6 Time to (de-)activate half of response regulators . . . 73
3.5.7 Diffusive flux at the promoter site . . . 73
3.5.8 Relation between the number of regulators and the number of DNA-binding sites . . . 74
3.5.9 Two sources of stochasticity . . . 76
3.5.10 Processivity . . . 77
3.5.11 Numerical simulations . . . 78
3.5.12 Bioinformatic analysis . . . 78

4 Methods for diffusion-limited problems 81
4.1 Abstract . . . 81
4.2 Introduction . . . 82
4.3 Regimes and models in biochemistry . . . 83
4.4 Test cases . . . 85
4.4.1 Gene expression . . . 85
4.4.2 Signal transduction . . . 87
4.5 Computational methods . . . 88
4.5.1 BD-level . . . 88
4.5.2 RDME-level . . . 89
4.5.3 CME-based . . . 90
4.6 Results . . . 90
4.6.1 Gene expression . . . 90
4.6.2 Dynamics of gene expression . . . 90
4.6.3 Reversible reaction of an isolated pair . . . 91
4.6.4 CheY diffusion . . . 93
4.6.5 Dynamics of CheY diffusion . . . 94
4.7 Discussion . . . 96

5 Discussion 101

A First-passage time basics 111
A.1 *pdf* and CDF . . . 111
A.2 Moments of the first-passage . . . 112
A.3 Moments and the survival probability . . . 112
A.4 Superposition of N independent random processes . . . 112
A.4.1 First order statistics . . . 113
A.4.2 k-th order statistics . . . 114
A.4.3 Excess lifetime . . . 114
A.4.4 Interarrival time CDF in an unsynchronized ensemble . . . 116

B Basics of diffusion-limited reactions **117**
B.1 Theory: single molecule 117
B.1.1 Incorporating reactions through boundary conditions 117
B.1.2 Solution for the spherically symmetric system 118
B.1.3 First-passage time 119
B.1.4 First-passage related to flux 119
B.1.5 Evaluation for 1D 120

C Mesoscopic models and computational methods **121**
C.1 Mesoscopic models 121
C.2 Computational methods 125

Bibliography **131**

Summary **147**

Samenvatting **151**

Dla laików **155**

Chapter 1

Introduction

Contents

1.1 Stochasticity fundamentals 4
1.2 First example 11
1.2.1 Physiological implications of bursts 13
1.3 Timing of biochemical reactions 17
1.3.1 Sequential processes 17
1.3.2 Diffusion-limited reactions 20
1.4 Organization of the thesis 21

Throw a pebble and... a cat from the window and you will hear just one thump in the courtyard, possibly preceded by a squeaking meow. Theoretically. The air friction will most likely slow down the furry pet a little while it desperately attempts to adjust the position before the inevitable. Prior to discovering the celebrated relation for the motion of a mass in a gravitational field one needs to consider a highly abstracted version of the problem. An easy procedure for an apt high school student in the XXI century, but a real challenge for those unaware of Newton's *Principia*. Turning the stone and the cat into point masses and letting them enjoy the fall in a vacuum, should do the trick and reveal that neither the object's mass, nor its shape affect the motion, the fact expressed by Galileo nearly 400 years ago: *All objects fall at the same rate in a vacuum.*

There are at least three conclusions to infer from this example. First, an animal, though very different in our perception from a dull stone, belongs to the same realm where the fundamental laws of physics apply. Quantum theory, currently the most fundamental description of matter, is as capable of explaining the formation of a covalent bonding of carbon atoms in a diamond lattice as the properties of a hydrogen bond to which nucleic acids and proteins owe their three-dimensional structure.

Secondly, questions about physical processes can be successfully addressed even with theories that inherently have limited scopes. While certainly required for describing phenomena on atomic scales, quantum mechanics (QM) looses its appeal when applied to a large collection of particles such as a stone or a cat. Although an architect designing a bridge might indirectly profit from QM through advancements in material science, he will certainly succeed without being aware of all the intricacies of the Schrödinger equation and the Hilbert space. Classical Newtonian mechanics, laws of which follow from the laws of QM in the limit of macroscopic systems, is sufficient to describe tensions and loads of a massive steel construction and the falling of cats.

The final observation regards inferring the underlying theoretical framework. When skimming through an introductory physics course book, students often come to a (hopefully superficial) conclusion that physics has actually very little to do with the real world.

Point masses, frictionless motion, approximations even to seemingly simple problems (physical pendulum, three-body problem) do not have an immediate counterpart in reality. Nonetheless, a significant idealization of a problem eases the discovery of fundamental laws. A movement of a block involves friction, heat transfer, possible deformations. In most of the cases, however, a good approximation can be obtained by omitting those effects and trading them for frictionless dynamics of point masses. Naturally, a model is realistic as long as it provides testable predictions but more importantly the idealized description allows to discover a more general framework.

Formulation of classical mechanics using the Lagrangian formalism, although applicable to a narrow class of systems without an approximation, gives insights into the structure of the physics theory. For instance, the law of conservation of energy is a very deep and far-reaching principle present in such diverse physical phenomena as electromagnetic radiation, general relativity or quantum mechanics. A relatively straightforward algebra reveals another deep result – the principle of least action. In fact, laws of motion of particles in gravitational fields, laws of electricity and magnetism, motion of particles in electric fields, all have a common basis: the principle of least action. Laws of thermodynamics are simply statistical claims about large numbers of degrees of freedom, given that those degrees of freedom are governed by the underlying principle of least action.

Although numerous details are deemed unimportant, simple models give particularly valuable insights into the problem; especially if the mathematical representation is tractable enough to allow for an informative and closed solution. Unfortunately, the amount of analytically solvable models in physics is limited. The situation in biological sciences seems even grimmer and the reasons for this are the following. For generations natural sciences have been dominated by strong belief in the power of reductionism. Rightly to be so. A long struggle to discover the nature of forces holding matter together resulted in a solid theoretical framework followed by a surge of then unthinkable applications. Parallel to these ground-breaking discoveries biologists continued to be faced with a multitude of new species, forms, animal behaviors and habits. A profound contribution of Darwin cleared the picture, although being developed almost the same year as Maxwell's equations (unfortunately) still continues to be the subject of many heated debates. Only dawn of molecular biology and genetics in the twentieth century shed new light on mechanisms reasoned a hundred years earlier. But then the problem became even more clear. The quest to reveal fundamental principles in biology has become marred by a staggering number of components constituting a living organism, the degree of interaction between the elements and finally the huge number of states exhibited by such an interacting network. Even a simple single-cell organism involves processes spanning many orders of magnitude: from nanosecond timing of structural motion of bio-polymers, to molecular clocks ticking at night-and-day intervals, or month or even year-long-lasting states of biochemical switches. All of these factors make it difficult to view the system as a collection of independent entities. But living organisms are subject to the laws of physics. Is it then possible to explain their design from the first principles using the same tools as in physics? Is it reasonable to assume that once we understand all the fundamental laws governing micro-world, the macroscopic description will follow [Binder 2009]?

In order to survive in the environmental niche, even the smallest organism has to reliably pass genetic information to the progeny, a multitude of intracellular processes needs to be coordinated in order to proliferate, metabolize, respond to changes in nutrient con-

centration or to sudden appearance of toxic compounds. On the other hand, colonization or adaptation to new environmental conditions requires changes in the phenotype accessible only through genetic mutations, which give rise to new metabolic, sensory or physical features. Some of the survival strategies involve genotypes capable of generating various phenotypes optimal for different environments. Various regulatory mechanisms exploit noise to randomize outcomes where variability is advantageous [McAdams & Arkin 1999]. In order to achieve all of these goals, regulation of biochemical processes evolved such that the organism may profit from retaining stochasticity inherent in biochemical reactions, while other structures evolve to suppress the noise which would disrupt precision of cellular physiology.

Thermal fluctuations may easily alter the rates of chemical reactions, local concentrations of reactants or the composition of genetic code [Dronamraju 1999]. In a number of experiments, isogenic populations starting from the same initial conditions have been propagated into cells with entirely different molecular makeup [Elowitz *et al.* 2002, Blake *et al.* 2006, Spencer *et al.* 2009] or proliferated into phenotypically distinct cellular entities with diversified biological functions [Balaban *et al.* 2004, Feinerman *et al.* 2008, Chang *et al.* 2008] – a convincing demonstration of how random molecular events may affect the macroscopic observable: the phenotype.

If molecular fluctuations are so abundant, it is truly remarkable that seemingly fragile molecular structures inside living systems reliably hold genetic information even for thousands of years, perform uncountable enzymatic reactions, protect the organisms against the influence of harsh environments and give rise to countless life forms, shapes and behaviors. As might be expected, the composition (or design) of an organism successfully populating a niche reflects the two tendencies mentioned earlier: flexibility allowing for acquisition of new features and robustness facilitating reliable information processing. For instance, sensing changes in the nutrient concentration in the environment of a simple bacterial cell is prone to uncertainties due to thermal noise. Attaining a level of signal-to-noise ratio that would allow for appropriate and possibly rapid cellular response requires a specific architecture of molecular structures (a sensor) and corresponding biochemical reactions (a signaling and a regulatory network). On top of that, the number of biomolecules involved in the response has to be such that the network itself is not activated by a sudden fluctuation in the concentration of reactants.

Thus, recognizing how organisms may benefit from molecular fluctuations and when these stochastic effects are suppressed gives valuable insights into how fundamental physical constraints shape physiology of biological systems in the course of evolution. For instance, fluctuations are attenuated through the structure of biochemical networks (e.g. cascade architecture of eukaryotic signaling networks [Thattai & van Oudenaarden 2002, Hooshangi *et al.* 2005, Bruggeman *et al.* 2009]) or through properties of the stochastic system itself (e.g. sequential processes described in more detail further in this chapter). The evolution of structures capable of reducing molecular fluctuations confers fitness advantage for the organism. The design of these structures such as the number of levels in the cascade or steps in the sequence, the number of molecules involved in the process reflects the direction that the evolutionary process has taken in order to overcome physical constraints in processing of the extracellular signal or limitations in the speed of chemical reactions.

Manifestation of stochastic phenomena in cellular processes has some practical im-

plications for the modeling community too. Analysis of problems involving stochasticity requires a more complex theoretical and computational apparatus than deterministic processes. Therefore, it is of great importance to recognize deterministic regimes in biochemical processes. Typically, for a large number of molecules the significance of molecular fluctuations diminishes. The complexity of theoretical description reduces dramatically and so does the computational cost. Large time scale separation confers a similar advantage: very fast fluctuations may be averaged out by much slower downstream processes effectively alleviating the need for a detailed stochastic description of the fast module.

The exposition in the following sections will focus on stochastic phenomena in a single cell. To address that, we shall introduce the framework for the analysis of stochastic events, namely the waiting-time or the first-passage time theory. We will draw parallels to familiar macroscopic approaches by illustrating the limits of the stochastic theory for a large number of molecules. By analyzing the effect of various waiting-time statistics in protein synthesis and degradation we shall gain insights into the source of precise timing in stochastic systems as well as the reason why a deterministic approximation works so well in theoretical biology. In particular, we shall discuss the exponential regime of the waiting times which allows for significant simplification of sequential and diffusion-limited biochemical processes. The exponential approximation contributes to a largely unexplored topic of the interface between stochastic and deterministic theoretical modeling in biology.

1.1 Stochasticity fundamentals

Biochemical transitions typically of interest for molecular systems biology involve macromolecules sized above $5\,nm$; a so called mesoscopic level. Confront that with $\approx 0.3\,nm$ – the effective diameter of a water molecule or an ion. An *E.coli*, a model prokaryotic organism contains $2,5 \times 10^{10}$ (25 billion!) water molecules and as little as ≈ 4 million proteins in a $2\,\mu m$-long ellipsoid cell. An enormous disparity of time scales of atomic interactions between small molecules (e.g. water, ions) and large bio-molecules (e.g. proteins, nucleic acid chains, membrane lipids) permits statistical claims about the effective forces exerted on the latter. Hence, the movement of biomolecules can be described much simpler by introducing a stochastic term accounting for random collisions of a protein, for instance, with much smaller molecules of the solvent. The complexity of the resulting equations of motion (Einstein-Smoluchowski diffusion equation) is reduced dramatically compared to the initial description accounting for the motion of all molecules, the solvent and the solute. If diffusive encounters of molecules are much faster than the time required to undergo a chemical transition, the spatial aspect of the problem can be discarded and an even simpler model materializes. The complication remains, however. Thermal fluctuations affecting the rates of chemical reactions remain in the form of probabilistic terms similar to those in the synthesis-degradation problem. The effect of fluctuations becomes particularly strong if the amount of molecules involved in a reaction is *small*. Here lies a difficulty of stochastic modeling. The boundary between a system where fluctuations in the species concentration are significant and a system in the macroscopic regime is not immediately apparent. It depends on relative timescales of processes in a biochemical network and the topology of the network, among other factors. Some of these issues we explore in Chapter 3 where we analyze a generic bacterial signaling network. The number of molecules sufficient to overcome diffusion limitation and to assure signal responses

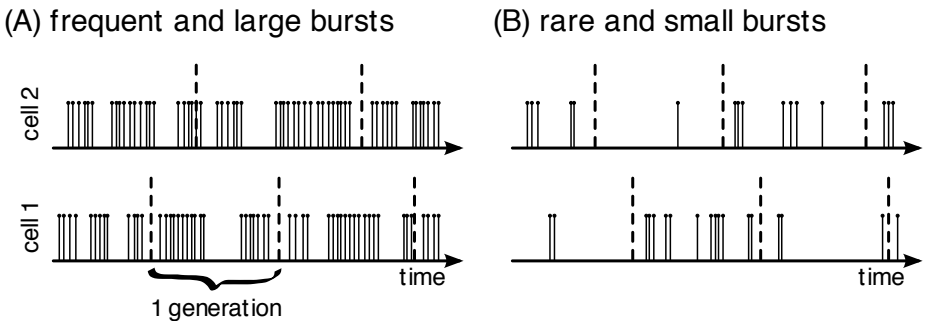


Figure 1.1: Stochastic protein synthesis may leave individual cells with very few or none of the product if bursts in protein synthesis are infrequent and small within a cell's generation. For the experimental evidence of this effect in *E.coli* see [Yu *et al.* 2006].

without large variability turns out to be around a few dozen.

Accounting for stochastic effects in biochemical processes requires an entirely different take compared to a much more familiar macroscopic description where variables correspond to the amount (or the concentration) of chemical species. Since a fluctuating variable takes on many values with various probabilities, the theoretical description should not only allow for computation of the mean but also account for the second moment. A complete description should even track the evolution of the whole probability distribution, which prescribes how frequently values of the random variable (admissible by a particular random process) occur. As we shall demonstrate further, the shape of the steady-state probability distribution of the level of a biochemical product, a protein for instance, can reveal valuable information about the underlying stochastic mechanism. It may also imply physiologically meaningful behavior of cells.

In the recent decade in a number of settings, experimentalists have been able to measure distributions of protein concentration taken as snapshots over entire cell population [Elowitz *et al.* 2002] as well as temporal changes of a fluctuating protein. Theoretical arguments have been made regarding the equivalence (or lack thereof) between the two averages [Tănase-Nicola & ten Wolde 2008]. According to the ergodic principle, these two should be equivalent. However, the nature of the cellular system itself may prevent the evolution of all degrees of freedom from reaching all possible states in the lifetime of the organism. This may happen if fluctuations are slow (Fig. 1.1). Hence, not only the stochastic model requires one to study the whole spectrum of fluctuations instead of the mean, but also the knowledge of statistics of individual events is required to understand the physics of intracellular processes. Since it is difficult to infer about the physiology of the single cell based on statistical claims about population averages, we shall extend the analysis of cellular stochastic processes with the study of temporal properties of random events. We will scrutinize the steady-state distribution of waiting times between biochemical events. Owing to advancements in single-cell, single-molecule experimental techniques these type of theoretical studies are increasingly gaining solid experimental evidence.

There exist many equivalent approaches to model stochasticity at the mesoscopic level (e.g. master, Langevin, Fokker-Planck equations). For reasons explained earlier, one particularly didactic way of looking at fluctuations is by considering waiting times. A generic

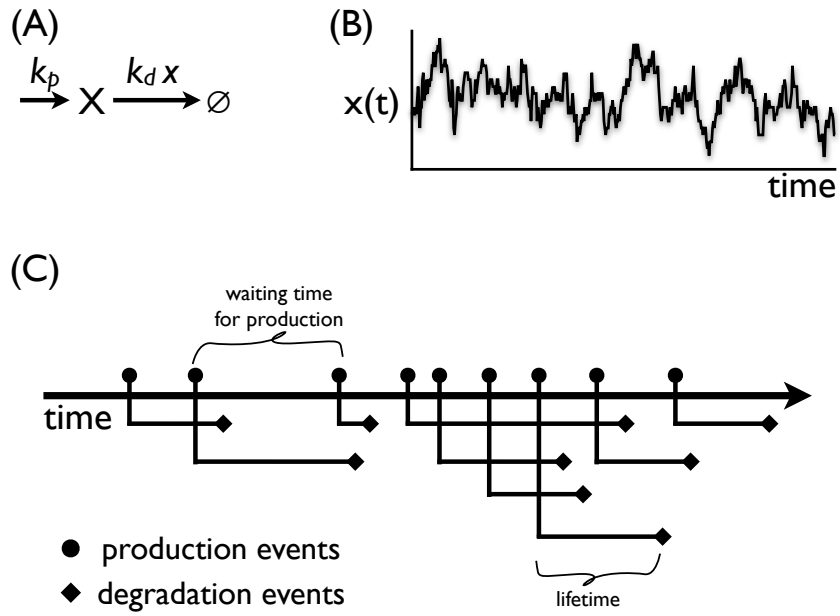


Figure 1.2: Stochastic production-degradation model. (A) Schematic representation of the model. Chemical species X is synthesized at the rate k_p and degraded at the rate $k_d \cdot x$, where x denotes the number of X molecules. The inverse of parameters k has an intuitive interpretation; $1/k_p$ is the mean interval between production events, $1/k_d$ is the mean lifetime of a single X molecule. (B) Sample stochastic time trajectory. Synthesis and degradation events are drawn from the exponential waiting time distribution. (C) Stochastic birth and death of individual molecules X.

birth-and-death model depicted in Fig. 1.2A includes only two stochastic processes. On average, the synthesis generates k_p elements of X per unit time. X can represent a species of a bio-molecule like a protein or a messenger RNA, for instance. In the latter of the two reactions every molecule X undergoes degradation. The average lifetime of such a molecule (the time between its synthesis and degradation) equals $\tau_d = 1/k_d$ time units.

If both, synthesis and degradation were taking place at exactly equal instances of time, the amount of molecules X, denoted by lower-case x, would remain constant at any given time (Fig. 1.3A). An instructive analogy would be a queue at the cash register. If a next customer joins (*synthesis*) the waiting people precisely at the moment the first person finishes paying (*degradation*), the length of the queue will not change. Unlike this deterministic situation, intervals between subsequent synthesis and degradation events are not fixed in a chemical system (nor in a real-life queue on a busy day in the supermarket). Precise timing of events is heavily influenced by thermal fluctuations affecting the molecules' relative position which in turn influences their propensity to overcome the free energy barrier and undergo a chemical reaction.

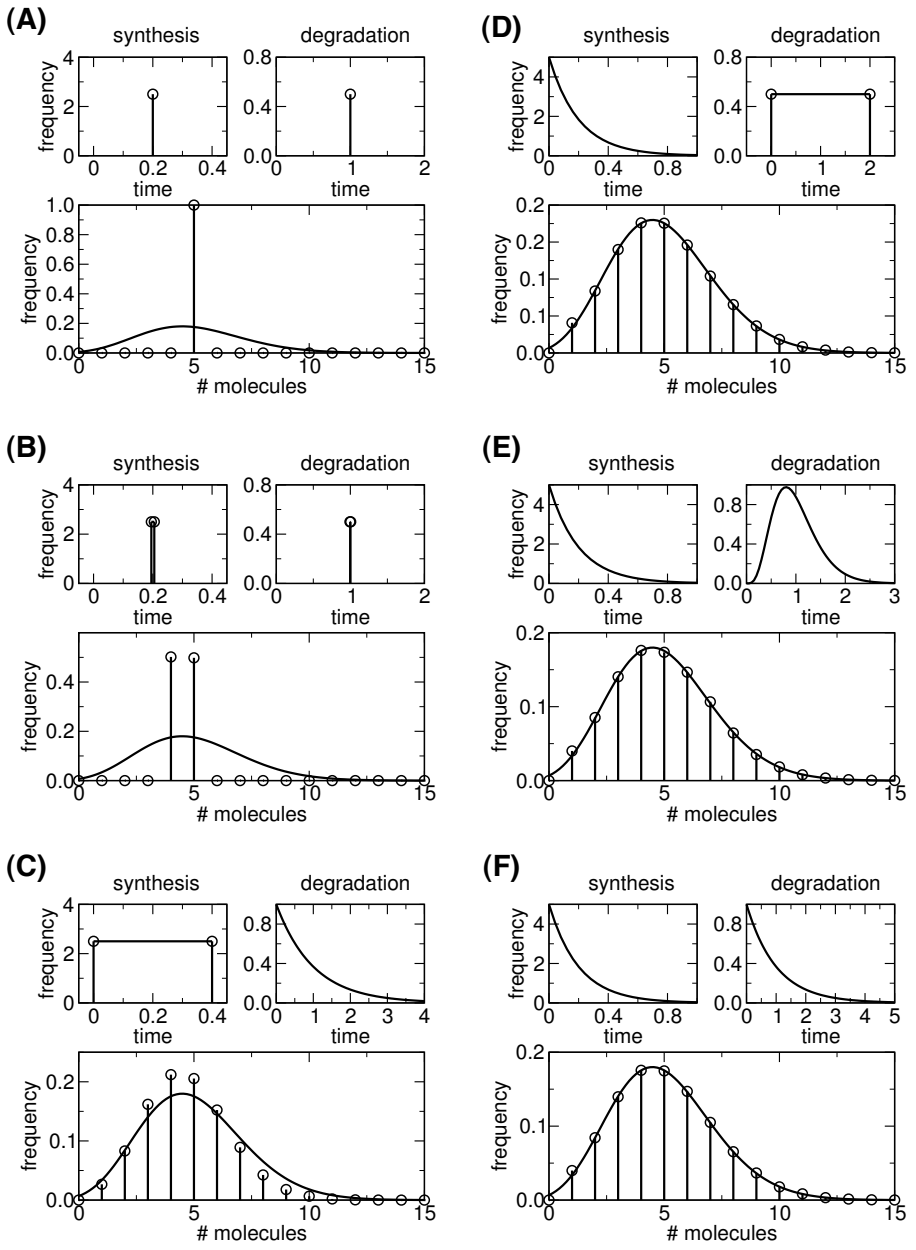


Figure 1.3: The effect of waiting time statistics on steady-state product distribution in a linear birth and death process. (A-F) Every panel consists of three subplots: the top-left contains a normalized histogram of intervals (waiting time distribution) in the synthesis process with the mean of 5 events per unit time, the top-right contains an analogous histogram for the life-time of a single molecule X – on average 1 event per unit time. The bottom graph depicts a normalized histogram of the amount of X in steady-state (steady-state X distribution). The solid curve in every panel depict the Poisson distribution – the distribution resulting from exponentially distributed synthesis and degradation processes (panel F).

Since the length of intervals between synthesis events and the lifetime of single molecules fluctuate, the steady-state number of X will change in time accordingly. Figs 1.2B and C illustrate a sample time-course. But fluctuations of waiting times may differ in origin and hence may have a different statistic. Processes such as mRNA or protein degradation, assembly of protein initiation complex, enzymatic reactions, diffusion, they all involve a series of steps to complete. While the average behavior can be conveniently characterized by a model where the sequence of events is aggregated into a single step with an effective mean completion time, higher moments are sensitive to the model's granularity. This is best illustrated in Fig. 1.3 where the steady-state X distribution has the same mean regardless of the production and the synthesis statistics, however its standard deviation varies significantly. We shall see more examples of this in sections below.

How does the statistics of waiting times for synthesis and degradation events affect the magnitude and frequency of steady-state fluctuations of X? As explained earlier, synchronized additions and subtractions of the product leave the level of X unchanged (Fig. 1.3A). The moment the synthesis intervals are no longer equally-timed, the degradation does not coincide with the production and the level of X fluctuates in the course of time. The system becomes stochastic. Instead of a single value as in the deterministic case, x may take a whole range with different probabilities. If the amount of X synthesized balances the amount of X degraded during the same period, the system is said to be in steady-state. Then, the distribution of x can be plotted in a straightforward manner by creating a normalized histogram from equally-timed measurements of x. Clearly, along with the increased randomness of waiting times, the amount of states reached by x rises and the steady-state product distribution widens (Fig. 1.3A-C). One waiting-time distribution is particularly privileged, however. Intervals between synthesis events drawn from the exponential function $(k_p \exp(-k_p t))$ result in the steady-state product distribution independent of the type of the waiting-time distribution for the degradation (Fig. 1.3D-F)! This phenomenon is known in queueing theory as *insensitivity*. Moreover, for all of these cases the steady-state X distribution is Poissonian – the distribution being a seminal result for the birth-and-death model with exponential waiting times.

The special status of the exponential waiting-time distribution becomes clear if one attempts to characterize a stochastic process mathematically. A consistent description resulting in the evolution equation – the master equation (a stochastic counterpart of the ordinary differential equation), requires a so called Markovian assumption. The assumption implies that the future evolution of a stochastic process depends solely on the current state of the system and that no memory of previous events exists – the stochastic process is *memoryless*. Take the synthesis of X, for example. Suppose that the last production event took place one second ago. According to the memoryless assumption, the probability that the next event occurs after an additional two seconds, given that one second of waiting has already passed, equals just the probability of an event taking place after two seconds regardless of the amount of waiting before. Using a bus stop analogy: the probability of a bus arriving after two minutes would be the same for newcomers as for passengers already waiting at the bus stop. On average, one would wait for a bus for the same amount of time, and it would not matter when during the period between two bus departures one had arrived. It contradicts common sense; people arriving later after the last bus departure *feel* that the bus should appear *sooner*. This is because real buses run

(usually) according to a schedule and intervals between their departures obey a different statistics!

The memoryless assumption constrains the type of the function from which the intervals between events are drawn. It turns out that the only function satisfying this condition is the exponential. This has far-reaching consequences. Any realization of a memoryless stochastic process (a so called Markov chain) implies waiting times between events being drawn from the exponential distribution (just like synthesis and degradation in Fig. 1.3F). The time evolution of such a process has a tractable analytical representation in the form of the master equation. In fact, stochastic processes as those depicted in Figs 1.3A-C where uniform waiting time distribution are used have no closed analytical solutions. Approximations are required which complicates the analysis and numerical simulations.

In the limit of large numbers of reactants the memorylessness is also reflected in the Law of Mass Action. Consider a volume V in which a linear decay of N_0 molecules of species X takes place ($X \rightarrow \varnothing$). The dependency of the continuous concentration $[x] = x/V$ on time is easily obtained as $[N_0] \exp(-k_d\, t)$ from which the inverse equation, the time to reach a given concentration $[x]$, follows:

$$t_x = \frac{1}{k_d} \cdot \log \frac{[N_0]}{[x]}\,. \tag{1.1}$$

In the stochastic description, the time to decay for a single molecule is distributed exponentially with an average $1/k_d$, i.e. $\exp(-k_d\, t)$. The probability for the first decay event in an ensemble of N molecules, can be computed by multiplying probabilities:

$$\begin{aligned} Pr(T^{(1,N)} > t) &= Pr(\text{1-st molecule out of N decays after t}) \\ &\equiv Pr(\text{all N molecules decay after time t}) \\ &= Pr(T^1 > t \,\text{and}\, T^2 > t \,\text{and}\, \ldots \,\text{and}\, T^N > t) \\ &= Pr(T^1 > t) \cdot Pr(T^2 > t) \cdot \ldots \cdot Pr(T^N > t) \\ &= e^{-N\, k_d\, t}. \end{aligned} \tag{1.2}$$

The resulting probability for the first degradation is also exponentially distributed, but with a smaller average, $1/(Nk_d)$. By calculating the probability for the second, third and k-th degradation one obtains the mean time after which $x = N - k$ molecules are left in the system, i.e. the stochastic equivalent of Eq. 1.1:

$$\langle t \rangle_x = \frac{1}{k_d} \sum_{i=(N-k)+1}^{N} \frac{1}{i}. \tag{1.3}$$

For large N these two equations (Eq. 1.1 and Eq. 1.3) are equivalent (Fig. 1.4).

The fact that the product of exponential functions is also an exponential greatly benefits the computational procedure. In a decay example, one (lengthy) algorithm would track the lifetime of every molecule. At the initial time zero, decay times would be drawn from the exponential distribution with the mean k_d for each of N particles. After ordering these next-decay times, the sequence of single-molecule subtractions from the ensemble would yield a sample time-dependent realization of the stochastic process. Such a trajectory could be obtained in a much simpler way with help of Eq. 1.2. The time for the first degradation event in a pool of N particles can be drawn from the exponential distribution

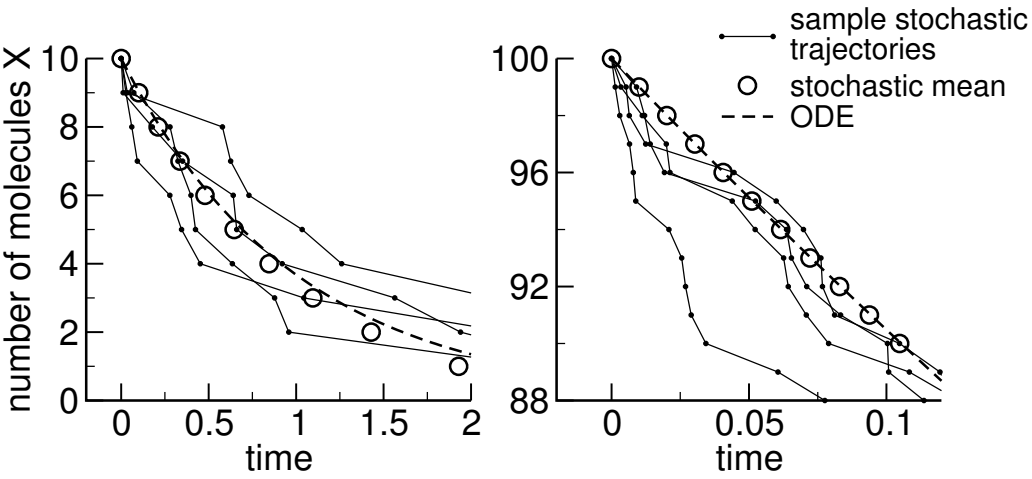


Figure 1.4: Comparison of the stochastic and the deterministic model of a decay reaction, $X \rightarrow \varnothing$. Initially, the system contains 10 (left panel) and 100 (right panel) molecules of species X, each having the same average lifetime τ_X. Stochastic mean (empty circles) is taken over 1000 trajectories simulated with the Gillespie algorithm. The discrepancy between the continuous model (dashed line) and the stochastic mean is apparent only for a small number of molecules (left panel).

with the decay constant $N \cdot k_d$. The exponential waiting time distribution guarantees the memoryless property of the process, thus the time for the next degradation is given by the same distribution but parameterized by $(N-1) \cdot k_d$, and so forth (Fig. 1.5). Contrary to the former scheme, the resulting algorithm does not scale with the number of molecules, N; in order to determine the time of the next decay reaction, only one random number generation is required. Both of these approaches to numerically simulate chemical reactions have been considered in the past and they form the basis of a *next-reaction* and a *direct* method [Gillespie 2007, Gibson & Bruck 2000].

The above holds only for exponentials. Any other form of the waiting-time distribution complicates significantly the scaling of the average first event time (the mean first-passage

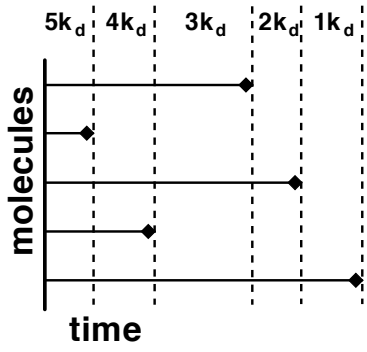


Figure 1.5: A linear decay model with exponential waiting times. One method to create a stochastic trajectory is to draw the next decay time from the exponential waiting time distribution, $k_d \exp(-k_d t)$. When ordered, these times form a sequence of decay events. Such an algorithm scales with the number of molecules. Alternatively, one can take the advantage of Eq. 1.2. First, the time of the first decay is drawn from the exponential distribution with the mean $N \cdot k_d$, then the next decay time is parametrized by $(N-1) \cdot k_d$, and so forth.

time) on N. As a result, the numerical simulation of such a system requires a much more involved implementation and (usually) more computational time. We shall come back to this point later.

Having in mind the severe constraint on the waiting time function imposed by the memoryless assumption one can appreciate the insensitivity of the steady-state product distribution to the type of the waiting time function describing degradation (Figs 1.3D-F). This result extends significantly the amount of solvable models and has a few interesting biologically relevant ramifications. Take mRNA or protein degradation, for instance. Both of these processes proceed in a sequential manner: the poly-adenylated tail of messenger RNA is removed step by step before the final degradation, attachment of ubiquitin monomers to the protein seals its fate and makes it recognizable by proteasomes [Pedraza & Paulsson 2008]. The waiting time statistics of sequential processes departs significantly from the exponential (a peaked, Gamma-like distribution). The insensitivity principle guarantees, however, that the steady-state distribution remains the same (Poissonian) as long as the waiting time distribution of the synthesis process is exponential.

A lot of intuition about biological systems relies on the memoryless property. In the following section we shall pursue some questions of central importance for understanding the origin of exponential approximation, the relationship between the waiting time distribution and steady-state properties such as the mean and noise. Turning to simple models should ease the exposition.

1.2 First example

The model from Fig. 1.2 is the simplest biochemical network with stochastic synthesis and degradation. In numerous biologically important cases, however, synthesis is affected by additional processes which, if collected into a single constant production rate k_p, would reproduce correctly only the average. One of such examples is activation of a gene by a transcription factor. It involves frequent association and dissociation events (a biomolecule makes contact with a site on DNA, forms a complex whose stability depends on the affinity of the binding site, and unbinds) interrupted by long periods during which the transcription factor engages in a long diffusive departure from the cognate DNA site. As a consequence, synthesis takes place only during those periods when the activating protein is bound; protein production occurs in bursts.

A coarse-grained model of such a switch-like product synthesis depicted in Fig. 1.6 omits details of association and dissociation events. At the microscopic level, these are typically affected by diffusion, electrostatics, hydrodynamic interactions or conformational changes of macromolecules. Nonetheless, this simplified description provides some valuable insights into how bursts of synthesis, and fluctuations in waiting times in general, affect the steady-state statistics. This relationship is largely unexplored for biochemical networks as most studies on the consequences of stochasticity for biological systems have focused on the noise in the steady-state level of molecular intermediates.

We assume that transitions between active and inactive states follow the exponential statistics and that production events in the active state occur at exponentially-distributed intervals. This might not hold true in general. A more detailed description could take into account a multi-step nature of gene activation: in order to render a gene active, a few proteins have to assemble a complex. This is the case in higher organisms

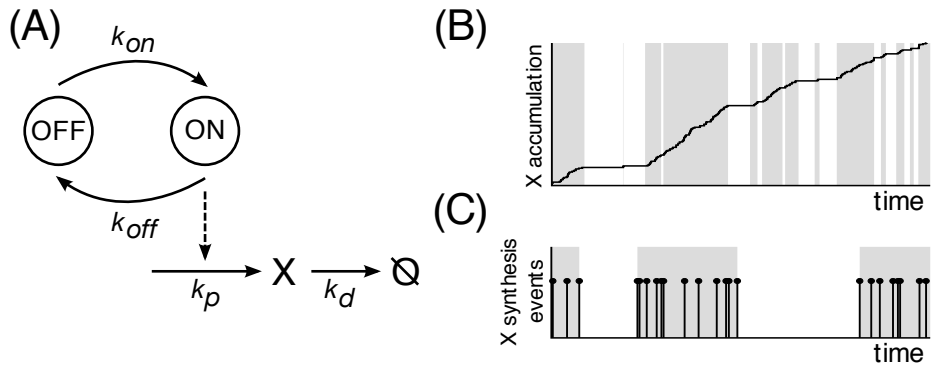


Figure 1.6: The simplest model of bursty production. (A) Synthesis of species X is modulated by a switch. (B) Accumulation of X in time (degradation is omitted here). Light regions correspond to the OFF state, grey represent the active ON state. (C) Synthesis of X takes place during the ON state only.

[Degenhardt *et al.* 2009] and hence switching from OFF to ON state has to be described by a non-exponential (Erlang-type) peaked distribution (the topic explored in more detail in Chapter 2). For simplicity we shall focus on a simpler case where waiting times of all four transitions in Fig. 1.6 are described by the exponential distribution.

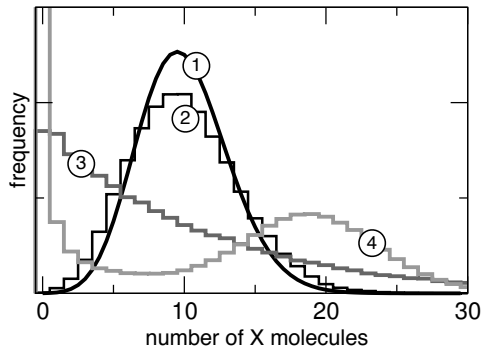
How is the overall waiting time distribution for the production events affected by the switching between synthetic activity and silence periods? In other words, what kind of probability distribution describes the statistics of product appearance given the silence periods interrupting an otherwise constant synthetic streak? Note that we assumed the lifetimes of the ON and OFF states and the intervals between production events in the ON state to have exponential distribution.

In case of a large time scale separation between the ON-OFF switch and the synthesis, i.e. many products are synthesized in the active state and the active state is comparably long to the inactive period, mostly two types of intervals are abundant: short ones corresponding to production events in the ON state, and long ones due to OFF periods (Fig. 1.6). Hence, the resulting distribution must consist of two differently-parameterized exponentials with weights depending on the time contribution from both states:

$$f(t) = w\, r_1\, e^{-r_1 t} + (1-w)\, r_2\, e^{-r_2 t}. \tag{1.4}$$

This heuristically derived equation is in fact an exact solution of the network in Fig. 1.6 for all time scales. The (non-trivial) relation of r_1, r_2 and w to kinetic parameters of the burst model in Fig. 1.6A can be found in the Appendix of Chapter 2.

We established that the stochastic production modulated by the ON-OFF results in the double-exponential waiting-time distribution. The simple birth-and-death model considered in Fig. 1.2 where synthesis events were spaced at exponentially distributed intervals produced a Poisson distribution of the product in the steady-state. Finally, we are able to address the question posed at the beginning of this chapter: how does this steady-state distribution change once the synthesis takes a different form, in this case the form of Eq. 1.4?



type	burst size	burst significance	<X>	STD_X
1 - Poisson	N/A	N/A	10	3.2
2 - burster	2.5	0.0	10	3.8
3 - burster	12.4	0.8	10	9.7
4 - burster	209	0.7	10	9.7

Figure 1.7: Steady-state product distribution affected by bursts. Increasing the burst size in the model from Fig. 1.6 induces a radical change of the steady-state distribution of X: from single-peaked to bimodal.

Sample analytical solutions for steady-state product distribution are depicted in the frequency plot in Fig. 1.7. All of the distributions have the same mean, whereas their shape differs depending on the size of synthesis bursts and their significance. Notably, for small burst size the distribution may take the bell-curve-like shape centered around the mean or, when bursts increase in size, it may peak twice: around the base state (zero) and the steady-state mean in the active state given by the ratio k_p/k_d. The existence of such a bimodal distribution is in no way reflected in the macroscopic description; the stochastic model reveals a whole new family of behaviors.

1.2.1 Physiological implications of bursts

Living organisms employ numerous strategies to strive in their environment. Spontaneous genetic mutations accompanied by the selection process lead to fixation of new, more "successful" genotypes. However, the time scale of these changes is too long to adjust to changes in nutrient abundance taking place on an hourly basis, for instance. A multitude of sensory networks capable of inducing and regulating gene expression have evolved to cope with changing conditions. Even for bacterial sensing largely relying on relatively simple two-component networks, a high number of such networks (typically around few dozen per cell) and the fact that they operate in parallel with numerous cross-talks have led some to postulate a rudimentary form of intelligence embedded in the sensory network [Hellingwerf 2005]. Whether their complexity is sufficient to exhibit neural network-like characteristics such as memory or learning is an intriguing concept awaiting further theoretical and experimental evidence.

While the ability to "intelligently" process environmental inputs by bacterial sensory machinery remains a hypothesis, the ability to reflect temporal pattern of extracellular conditions in the structure of regulatory network has been recently demonstrated experimentally [Mitchell *et al.* 2009]. Two model organisms, *E.coli* and *S.cerevisiae*, used in this study evolved to be capable of activating parts of its machinery in anticipation of the sequence of stimuli. Random, unpredictable fluctuations in environmental conditions pose a grander challenge for organisms, however. Likewise, rapid but infrequent (on

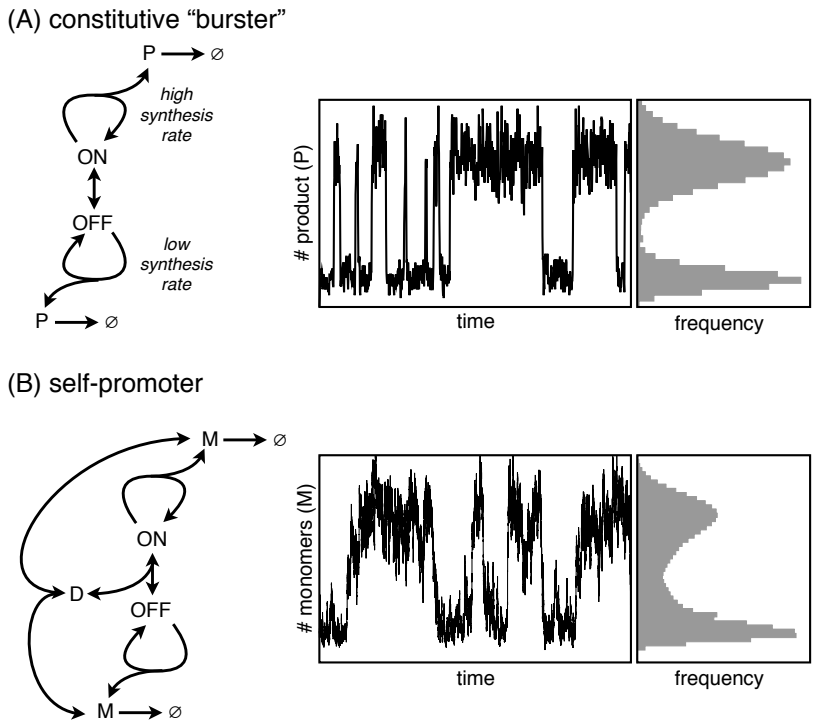


Figure 1.8: Bimodal behavior of the steady-state product distribution may arise in a nonlinear dynamical system as well as in a stochastic switch. (A) Spontaneous stochastic switch, a "burster", with constitutive production of P. If ON and OFF states are long enough to establish their respective quasi steady-states, the system will "flip" between two distinct states. (B) A classical mechanism leading to bimodality – a positive feedback. Product M promotes its own synthesis by forming a dimer D which further induces the active ON state. Dimerization introduces nonlinearity which may lead to bistability if none of the processes acts too strongly [Isaacs *et al.* 2003].

cell's generation scale) cues inducing irreversible lethal changes require a different type of response system. One solution is to maintain a multitude of regulatory mechanisms continuously prepared to face a vast array of environmental challenges. But even then, a cell fully equipped with sensory networks activating appropriate genes may respond to a change only if integration of the extracellular signal lasts long enough to average out noise inherent in the sensing procedure. If the time required for gathering sufficient information about the environment exceeds the generation time of a single cell, such a cell is never capable of reacting properly, which compromises the fitness of the population as a whole. Diversification of phenotypes may be a useful strategy to overcome this limitation.

Fitness advantage of a phenotypically heterogeneous population exposed to fluctuating conditions depends on the relative time scale of intracellular processes and changes in environmental cues [Kussell & Leibler 2005]. The choice of the strategy arguably depends on cells' ability to gather information about the outside world. Cells switching their pheno-

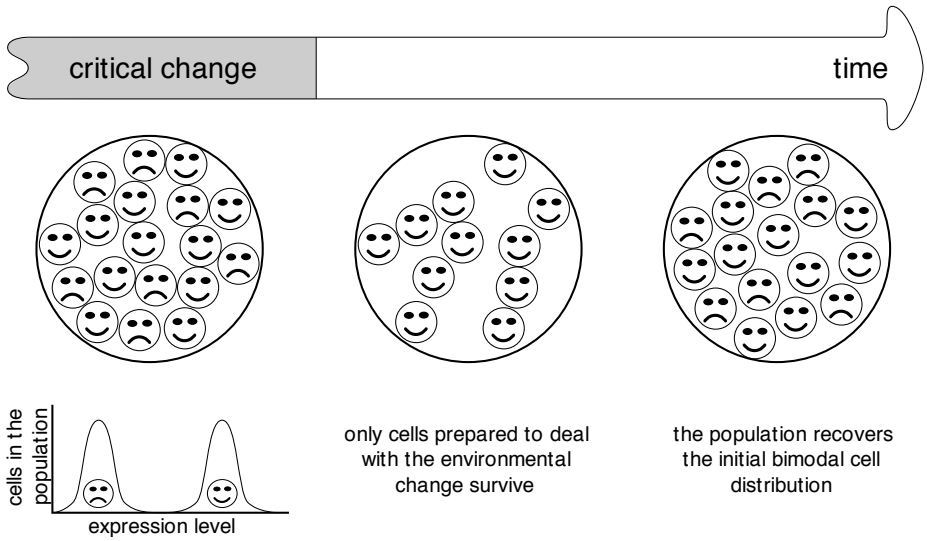


Figure 1.9: Population heterogeneity increases survival rate. In the simplest case, the abundance of the protein dealing with environmental challenge (e.g. neutralizing a toxin) has a bimodal distribution due to stochastic gene expression or due to feedback regulation, a so-called *feedback-based multistability* [Smits *et al.* 2006]. The example of the latter mechanism could be a positive feedback in the sporulation network of *Bacillus subtilis*. More generally, bi- or multi-modality could refer to meta-states where gene expression profiles of many genes undergo significant changes simultaneously [Chang *et al.* 2008]. Bimodality results in a stable, phenotypically heterogeneous population. Only individuals "prepared" for the upcoming change in the environment survive. The initial heterogeneity is recovered after few generations.

types with the same frequency as changes in the environment have a greater survival rate as opposed to populations staying out of sync with the surroundings [Acar *et al.* 2008]. This observation hints at a possible choice of strategies aiming for adaptation. In environments with rapid, irregular or extreme changes the cost of maintenance of elaborate sensory networks may exceed the benefits of a higher survival rate. Additionally, the response of the network to a rapid change may be unsatisfactorily slow thus promoting a simpler solution. An isogenic (having the same set of genes) population could, for instance, increase its fitness by generating subpopulations by stochastic modulation of gene expression; each of these subpopulations performs sub-optimally in the "average" environment but is able to survive critical changes – a strategy known as bet-hedging (for a comprehensive review see [Davidson & Surette 2008]).

Bimodality in the gene product level as depicted in Fig. 1.8 is one of the elementary mechanisms behind induction of population diversity in a population of isogenic cells. Two peaks of the steady-state product distribution have been traditionally attributed to bistability in the dynamical system resulting from some form of feedback regulation; a so-called *feedback-based multistability*. A thorough review concerning this type of mechanisms can be found in [Smits *et al.* 2006]. Here, we discuss a different case where two

distinct levels of a protein across the population emerge solely due to stochastic gene regulation (Fig. 1.8) [Kepler & Elston 2001, To & Maheshri 2010]. Cellular heterogeneity due to bursty protein production may arise only if the frequency of ON-OFF transitions is comparable to (or slower than) the generation time of the cell (Fig. 1.1); fluctuations are "slow" [Sigal *et al.* 2006]. Only them, extensive inactivity periods during protein synthesis may leave some cells in the population lacking these molecules, while other cells may contain the entire production burst [Yu *et al.* 2006].

The emergence of stochastic bet-hedging strategy in a bacterial population has been recently observed in an elegant experiment [Beaumont *et al.* 2009]. After 15 rounds of subjecting cells to two opposing environments, each favoring a different phenotype, a genotype evolved capable of stochastic switching between the two conditions. Stochastic gene expression has been also shown to confer fitness advantage in a population of yeast cells exposed to antibiotic stress [Blake *et al.* 2006]. Due to bursts in the production of the protein conferring resistance to an antibiotic, (at least) part of the population is in the position to respond rapidly (Fig. 1.9). This translates to an overall increase in the survival rate of the whole population. Survival upon antibiotic treatment in general has been attributed to population heterogeneity stemming from stochastic switching between two phenotypes with distinct survival rates, a phenomenon known as bacterial persistence [Balaban *et al.* 2004, Kussell *et al.* 2005, Bishop *et al.* 2007]. Similarly, Sorger and colleagues [Spencer *et al.* 2009] have demonstrated how cancer cells escape drug treatment thanks to stochastically induced cell-to-cell variability. Studies like this should help answering a long-standing question of why seemingly identical cells respond differently to a drug? In a remarkable experiment of Cohen et al., levels and locations of approximately 1000 endogenous proteins have been tracked in individual cells after admitting a chemotherapy drug [Cohen *et al.* 2008]. The presence of the drug evoked a higher variability in protein levels and in some cases concentrations exhibited a bimodal distribution. The low and high levels in protein concentration corresponded to survival or death of a cell. Thus, the presence of the environmental stress induced a stochastic strategy, which allowed part of the cancer cell population to escape their deadly fate. Phenotypic diversity in a clonal population is rarely a result of fluctuations in the expression of a single gene, however. Cellular states with distinct functionalities (phenotypes) usually correspond to transcription profiles differing in the expression of many genes. Slowly-decaying fluctuations may promote reversible transitions between such meta-states implying various cell fates as has been demonstrated for mammalian progenitor cells [Chang *et al.* 2008].

Diversification of the microbial population benefits its survival rate upon environmental changes but also allows these simple organisms to solve complex tasks. In one of such systems, a population of Salmonella bacteria split stochastically into two subpopulations. One of the phenotypes facilitates infection in the gut lumen by triggering the inflammation. By doing so it contributes for the "public good" of the whole bacterial invasion, however, these bacteria get eliminated by the host's immune response [Ackermann *et al.* 2008]. This self-destructive cooperation elegantly illustrates the population benefit of phenotypic noise and its role in bacterial pathogenesis. It is also an interesting contribution to recent experimental studies on the evolution of cooperation by engineering simple bacterial ecologies [Santorelli *et al.* 2008, Gore *et al.* 2009, Khare *et al.* 2009].

1.3 Timing of biochemical reactions

Thermal fluctuations, diffusion in crowded cellular environment, small numbers of the reactants, all of these affect rates of chemical reactions and contribute to the overall stochasticity in biochemical processes. Such inherent randomness may be selected in an evolving system if it increases the fitness of the organism [McAdams & Arkin 1999]. But living organisms are not falling apart, nor is their functioning completely erratic. Quite the contrary, their survival requires transfer of genetic information between generations, protection of this information against damages, processing of intracellular and environmental queues. Cellular events such as cell cycle, prediction of day-night rhythms, or anticipation of sequential environmental changes [Mitchell *et al.* 2009] require temporal synchronization between various processes.

Numerous mechanisms have evolved to achieve robust functioning despite biochemical noise. Steep input-output relationship in sensory systems, ultrasensitivity, arises in networks organized hierarchically [Hooshangi *et al.* 2005, Bruggeman *et al.* 2009, Cluzel *et al.* 2000, Tănase-Nicola *et al.* 2006]. Arrangement of enzymatic reactions in cycles helps to reduce the variance in the lifetime of molecules, hence increasing the timing accuracy [Qian 2006, Li & Qian 2002]. Molecular ratchets consisting of many irreversible steps give means to discriminate between states of similar free energies as in kinetic proofreading [Hopfield 1974, Salazar & Höfer 2009]. A ratchet formed by the state of DNA methylation coupled to regulation of DNA replication increases temporal precision of the cell cycle [Collier *et al.* 2007, Shen *et al.* 2008]. Irreversible path in protein evolution settles new functionality of complex architectures and possibly restricts other potential paths that could be explored under selection [Bridgham *et al.* 2009]. Redundancy in the structure of biochemical networks, motifs in feedback regulation [Legewie *et al.* 2008], feedback loops or cascades, large numbers of molecules, all of these increase functional stability of cellular processes [Alon 2006, Wagner 2007].

1.3.1 Sequential processes

For a single molecule noise, or variance scaled with the squared mean, in the time to complete a biochemical transition (the waiting time) decreases if the transition is divided into substeps [Li & Qian 2002, Pedraza & Paulsson 2008]. If the timing of all N processes in the sequence is of similar duration, the resulting waiting time distribution becomes narrower than the waiting time distribution of a single-step transition (Fig. 1.10). For a simplified case where the rates of all intermediate reactions are equal, the noise depends inversely on the length of the sequence.

A process with events drawn from a non-exponential waiting time distribution does not have a straightforward counterpart in a macroscopic model described by ordinary differential equations (ODEs). Macroscopic biochemical models assume exponential waiting times, i.e. a Markov chain is the underlying stochastic description of chemical reactions (see Eqs. 1.1 and 1.3). Divergence from the memoryless distribution has a major consequence for the steady-state product distribution (Fig. 1.3) but also for the transient as illustrated by the linear decay reaction in Fig. 1.11. If the lifetime of a molecule is distributed according to a non-exponential function, in this case a Gamma distribution arising due to a step-wise degradation, the decay proceeds much faster than in an equiva-

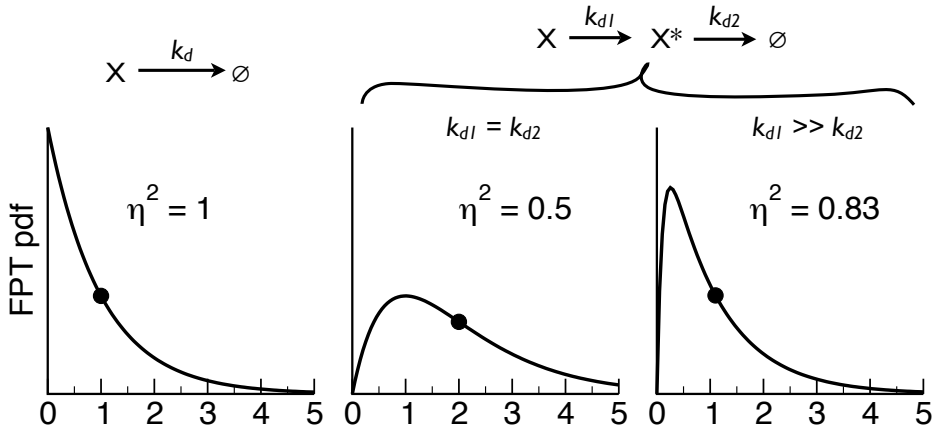


Figure 1.10: Waiting time distribution (the first-passage time probability density function – FPT pdf) and noise (η^2, variance over squared mean) in a sequential process. Black dot indicates the mean of the distribution. As an example we consider a decay reaction of species X. (Left) A single molecule X decays in a single transition with rate constant k_d. The corresponding waiting time distribution is an exponential with mean and noise equal 1. (Middle) A single molecule X decays in two substeps with rates k_{d1} and k_{d2}, respectively. Each substep has an exponential waiting time distribution. The overall distribution is a result of a convolution of the two distributions of the substeps. In this case, both rates equal 1, which results in the minimal noise and the narrowest distribution for such a sequence. (Right) Time scales of the two processes differ ($k_{d1} = 1$ and $k_{d2} = 0.1$). Noise approaches the value of the dominating step, i.e. the one with the larger mean.

lent process where the mean lifetime of the molecule is the same but the reaction consist of only a single step.

The above discrepancy between the time evolution of one-step and multi-step decay processes illustrates an important modeling issue. Very frequently, a description of biochemical reactions involves "bundled" steps which in fact consist of a number of shorter steps. Formation of a transcription initiation complex, mRNA or protein degradation, signal transduction are just a few of such processes. As we have seen, a more detailed description accounting for these sub-steps has a significant effect on the steady-state noise or the time evolution of the system. Notably, in the latter case a mere expansion of the macroscopic model with additional linear reactions describing all the steps in the process will not match the stochastic transient. Reaction rates in such an expanded model would require intricate dependencies on state variables.

Timing of biochemical events becomes more precise if processes are aligned in a sequence. However, the time to complete such a sequence increases with the amount of steps. Therefore, a gain in noise reduction may be offset by an extended duration of the process and possibly a higher cost of production and maintenance of proteins that make up the reaction chain. The simple *exit time* process from Fig. 1.12 provides an illustration of these tradeoffs. Molecules of species X are synthesized and degraded in first-order reactions. Further action is triggered only if a certain amount of X is accumulated; thus,

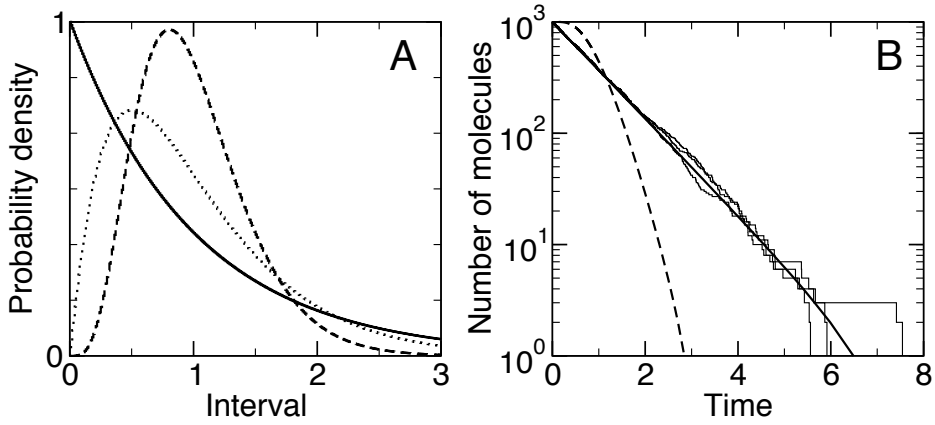


Figure 1.11: The effect of the peaked waiting time distribution on a decay reaction. (A) Waiting time distributions for a Poisson process (exponential distribution - solid line), the sum of two exponentially distributed random variables (gamma distribution with $k = 2$ steps - dotted), and the sum of five variables (gamma with $k = 5$ steps - dashed). All distributions have the same mean equal to 1. (B) Stochastic dynamics of a decay reaction, $X \to \varnothing$. Waiting times for a decay of every X molecule are drawn from the exponential distribution (solid line) and from a 5-step ($k = 5$) gamma distribution of the same mean equal to 1 (dashed). Averages are taken over 100'000 trajectories. Additionally, three arbitrary stochastic trajectories are drawn for the exponential case (light solid).

a sequence of synthesis events is required.

The higher the threshold in the exit time process, the longer the time to reach it; the effect enhanced by the degradation which forces the build-up level down to the initial zero value. The width of the exit time distribution becomes narrower with the increasing threshold level only to widen again in the regime where the threshold is comparable to or exceeds the steady-state mean of the synthesis-and-degradation. The narrowest exit time distribution corresponds to the smallest variability coefficient, or noise; the timing for this threshold value is the least stochastic.

An example of a biological process described by this model is the switching of the rotation direction of the flagellar motor in bacteria such as *E.coli*. Microbes are typically equipped with several independent motors. Their coordinated rotation is necessary to propel attached flagella and to direct the organisms towards regions more abundant with nutrients. How can a set of uncoupled motors can start their motion synchronously? One plausible explanation could be the threshold mechanism and its timing properties. A change in extracellular nutrient concentration gradient triggers a release of the activated signaling protein (phosphorylated form of CheY) at the cell's "nose" – a relatively small region at one end of the cell (Fig. 4.3). Sequential build-up of these proteins at the motor site is required to change the direction of the flagella rotation. As shown in Fig. 1.12, the timing might be very precise if the activation threshold is placed slightly below the steady-state concentration. Consequently, a simultaneous switch of the coordinated motion can arise without any physical coupling of the motors (sic!)

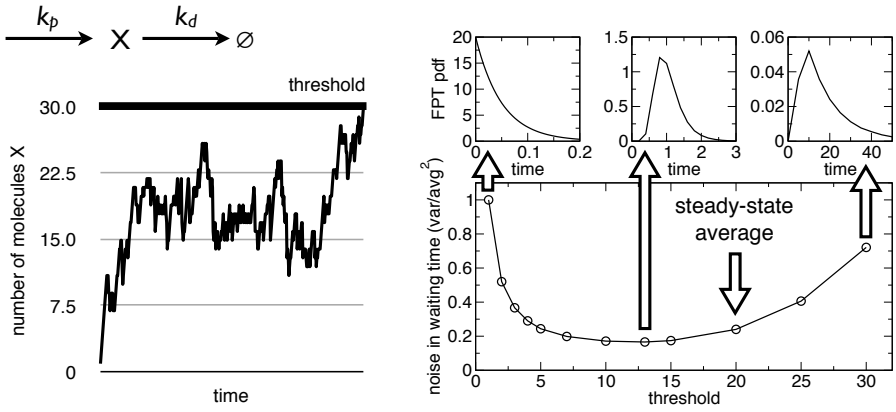


Figure 1.12: Reaching the threshold in a simple synthesis and degradation reaction. (A) An illustration of a single stochastic trajectory crossing the threshold level of molecule X. (B) Distributions and noise in the time to reach the threshold value of X. The steady-state average in the synthesis-degradation model equals 20; $k_p = 20\,[1/T]$, $k_d = 1\,[1/T]$. The smallest noise corresponds to the most peaked waiting time distribution.

1.3.2 Diffusion-limited reactions

Another wide class of processes giving rise to non-exponential waiting time distributions is a diffusive encounter of two reacting molecules (Fig. 1.13). In fact, diffusion can be perceived as a more general form of a sequential process where an infinite number of possible diffusive trajectories (sequences of small jumps in space) overlap. For a pair of initially separated molecules, the distribution of times to bind follows a peaked function similar to the distribution arising in the chain of chemical reactions.

The peakedness, and hence the smaller noise of the waiting time distribution can be further increased if many diffusing molecules perform the search simultaneously (Fig. 1.14). The increased precision in timing could be advantageous for signaling systems. When a change in extracellular conditions takes place, the reception of the signal by membrane sensors is proceeded by a race of a number of activated molecules (so called response regulators) towards the center of the cell. If one of the regulators successfully binds to a DNA site, gene expression can be altered in order to evoke appropriate physiological response. Due to a narrow waiting time distribution resulting from simultaneous diffusion of many molecules, the response will arguably vary only little among the cellular population allowing for a higher survival rate in face of a possibly fatal environmental change. However, mean and noise reduction is likely offset by the increased cost of protein production resulting, an issue investigated in detail in Chapter 3.

Dealing with non-exponential waiting time distributions for diffusion processes is cumbersome for only the Laplace form of the function is known analytically in most cases. Additionally, derivation of multi-particle distributions (according to Eq. A.13) requires knowledge of the explicit time-domain solution which often can be obtained only by means of numerical inversion. Many problems in biology, however, involve diffusive searches for small, compared to the entire volume of the cell, targets such as DNA binding site or a

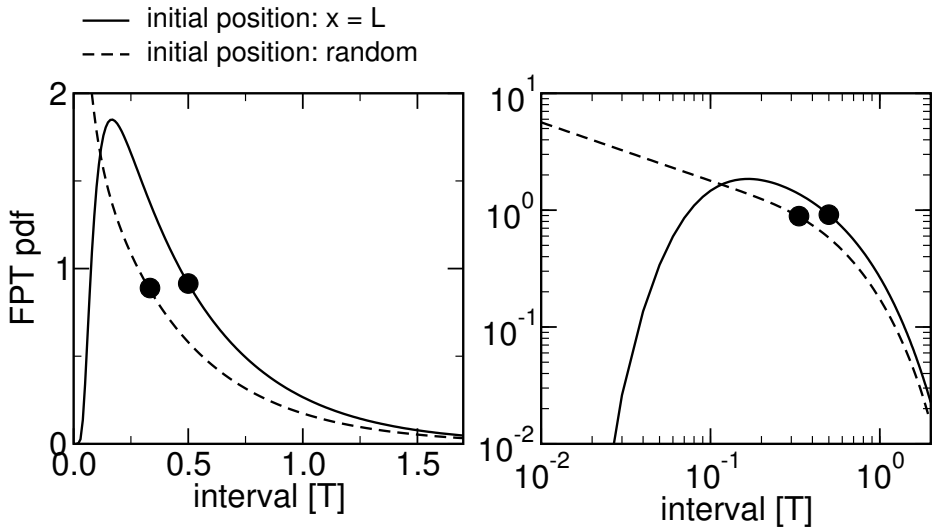


Figure 1.13: An illustration of the waiting time *pdf* for a molecule reaching the $x = 0$ end of the 1D domain of length L; diffusion coefficient $D = 1$. Results for two initial conditions are shown: the molecule initiates at the opposite end ($x_0 = L$) to the target, the molecule initiates at a random position inside the domain. Both panels show the same functions but in different coordinates. An explicit solution, Eq. B.17, has been plotted for $x_0 = L$, and a numerical inverse Laplace transform [Stehfest 1970] of Eq. B.18 for a random initial condition. Similar results are available for a three-dimensional spherical domain. Dots mark the mean.

localized molecule. Such searches involve so many steps that the initial configuration has only a minor effect on the duration of the search. The process becomes effectively memoryless, and the waiting time distribution, although still peaked, is wide enough to allow for the approximation with the exponential function [Dobrzyński 2008] (see Fig. 1.14 – a single-particle waiting time distribution can be well approximated by an exponential function). As a result, variability of the time to reach the target (the first-passage time) increases and approaches that of the Poisson process (Fig. 1.15).

The memoryless approximation of search trajectories is a valuable simplification for modeling and simulations of spatially resolved problems. In this regime, a diffusion-limited reaction involving an encounter of two random walking biomolecules can be described by a first-order transition where the waiting time follows a memoryless exponential function. Costly computer simulations that are typically required to tackle problems involving diffusion may be therefore superseded by a much simpler model quickly solvable by kinetic Monte Carlo methods such as the Gillespie algorithm.

1.4 Organization of the thesis

The generic burst-generating mechanism discussed in Fig. 1.6 was used to study statistics of intervals between syntheses of product X. In biological context, the ON/OFF switch represents the simplest model of gene activation and X is the protein or messenger

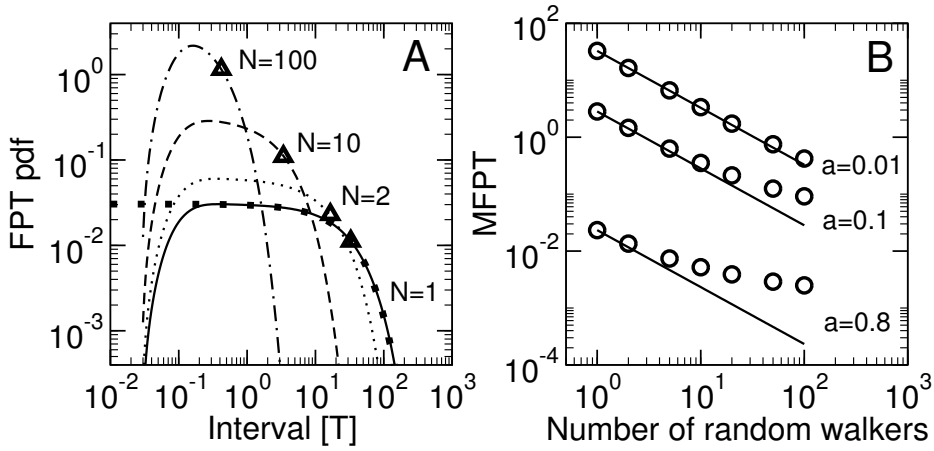


Figure 1.14: The effect of many random walkers (e.g. response regulators) on the first-passage time (FPT) *pdf*. Molecules start simultaneously at the outer membrane, $r_0 = R_{cell}$; the first-passage occurs when the first out of N walker reaches the inner sphere of radius $a = 0.01R_{cell}$. For a single walker, $N = 1$, the FPT *pdf* corresponds to that in Fig. 3.8; the exponential approximation is indicated by black dots. Triangles denote the mean of the distributions. Mean and noise (not shown) in the waiting time decrease with the increasing amount of random walkers. (B) The mean first-passage time (MFPT) as function of the number of molecules for different target sizes a. Solid lines denote the exponential-like scaling $1/N$, which holds for small target sizes. Circles are computed using order statistics for the analytical result: numerical Laplace inversion of Eq. 3.5 to obtain a single particle FPT *pdf*, substitution of $f(t)$ into Eq. A.13, finally computation of moments of $f^{(1,N)}(t)$.

RNA (mRNA), the intermediate synthesized during protein production. In **Chapter 2** we extend this picture and study how additional processes involved in protein synthesis influence the overall stochasticity of gene expression. We focus on a more realistic case, which includes transcription elongation, a process consisting of a multitude of small steps. We perform extensive computer simulations to demonstrate how transcription elongation can attenuate or enhance fluctuations introduced upstream of this process. In order to quantify our analysis we apply the first-passage time theory and introduce indexes for burst size, duration and significance, which have a clear analytical representation for the generic burst model. This chapter is based on published material [Dobrzyński & Bruggeman 2009].

Our focus on timing properties of stochastic events continues in **Chapter 3**. Here we look closer into a spatially-resolved system: a two-component signaling network. This relatively simple structure is responsible (among other tasks) for sensing changes in nutrient abundance outside of the cell and for inducing appropriate physiological responses within the cell. The response involves diffusion-limited searches of spatially-fixed targets by biomolecules. One of the main questions we set out to answer is how many molecules are required to facilitate a quick and robust response. Once again we employ the first-passage time theory and make use of the memoryless property of diffusion-limited reactions. Parts of the background theory used in this chapter has been published in form of conference

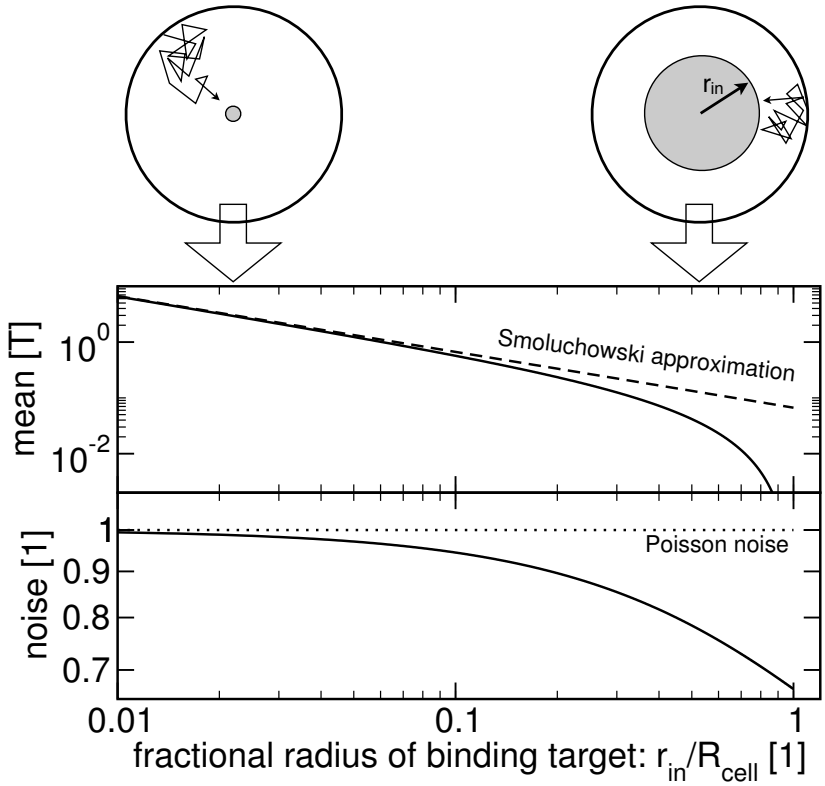


Figure 1.15: Validity of Smoluchowski approximation in the model of target search. The mean and the noise are shown for the first-passage time of a molecule diffusing from the outer boundary at R_{cell} (e.g. cell's membrane) to a target (e.g. a DNA-binding site or cell's nucleus) of radius r_{in} in the center of a spherical cell. The target search depicted here is a model of the first step in the cellular signal response (discussed in more detail in Chapter 3). For targets much smaller than the cell's radius ($r_{in}/R_{cell} \ll 0.1$) the mean search time is well approximated by the estimation derived from Smoluchowski diffusion-limited rate constant K_D, i.e. $\tau_D = 1/K_D = V_{cell}/4\,\pi\,D\,r_{in}$. Note that since noise is a dimensionless quantity it is also independent of the diffusion constant D.

proceedings [Dobrzyński 2008].

Likewise, in **Chapter 4** the analysis of the first-passage time distribution for a bimolecular reaction allows for a thorough comparison of spatial-stochastic computation methods. These methods are used in order to obtain statistics of biochemical processes involving diffusion. Higher moments and steady-state distributions of reactants are typically computed this way. We analyzed available methods which assume different underlying physical models of the diffusion and identified sources of discrepancies in their results. This chapter is based on published material [Dobrzyński *et al.* 2007].

Three appendices include mathematical definitions and derivations of essential equations discussed throughout the thesis. **Appendix A** covers basic concepts of the first-passage time theory such as the first-passage time probability density function, com-

putation of the first and the second moment of the distribution, and order statistics. **Appendix B** focuses on the first-passage time theory in the context of diffusion-limited reactions. Examples of derivations for one-dimensional geometries can be found there. The methods presented there along with the order statistics from **Appendix A** are used extensively in **Chapter 3** to compute distributions and moments of first-passage times for many diffusing molecules in a signaling network. Derivations specific to this three-dimensional problem are included in **Section 3.5** – Materials and Methods. **Appendix C** includes a detailed description of models and computational methods along with algorithms applicable to modeling of spatial stochastic problems in biology.

Chapter 2

Elongation dynamics shape bursty transcription and translation

Contents

2.1 Abstract **25**
2.2 Introduction **26**
2.3 Results **27**
2.3.1 Analytical expression of the waiting time distribution 27
2.3.2 Measures for characterization of bursts 29
2.3.3 Motor-protein traffic jams along biopolymer chains 31
2.3.4 Pausing of motor proteins can generate bursts 33
2.3.5 Aggregative behavior of multiple burst-generators 35
2.4 Discussion **36**
2.5 Materials and methods **38**
2.5.1 Statistics of the arrival process 38
2.5.2 The limit of a large time scale separation 40
2.5.3 Moments of the first-passage time *pdf* 40
2.5.4 Quantitative characterization of bursts 41
2.5.5 Non-exponential waiting time distribution for the switch 46
2.5.6 Interarrival time CDF in a pool of unsynchronized IPPs 48
2.5.7 Progression of motor proteins along the polymer 48

2.1 Abstract

Cells in isogenic populations may differ substantially in their molecular make up due to the stochastic nature of molecular processes. Stochastic bursts in process activity have a great potential for generating molecular noise. They are characterized by (short) periods of high process activity followed by (long) periods of process silence causing different cells to experience activity periods varying in size, duration and timing. We present an analytically solvable model of bursts in molecular networks, originally developed for the analysis of telecommunication networks. We define general measures for model-independent characterization of bursts (burst size, significance, and duration) from stochastic time series.

Inspired by the discovery of bursts in mRNA and protein production by others, we use those indices to investigate the role of stochastic motion of motor proteins along biopolymer chains in determining burst properties. Collisions between neighboring motor proteins can attenuate bursts introduced at the initiation site on the chain. Pausing of motor proteins can give rise to bursts. We investigate how these effects are modulated by the length of the biopolymer chain and the kinetic properties of motion. We discuss the consequences of those results for transcription and translation.

2.2 Introduction

The stochasticity of molecular processes contributes to heterogeneity in populations of isogenic cells. Cellular heterogeneity is manifested by differences in the copy numbers of molecules and in the timing and duration of processes. Recent advances in single-cell measurement have facilitated the quantification of stochastic phenomena [Elowitz *et al.* 2002, Ozbudak *et al.* 2002, Golding *et al.* 2005, Yu *et al.* 2006] (reviewed in [Kaern *et al.* 2005, Kaufmann & van Oudenaarden 2007]). Together with models and theory much insight has been obtained into the sources of noise and how particular network designs contribute to noise suppression and amplification [Elf & Ehrenberg 2003, Simpson *et al.* 2003, Paulsson 2004, Pedraza & Paulsson 2008].

Stochasticity of gene expression has been described by distributions of macromolecules in a population of cells [Elowitz *et al.* 2002, Ozbudak *et al.* 2002]. Whether averaging over a population captures the entire spectrum of molecular fluctuations a particular cell experiences over one generation, depends on magnitudes and rates of fluctuations. If these are slow but high in amplitude, the required averaging duration may extend over a generation span [Rosenfeld *et al.* 2005]. Then, a single cell may not even be able to reach protein states accessible to other members, thus rendering cell-cell protein level distributions uninformative with respect to behavior of genetic circuits [Sigal *et al.* 2006]. In such cases, waiting times for individual birth and death events need to be monitored in order to assess physiological constraints on a single-cell level. This stochastic nature of waiting times will be our focus. Little analytical theory has been developed to deal with this phenomenon despite its relevance for single cell behavior.

The waiting times in a first-order process with rate constant k follow an exponential distribution; the mean waiting time between events and its standard deviation are equal to $1/k$. The waiting time for an event is no longer exponentially distributed if it is regulated by another process. This mechanism underlies bursts in synthetic activity. The interrupted Poisson process (IPP) was introduced to study bursts in queuing and telecommunication theory [Kuczura 1973]. In an IPP, a stochastic switch modulates a process with exponentially distributed waiting times. Depending on the time scale separation between the process and the switch, multiple time scales may appear in waiting times for production events.

Bursts have received ample attention in the biophysics literature [Kaern *et al.* 2005, Pedraza & Paulsson 2008, Colquhoun & Hawkes 1982, Walczak *et al.* 2005, van Zon *et al.* 2006, Mitarai *et al.* 2008]. These studies tend to focus predominantly on the protein number distributions, but do not analyze the distributions for waiting times in much depth. We show that such statistics are relevant for burst characterization and the molecular mechanisms giving rise to bursts.

Bursts have been experimentally observed for synthesis of mRNA and protein [Golding *et al.* 2005, Yu *et al.* 2006, Raj *et al.* 2006, Cai *et al.* 2006, Chubb *et al.* 2006, Newman *et al.* 2006, Bar-Even *et al.* 2006]. They are characterized by rapid productions of a number of mRNA or protein molecules during short time intervals. Periods of synthetic silence occur between bursts. Bursts may give rise to significant disturbances of cellular physiology depending on burst size and the duration of synthetic silence and activity. Even though the benefit of bursts needs to be analyzed further, they could be beneficial for cells living in rapidly fluctuating environments [Acar *et al.* 2008]. Bursts may give rise to a bimodal distribution of protein expression across cell populations [Friedman *et al.* 2006]. Thereby, two sub-populations could emerge having different adaptive potentials.

We apply the analytical theory of IPPs to a molecular mechanism for bursts. In order to identify and characterize bursts we derive three new indices: burst size, duration, and significance. We demonstrate how motor protein trafficking along biopolymer chains (such as mRNA polymerase and DNA polymerase along DNA, ribosomes along mRNA, and cargo-carrying dynein along microtubuli) can generate bursts depending on the length of the biopolymer and stochasticity of initiation and motion. We show that motor proteins can generate bursts by pausing or by memory of initiation bursts.

2.3 Results

2.3.1 Analytical expression of the waiting time distribution

In this section, we study a small network to gain insight into burst-generating mechanisms. This will allow us to derive general indices for the characterization of burst properties. These indices will be applied to characterize biological mechanisms.

The network consists of a source switching between an inactive *OFF* and an active *ON* state according to a Poisson process (Fig. 2.1A). *OFF* and *ON* periods are defined on the level of the switch (see [Walczak *et al.* 2005, Mitarai *et al.* 2008] for the discussion of mechanisms giving rise to genetic switches). In the active state, production of P, e.g. mRNA or protein, occurs at exponentially distributed intervals of length $\tau_{ini} = 1/k_{ini}$. The average *ON* period lasts for $\tau_{on} = 1/k_{sw}^{-}$. Production periods are interrupted by transitions to the *OFF* state. Each rate constant corresponds to the inverse of the mean first passage time for a complex kinetic mechanism. We assume that it follows an exponential waiting time distribution.

Product generation is bursty if it occurs many times during one *ON* state. In addition, the duration of the *OFF* state ($\tau_{off} = 1/k_{sw}^{+}$) should be longer than or comparable to the *ON* period. Under these conditions, waiting times display two time scales (Fig. 2.1B). Since our interest is the statistics of intervals between production, the degradation of P does not play a role.

The mechanism discussed here specifies an interrupted Poisson process (IPP) investigated in the field of queuing theory [Kuczura 1973]. An IPP is a Poisson process for event occurrence (*arrivals*) modulated by a random switch. In this framework a gene that switches between an *ON* and *OFF* state as function of a transcription factor would be considered the source. Arrivals would, for instance, correspond to initiations of transcription giving rise to elongating RNA polymerases. The IPP theory provides the probability density function (PDF) for waiting times between production events, $f_X(t)$, with the

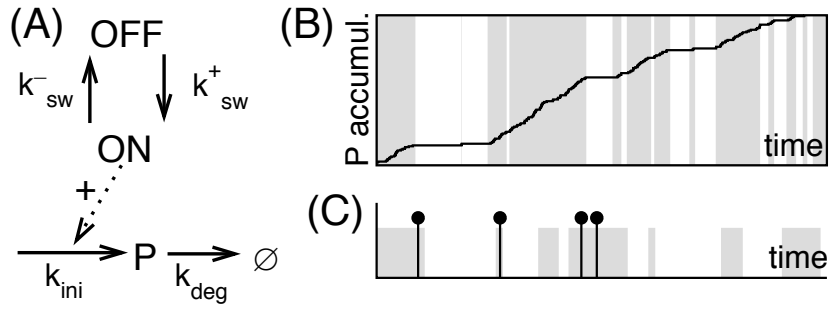


Figure 2.1: Bursts generated by the minimal model. (A) The network consists of a switching source and a Poissonian product generator. Full arrows denote reactions. Product P is synthesized only in the *ON* state. k^+_{sw}, k^-_{sw}, k_{ini} and k_{deg} denote the *ON* switching, *OFF* switching, production, and degradation rate constant, respectively. (B) Simulation of bursty accumulation of P. Two time scales correspond to uninterrupted and interrupted production events. Bars denote *OFF* (white) and *ON* (grey) state. (C) Waiting times for non-bursty production events (vertical lines). On average, 1 P is produced during the *ON* state. The resulting intervals between production events correlate weakly with *OFF* and *ON* states.

stochastic variable X as the waiting time with value t. It is instructive to realize that the PDF can become larger than 1 (it is *not* a probability) and that $\int_0^\infty f_X(t)dt = 1$. The probability of an interval between consecutive events being within $[t, t+dt]$ equals $f_X(t)\,dt$. The PDF of an IPP is a weighted sum of two exponential distributions,

$$\begin{aligned} f_X(t) &= P[X \in (t, t+dt)]/dt = w_1 r_1 e^{-r_1 t} + w_2 r_2 e^{-r_2 t}, \\ r_{1,2} &= \left(K \pm \sqrt{K^2 - 4k_{ini}k^+_{sw}}\right)/2, \quad r_1 > r_2, \end{aligned} \tag{2.1}$$

where $K = k^+_{sw} + k^-_{sw} + k_{ini}$, and the weight factors $w_1 = 1 - w_2 = (k_{ini} - r_2)\,/\,(r_1 - r_2) \in (0, 1)$; derived in Section 2.5.1 of Materials and Methods. The PDF can reveal the presence of two times scales in a stochastic time series (Fig. 2.2, first row).

The length of intervals between production events follows from the superposition of two independent processes: production during a single *ON* state, and periods of synthetic inactivity. The latter may result from multiple switches between *ON* and *OFF* states without producing any P. This is the case if the mean number of productions per *ON* state is small, i.e. $k_{ini} \approx k^-_{sw}$ (Fig. 2.1C). Accordingly, synthetic activity and silence periods (at the level of P production) do not strictly overlap with *ON* and *OFF* states of the switch.

Characteristic time scales of the fast and the slow process appear in Eq. 2.1 as rates r_1 and r_2. Weight factors w_1 and w_2 are the probabilities to observe the short period (mean duration $1/r_1$) and the long period (mean $1/r_2$) between P productions, respectively. For large time scale separation, i.e. when k^+_{sw} and k^-_{sw} are much smaller than k_{ini} (last column in Fig. 2.2), the rates become $r_1 \approx k_{ini}$ and $r_2 \approx k^+_{sw}$. In this regime, the *expected burst size*, β_e, equals the number of initiations per *ON* state, i.e. k_{ini}/k^-_{sw}.

The point of time scale separation we refer to as τ_X (Fig. 2.2). At this interval the

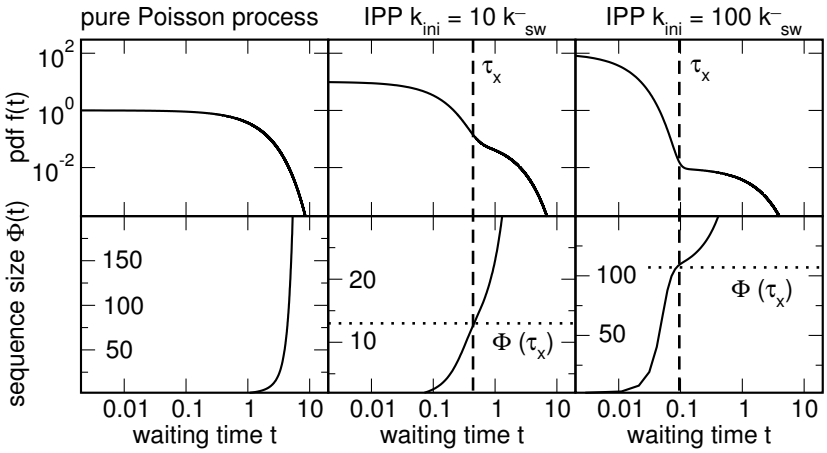


Figure 2.2: Theoretical analysis of the minimal burst model (Fig. 2.1). Columns correspond to different parameterizations. First row: the waiting time PDF for a pure Poisson process and for two IPPs with different burst characteristics (Eq. 2.1). The vertical dashed lines indicate the threshold of time scale separation τ_X. Second row: sequence size function (Eq. 2.4). The Φ evaluated at τ_X yields the burst size β (horizontal dotted lines).

probability to observe a waiting time resulting from the fast and the slow process is identical (two terms of Eq. 2.1 are equal). Two time scales are observable if $k_{ini} > k_{sw}^{+} - k_{sw}^{-}$; only then $\tau_X > 0$ (Eq. 2.18 in Materials and Methods).

The mean waiting time determined from Eq. 2.1 corresponds to the macroscopic estimate (Eqs 2.12 in Materials and Methods),

$$\langle t \rangle = \int_0^\infty t\, f_X(t) dt = \frac{w_1}{r_1} + \frac{w_2}{r_2} = \tau_{ini} \left(\frac{\tau_{on}}{\tau_{on} + \tau_{off}} \right)^{-1} \tag{2.2}$$

The inverse of this equation equals the mean arrival rate, $k_{ini} \langle \mathrm{ON} \rangle$. It has the interpretation of the expected burst size divided by the duration of a single switch cycle. The noise in the waiting time is given by

$$\eta_t^2 = \mathrm{VAR}/\mathrm{AVG}^2 = \sigma_t^2 / \langle t \rangle^2 = 1 + 2\,\beta_e \langle \mathrm{OFF} \rangle^2. \tag{2.3}$$

The second term expresses the deviation of IPP from a Poisson process. It is small if either the *ON* state is short-lived (β_e decreases) or silence periods are negligible.

2.3.2 Measures for characterization of bursts

The size of a burst and the burst duration are relevant burst properties. For biological applications they need to be determined on the basis of a stochastic time series as the mechanism underlying bursts is typically unknown.

The *burst size* β is defined as the mean number of production events not interrupted by a long inactivity period. Inactivity periods (interruptions) occur as often as bursts.

The total number of production events divided by the number of interruptions yields the burst size. In order to determine the time scale of interruptions and hence their number, we define a *sequence size* function $\Phi(\vartheta)$,

$$\Phi(\vartheta) = \frac{\#\text{ total events}}{\#\text{ intervals longer than }\vartheta} = \frac{n_a}{n_a P[X > \vartheta]} = \frac{1}{1 - F_X(\vartheta)} \tag{2.4}$$

where $F_X(\vartheta)$ is the cumulated distribution function (CDF), i.e. $F_X(\vartheta) = \int_0^\vartheta f_X(t)dt$. For a given threshold ϑ, the function yields the sequence size such that events are grouped into sequences interrupted by intervals longer than ϑ. Due to the time scale separation, there is a specific interval ϑ_b for which $\Phi(\vartheta)$ equals the burst size β. The value of ϑ_b can be determined on the basis of the functional dependence of $\Phi(\vartheta)$ as illustrated in Fig. 2.3.

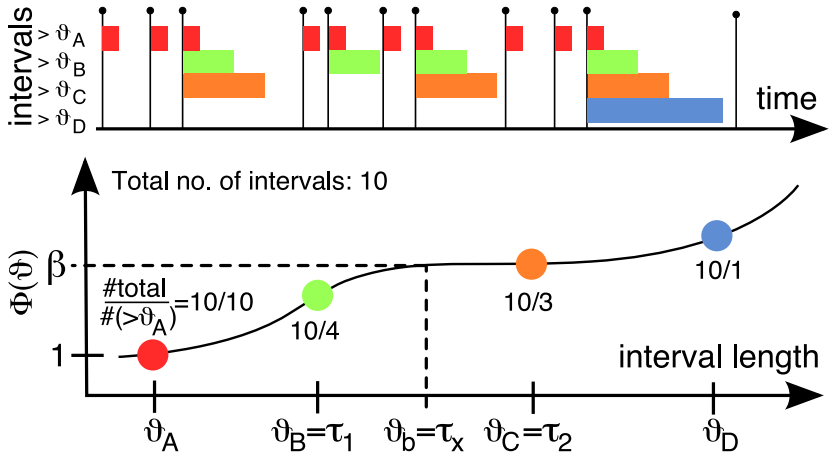


Figure 2.3: Cartoon of a sequence size function Φ. (Top) time series of production events. Horizontal bars denote intervals longer than thresholds $\vartheta_{A,B,C,D}$. (Bottom) For a given ϑ, Φ is constructed by dividing the total number of intervals by the number of intervals longer than ϑ. Time scale separation introduces a regime where ϑ is longer than intervals within bursts but shorter than interruptions between them; a plateau appears. The point of time scale separation τ_X lies in the middle of two inflection points $\tau_{1,2}$ determined from the second derivative of Φ. The value of Φ at τ_X is the burst size β.

As a measure for burst size β we evaluate the sequence size function at the interval of time scale separation, τ_X, which has a straightforward interpretation for the minimal burster from Fig. 2.1. This point lies in the middle of two intervals τ_1 and τ_2 ($\tau_1 < \tau_2$); they correspond to the change of $\Phi(\vartheta)$ from convex to concave to convex as function of ϑ, respectively (Eqs 2.19–2.21).

The burst size β is greater than the expected burst size β_e as the latter excludes the possibility of an *ON* state without a production event. Both measures are approximately equal for a large time scale separation: $\beta = \beta_e + \mathcal{O}(\log k_{ini})$, if $k_{ini} \gg k_{sw}^+$ and $k_{sw}^- \approx k_{sw}^+$. Fig. 2.2 illustrates the PDF and $\Phi(\vartheta)$ for different parameter regimes of the minimal

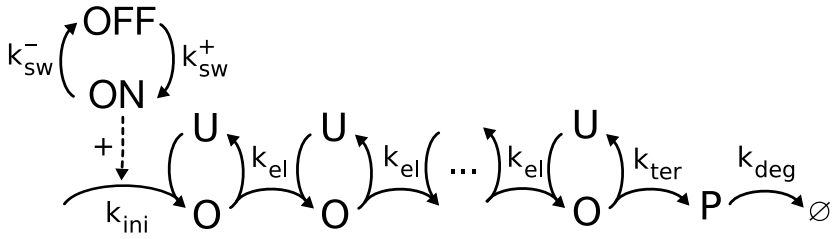


Figure 2.4: Canonical model of macromolecular trafficking along a biopolymer. In the *ON* state of the switch, proteins initiate elongation with a rate constant k_{ini}. Elongation occurs with a rate constant k_{el}. "O" and "U" denote occupied and unoccupied state of the site, respectively. Motors leave the chain with a rate constant k_{ter} and accumulate a product P. The product is degraded with a rate constant k_{deg}.

model. Bursts become more pronounced for high k_{ini} over k_{sw}^- ratios (increased time scale separation). The behavior of the sequence size function for non-exponential waiting times is explored in Section 2.5.5. In short, a gamma-distributed waiting time for the *OFF* to *ON* transition increases the time scale separation. The applicability of the indices is not affected.

Once the burst size is known, the *duration of a burst*, τ_β, can be obtained by multiplying β by the mean waiting time within a burst ($1/r_1$ in the minimal model). Whether the interval is part of a burst can be deduced using the threshold of time scale separation, τ_X (determined on the basis of the waiting time PDF). In addition to burst size, burst significance is important. Bursts loose significance if the interruption period becomes comparable to intervals within a burst. To quantify this we introduce a dimensionless *significance coefficient* $\xi = 1 - \tau_1/\tau_2$, $\xi \in (0, 1)$.

We use all three measures, β, τ_β, and ξ to analyze stochastic time series for more complex schemes. The advantage is that the indices are mechanism-independent. In addition to their unbiased nature, they have a clear mechanistic interpretation for the minimal burster. This property facilitates interpretation of yet unidentified mechanisms giving rise to bursts.

2.3.3 Motor-protein traffic jams along biopolymer chains

Bursts have been observed experimentally for single-cell synthesis of mRNA and protein [Golding *et al.* 2005, Yu *et al.* 2006, Raj *et al.* 2006, Cai *et al.* 2006, Chubb *et al.* 2006] (review [Kaufmann & van Oudenaarden 2007]). For such cases, free energy driven motion of a catalytic motor protein along a biopolymer template is required. Here, we investigate the role of the stochasticity in the initiation of motion and in the motion itself for the observation of bursts at the end of the chain. We will consider different lengths of the polymer and kinetics of initiation and transport.

Fig. 2.4 shows a canonical 1-D macromolecular trafficking model. It contains the switching source, as described in the previous section, followed by sites on the polymer. Each site can be occupied by a single motor, moving forward only, with the elongation rate constant k_{el} (sites per time).

Evidently, motors cannot pass each other and shall collide. We also consider the motor

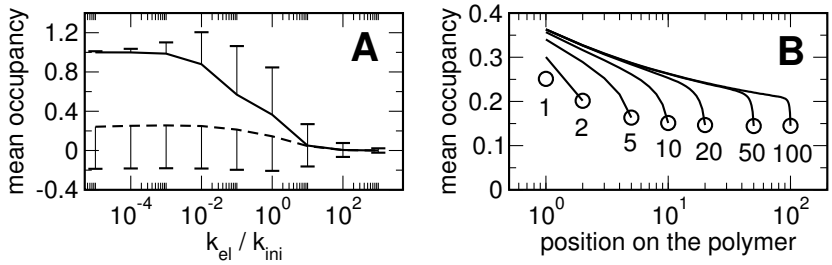


Figure 2.5: Mean site occupancy: (A) at the beginning (solid line) and at the end (dashed) of a 100-site polymer ($k^+_{sw} = k^-_{sw} = 1\,[1/T]$, $k_{ini} = 100k^-_{sw}$); error bars, standard deviation, (B) along 1 to 100-site polymers (dotted), numbers indicate the length, circles mark mean occupancy at the polymer's end; parameters as in (A), and $k_{el} = k_{ini}$.

occupying more than one site (Section 2.5.7.2). Previously we discussed conditions for bursts to emerge at the start of the polymer. Whether bursts are preserved at the end of the chain depends on the characteristics of the stochastic motion.

In Fig. 2.5A we consider a polymer of length 100 sites, preceded by a *bursty* switch with 100 initiations per *ON* state, on average. We plot the mean occupancy of sites at the beginning and at the end of the polymer as function of a dimensionless ratio, k_{el}/k_{ini}. If a motor protein travels many sites between initiations during a single *ON* state ($k_{el}/k_{ini} \gg 1$), its progression is not hampered by collisions. A traffic jam arises at the beginning of the chain if the number of traversed sites during consecutive initiations is small ($k_{el}/k_{ini} \ll 1$). This results in high mean site occupancy at the initial segment of the polymer. Congestion weakens at the end of the polymer because motors have less partners ahead of them. As shown in Fig. 2.5B, the net occupancy gradient along the chain increases with the polymer length.

If collisions are significant, the time scale separation generated at the initiation stage is disrupted. The frequency of P production becomes exclusively determined by the motor protein progression at the end of the chain. The effect intensifies as the length of the polymer increases as illustrated in Fig. 2.6A.

Due to collisions, the rate at which motors leave the polymer becomes smaller than k_{ini}. The mean interval between productions (the inverse of the macroscopic flux) becomes longer for longer chains (panel B). Additionally, the standard deviation becomes comparable to the mean waiting time. The process becomes exponentially distributed and the memory of the state of the switch is lost entirely.

Motor protein collisions disrupt inactivity periods of initiation bursts. As a result, intervals between production events become comparable. In this regime, the burst size β measured at the end of a long polymer is seemingly larger than for a short one (panel C). Therefore, it is necessary to aid the measurement of bursts with the significance index, ξ. As the length of the polymer increases, the significance of bursts diminishes (panel D). Below, we introduce a mechanism that can recover bursts at the output even if no bursts occur at the input.

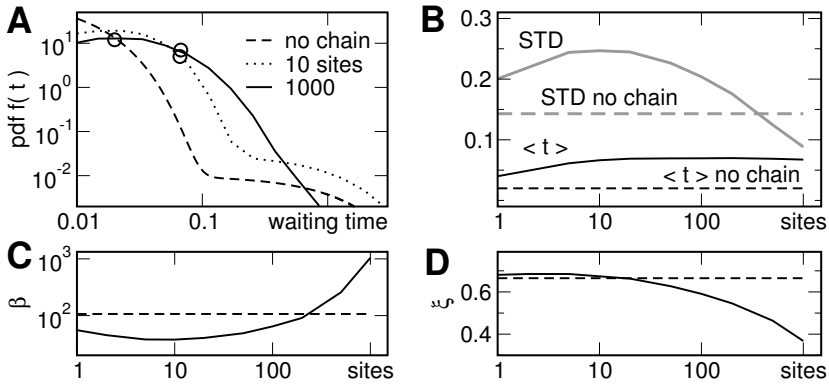


Figure 2.6: Analysis of the canonical model of macromolecular trafficking. Gillespie simulations of at least $1e5$ events, $k_{sw}^{+} = k_{sw}^{-} = 1\,[1/T]$, $k_{el} = k_{ini} = 100k_{sw}^{-}$. The dashed line in all panels marks the minimal burst model (no elongation). (A) The waiting time PDF for 10 and 1000-site polymers. Circles, the mean waiting time. (B) The mean waiting time and its standard deviation, (C) burst size, and (D) burst significance for different chain lengths.

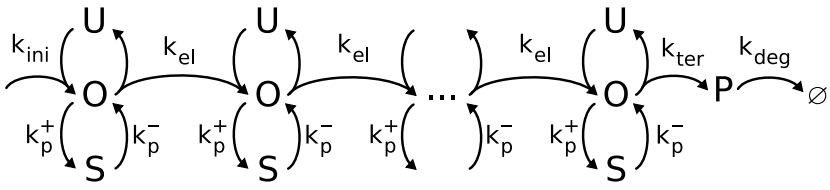


Figure 2.7: Model of macromolecular trafficking with pausing motor proteins. Proteins initiate motion with a fixed rate constant k_{ini}. A site can be unoccupied (symbol 'U'), occupied ('O') by a motor, or occupied by a motor in a paused state ('S'). Motors can pause at a rate k_p^+ at every site. The lifetime of the paused state is $1/k_p^-$. Other parameters are the same as in Fig. 2.4.

2.3.4 Pausing of motor proteins can generate bursts

We shall consider a polymer without a switching source that modulates initiation. Initiation takes place at a fixed rate constant k_{ini}. Such a model has only one time scale, thereby no bursts at initiation can occur. We consider motor protein pausing along the chain as a potential burst-generating mechanism. At every site, a motor can switch at a rate k_p^+ to a paused state that lasts $1/k_p^-$ (Fig. 2.7). This mechanism is known to occur for RNA polymerase [Yin *et al.* 1999, Artsimovitch & Landick 2000, Bar-Nahum *et al.* 2005, Greive & von Hippel 2005, Hatoum & Roberts 2008] and ribosomes [Hayes & Sauer 2003, Sunohara *et al.* 2004, Buchan & Stansfield 2007, Galburt *et al.* 2007, Wen *et al.* 2008].

Pausing of a single motor causes congestion due to its collision with consecutive motors during its dwell time. This allows for the build up of a burst packet. The packet can survive until the end of the chain only if the pausing frequency is low for a given chain length L. If this is not the case, there is a high probability that proteins within the potential

burst will also pause and thus divide the packet (Fig. 2.8A, curves for $k_p^+ = 100\,[1/T]$). Another requirement for bursting concerns the lifetime of the paused state. If too short, compared to the initiation rate k_{ini} and the elongation rate k_{el}, the consecutive proteins will not catch up (Fig. 2.8A, curves for $k_p^- = 100\,[1/T]$). Time scale separation in the waiting times at the end of the chain, and hence bursts, appear only when motors do not pause too frequently during the elongation, and when the paused state is sufficiently long-lived (Fig. 2.8A, solid curve for $k_p^+ = k_p^- = 1\,[1/T]$).

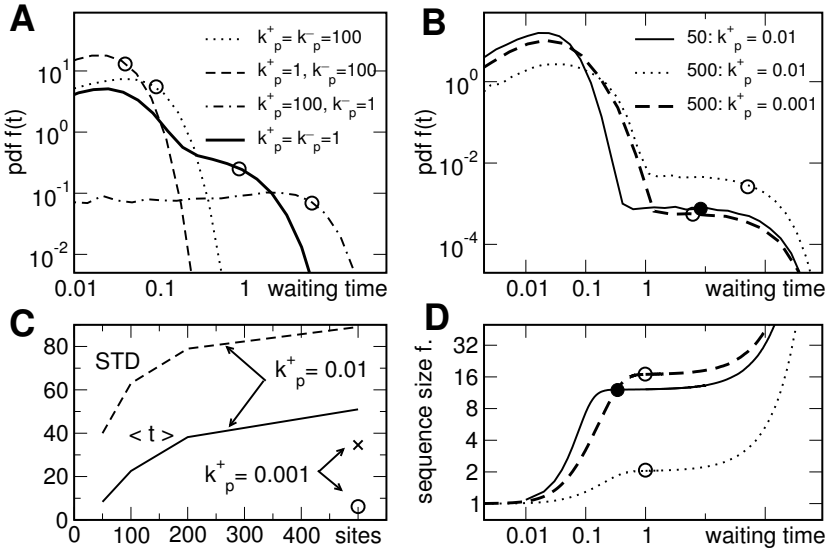


Figure 2.8: The effect of chain length and pausing parameters on waiting time statistics for the model with pausing. Data obtained from Gillespie simulations of 1e6 events. Common parameters: $k_{ini} = 100\,[1/T]$, $k_{el} = k_{ter} = k_{ini}$. (A) The appearance of the time scale separation due to the pausing of proteins. Chain length: 100 sites. Bursts arise when (i) k_p^+ allows for only few pauses during the elongation and (ii) $1/k_p^-$ is long enough for many initiations to occur (solid line). Circles, the mean waiting time. (B) Waiting time PDF for chain lengths 50 and 500 sites. The lifetime of the paused state is fixed: $1/k_p^- = 100\,[T]$. At a pausing rate constant $k_p^+ = 0.01\,[1/T]$, the increase in the chain length increases the probability of multiple pauses during the progression: bursts disappear (dotted line). Reduction of k_p^+ to $0.001\,[1/T]$ recovers bursts (dashed line). Circles, $\langle t \rangle$. (C) The mean waiting time and its standard deviation as function of chain length. Lines: $k_p^+ = k_p^- = 0.01\,[1/T]$. Symbols at $L = 500$ correspond to the dashed line in (B): $k_p^+ = 0.001\,[1/T]$. (D) Sequence size function for polymers as in B. Circles, β.

The length of the biopolymer chain affects the statistics of bursts as well. The addition of sites increases the probability that a single motor pauses a number of times during its progression. Thereby, such a motor destroys the burst it was part of. The effect is equivalent to an increase in k_p^+ at a fixed L. As a result, the waiting times lack the short time scale originating from frequent product initiation. Instead, they are dominated by the lifetime of the paused state. The mean waiting time $\langle t \rangle$ and its standard deviation increase with increasing L, and the time scale separation becomes less pronounced (Figs. 2.8B and

C). Since burst size β decreases, bursts tend to disappear for longer chains. They can always be recovered by decreasing the probability of a single motor protein to pause many times during its progression. As an illustration we shall change the pausing rate for the longest chain considered, $L = 500$. If we set k_p^+ 10 times smaller than the value used for $L = 50$, the time scale separation is recovered and β increases (Figs. 2.8B and D, dashed line). This indicates that pausing may prevent or promote bursts depending on the properties of the biopolymer.

2.3.5 Aggregative behavior of multiple burst-generators

Statistics of bursts change when they arise from the simultaneous activity of a number of independent burst-generating mechanisms. In biological terms, this superposition may describe the transcription of an mRNA from independent copies of a gene or the translation of protein from a number of mRNAs. Here we focus on the extension of the simple model of bursts to a superposition of many independent interrupted Poisson processes.

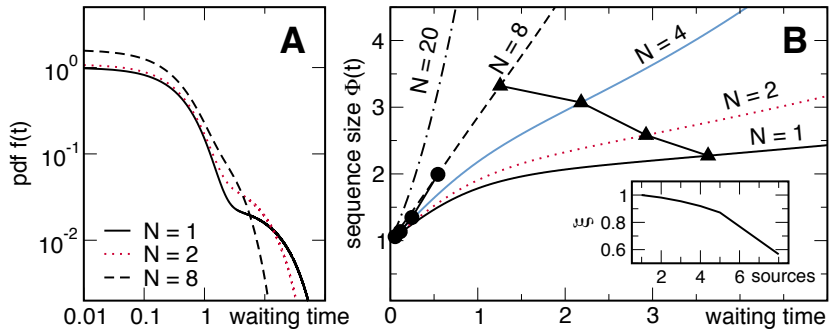


Figure 2.9: Superposition of independent bursters ($k_{sw}^+ = 0.1\,[1/T], k_{sw}^- = 1\,[1/T], k_{ini} = k_{sw}^-$). (A) Analytical waiting time PDF as a function of the number of burst sources. The curve becomes almost exponential for 8 sources. (B) Analytical sequence size function for 1 to 20 sources. The two roots of its second derivative (solid circles and triangles) vanish for more than 8 sources. (Inset) burst significance as function of the number of sources.

We choose parameters such that a single burster initiates 1 product per *ON* state ($k_{sw}^+ = 0.1\,[1/T]$, $k_{sw}^- = 1\,[1/T]$, $k_{ini} = k_{sw}^-$). The resulting burst size is small, $\beta \approx 2.3$. The waiting time PDF almost completely looses its double-exponential character for more than 8 sources (Fig. 2.9A). Although the burst size increases for many sources (panel B), their significance ξ diminishes until the sequence size function Φ becomes always convex and ξ can no longer be evaluated (inset of panel B). This indicates that clustering of events into bursts no longer occurs: waiting times follow a single-exponential distribution. Analytical results for the pooled mechanism, the PDF and the Poissonian behavior in the limit of many independent IPPs, can be found in Section 2.5.6. These results are particularly insightful for the conditions for bursts in the protein level. A burst of a few mRNA transcripts does not necessarily cause a protein burst. It depends on temporal correlation between translation that occurs independently on each transcript.

2.4 Discussion

For most applications the exact burst-generating mechanism is not known or too complex to handle analytically. To overcome this problem we defined three measures for the characterization of bursts: *burst size*, *duration*, and *significance*. We stress the usefulness of the significance measure, as large bursts can arise at a negligible time scale separation. The indices have a transparent interpretation for the system in Fig. 2.1 and allow for model-independent analysis of more complicated mechanisms. Additionally, we offer a rigorous method to obtain the indices from stochastic time series. We applied those measures to investigate the influence of the stochastic motion of motor proteins along a biopolymer on bursts of product release at the end of the chain, e.g. of protein, cargo-vesicles or mRNA. Our study was inspired by the experimental discovery of bursts in transcription and translation [Golding *et al.* 2005, Yu *et al.* 2006, Raj *et al.* 2006, Cai *et al.* 2006, Chubb *et al.* 2006]. We found that bursts at the input of the chain tend to be smoothed out by longer chains due to congestion of motor proteins. Due to collisions time scales within and between bursts become comparable causing the burst size to be larger but of less significance. Hence, the stationary output flux decreases with the length of the chain. At a fixed initiation rate, bursts can emerge due to pausing of motor proteins.

We discussed two mechanisms that could give rise to bursts in production. Bursty transcription initiation was the first one. The second considered motor protein pausing as a source of bursts. How to distinguish these two mechanisms experimentally and how can existing data be interpreted in this light? We shall consider this in more detail for transcription. Promising mechanisms for initiation-induced bursts in mRNA production are genes controlled by strong repressors occasionally leaving the promotor to allow a few RNA polymerases to initiate transcription as suggested in [Golding *et al.* 2005, Yu *et al.* 2006]. On average, every τ_{ini} minutes an RNA polymerase initiates elongation if the gene is in the *ON* state, i.e. in the presence of an activating transcription factor or in the absence of a repressor. During this time, polymerase traverses $k_{el}\,\tau_{ini}$ nucleotides. If no significant congestion occurs along the DNA, the mean waiting time $\langle t\rangle$ for polymerases at the end of the chain is proportional to τ_{ini} (Eq. 2.2); the time scale of initiation bursts (if present) is preserved at the end of the chain. If motors collide during their progression this relationship is destroyed. Fig. 2.10 illustrates this for switch parameters corresponding to measurements by Golding *et al.* [Golding *et al.* 2005]; $\tau_{on} = 6$, $\tau_{off} = 37$ min. This figure displays the dependency of $\langle t\rangle$ on the initiation time τ_{ini} for the minimal, canonical trafficking, and detailed model (with initiation switch and pausing). Assuming that the experimental system studied by Golding *et al.* is in the regime where collisions do not disrupt the proportionality between $\langle t\rangle$ and τ_{ini}, their measured waiting time within a burst corresponds to $\tau_{ini} = 2.5$ min. For these parameters, Fig. 2.10 indicates that most of the control on the waiting time is exerted by the initiation; pausing properties do not affect $\langle t\rangle$. The calculated burst size (3.6) is close to the experimental result (≈ 2.2).

How would pausing of RNA polymerases and altered initiation rates (different genes) change this picture? In Fig. 2.10, we investigate this using realistic pausing parameters. The panel for burst size shows that initiation bursts are reduced by pausing. Spontaneous collisions and hence deviation of $\langle t\rangle$ from linear dependence (Eq. 2.2) occur only at high initiation rates. The discrepancy is larger if pausing is considered. Therefore, pausing is a more plausible mechanism for such deviations in real biological systems where τ_{ini} exceeds

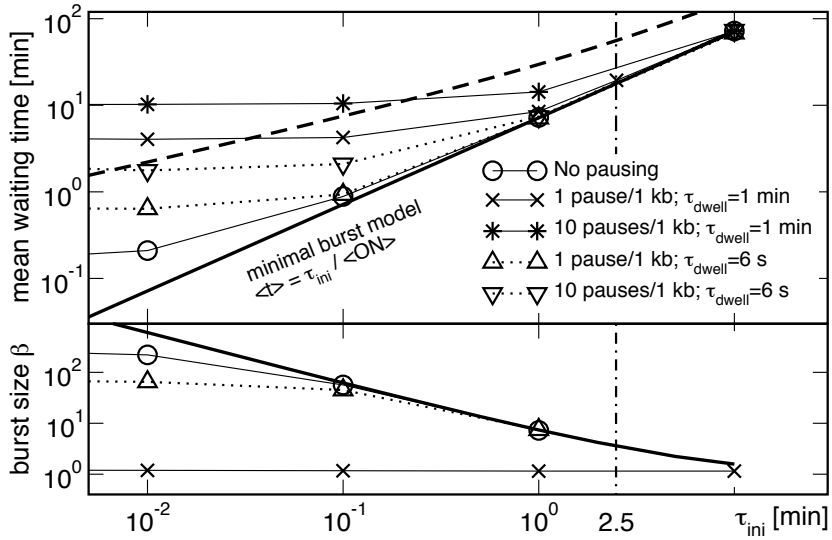


Figure 2.10: Detailed model of elongation: mean waiting time $\langle t \rangle$ for mRNA production as function of initiation intervals $\tau_{ini} = 1/k_{ini}$. The gene consists of 1000 nucleotides, RNA polymerase occupies 50 nt [Voliotis *et al.* 2008]. Parameters of the initiation switch are $\tau_{on} = 6$ and $\tau_{off} = 37$ min [Golding *et al.* 2005]. Elongation occurs at 50 nt/s [Greive & von Hippel 2005]. Means were obtained from Gillespie simulations of at least $1e4$ events. Solid line without symbols denotes the mean, and the dashed line one standard deviation plus the mean for the minimal burst model (no elongation). The vertical line indicates the mean waiting time of mRNA in the experiment of Golding *et al.* [Golding *et al.* 2005]. Deviation from the solid line results from collisions of RNA polymerases.

0.1 min. This makes pausing a potent target for regulation, e.g. by NusA and NusG, in accordance to recent experiments [Artsimovitch & Landick 2000, Bar-Nahum *et al.* 2005, Greive & von Hippel 2005, Hatoum & Roberts 2008]. Voliotis *et al.* [Voliotis *et al.* 2008] have postulated collision-induced bursts caused by backtracking of RNA polymerases. Fig. 2.10 offers a convenient method to determine whether collisions induced by pausing or spontaneous are additional controlling processes besides initiation.

Which genes are likely to generate bursts? Strong repressors could induce bursts according to bursty initiation as described above. In experimental studies this mechanism yielded a small burst size [Golding *et al.* 2005, Yu *et al.* 2006]. If the lifetime of mRNAs and proteins is shorter than the *OFF* period, transient bursts ("puffs") are produced. Savageau's demand principle [Savageau 1998] predicts that infrequently used genes are regulated by repressors to prevent them from accumulating mutations. Such systems should be susceptible to bursts, e.g. repressor-regulated operons of prokaryotic signaling networks.

Highly-activated genes under the control of an activating transcription factor can give rise to bursts by the pausing mechanism (Fig. 2.8). Such systems are predominantly in the *ON* state (e.g. by way of enhancer dependence) and have an approximately con-

stant and high initiation rate. This should make them prone to bursts induced by pausing. Such a mechanism has not been experimentally observed so far. *E. coli*'s *rrn* genes (coding for ribosomal RNA) would be likely candidates; for they are among the genes with highest expression activity [Bremer *et al.* 2003]. In mammalian systems, poised and paused polymerases occur often and could underlie the experimentally observed bursts [Chubb *et al.* 2006, Wolf *et al.* 1995, Price 2008].

Typically, products are not synthesized in a single macromolecular process. We showed that if bursts are generated independently, the resultant burst tends to loose significance with an increasing number of sources. Hence, mRNA bursts are more likely to occur than protein bursts, as protein is typically produced from a few transcripts simultaneously. Depending on the significance and size of the burst generated by a single catalytic process, this rules out significance of bursts in, for instance, metabolism where the copy number of catalytic proteins is thousands.

Bursts are a powerful mechanism to generate cellular heterogeneity. Key processes such as transcription and translation are particularly prone to generate bursts. How biological systems manage to function reliably in the presence of bursts, whether they actively suppress them or control their characteristics remains to be experimentally shown. Exact burst properties and their functional consequences depend on the relative times scales of initiation, elongation, and termination processes. The analytical theory we presented gives insight into generic burst properties. The proposed burst indices allow for quantitative studies of specific systems and their comparison.

Acknowledgments

Helpful input has been provided by J.G. Blom, J. Vidal Rodríguez, H. Westerhoff and P.R. ten Wolde. This work was supported by the Netherlands Organisation for Scientific Research (NWO), Computational Life Sciences project NWO-CLS-635.100.007. FJB thanks the Netherlands Institute for Systems Biology and NWO for financial support. Finally, we thank the anonymous reviewers for their very insightful comments.

2.5 Materials and methods

2.5.1 Statistics of the arrival process

We are interested in intervals between product arrivals for the interrupted Poisson process (IPP). The statistic is independent of the degradation rate k_{deg}. Derivation of the waiting time probability density function is in fact a first-passage time problem [Redner 2001]. The original derivation of the *pdf* for the IPP process was obtained by Kuczura [Kuczura 1973, Milne 1982]. Here we obtain the same result using a much simpler approach.

An interval t between two arrivals is the time to release the $(n+1)^{th}$ product at time t given that the n^{th} product arrived at the initial time $t = 0$. In other words, we are looking for a probability density of the occurrence of the *first* event in the infinitesimal interval $(t, t + \Delta t)$, if we set the initial time $t = 0$ at the moment of the previous arrival.

We denote the probability that there were k arrivals in time interval $(0, t]$, given that an arrival occurred at $t = 0$ as $p_{k,\,\varepsilon}(t)$. Variable $\varepsilon = \{0, 1\}$ describes whether at the

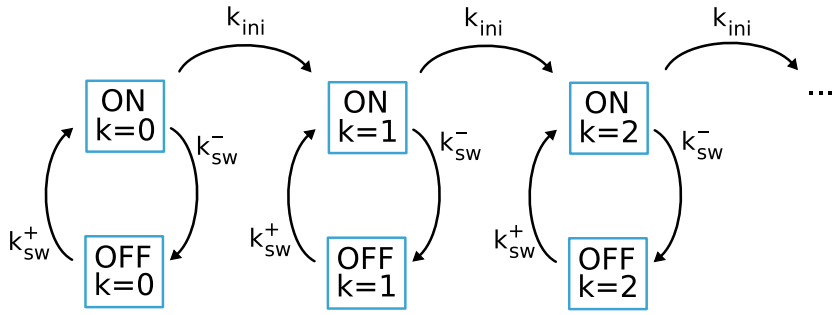


Figure 2.11: Scheme of discrete states in the IPP. A single box represents a state $p_{k,\,\varepsilon}\,(t)$ where the switch is in the *ON* ($\varepsilon = 1$) or *OFF* state ($\varepsilon = 0$). Index k denotes the number of arrivals in time interval $(0, t)$, given that an arrival occurred at $t = 0$.

particular time t the switch is in the *OFF* or *ON* state, respectively. The following set of master equations governs the evolution of the system between discrete states (Fig. 2.11):

$$\begin{aligned}
\frac{d}{dt}p_{0,\,1}\,(t) &= k^{+}_{sw}\,p_{0,\,0}\,(t) - \left(k_{ini} + k^{-}_{sw}\right) p_{0,\,1}\,(t) \\
\frac{d}{dt}p_{k,\,1}\,(t) &= k^{+}_{sw}\,p_{k,\,0}\,(t) - \left(k_{ini} + k^{-}_{sw}\right) p_{k,\,1}\,(t) + k_{ini}\,p_{k-1,\,1}\,(t)\,, \quad k = 1, 2\ldots \\
\frac{d}{dt}p_{k,\,0}\,(t) &= -k^{+}_{sw}\,p_{k,\,0}\,(t) + k^{-}_{sw}\,p_{k,\,1}\,(t)\,, \quad k = 0, 1\ldots
\end{aligned} \tag{2.5}$$

with the initial condition $p_{0,\,1}\,(t = 0) = 1$.

The first-passage probability includes all *incoming* fluxes due to transitions that drive the system from the initial condition to state $p_{1,\,1}\,(t)$ for the first time:

$$f_{1,1}(t) = k_{ini}\,p_{0,\,1}\,(t) \tag{2.6}$$

Note, we do not add the flux $k^{+}_{sw}\,p_{1,\,0}\,(t)$. The system visits the state $p_{1,\,1}\,(t)$ prior to arriving at the state $p_{1,\,0}\,(t)$.

Solving the problem, requires only two master equations for $k = 0$, which is a simple homogeneous system of linear ODEs. The solution, obtained in a standard manner, is the sum of two exponentials since there are only two variables (eigenvalues).

We take a different approach to illustrate the usage of the Laplace transform and its asymptotic properties. The Laplace transform allows to obtain a set of algebraic equations:

$$\begin{aligned}
s\tilde{p}_{0,\,1}\,(s) - p_{0,1}(t = 0) &= k^{+}_{sw}\,\tilde{p}_{0,\,0}\,(s) - \left(k_{ini} + k^{-}_{sw}\right)\tilde{p}_{0,\,1}\,(s) \\
s\tilde{p}_{0,\,0}\,(s) - p_{0,0}(t = 0) &= k^{-}_{sw}\,\tilde{p}_{0,\,1}\,(s) - k^{+}_{sw}\,\tilde{p}_{0,\,0}\,(s) \\
\tilde{f}_{1,1}(s) &= k_{ini}\,\tilde{p}_{0,\,1}\,(s)
\end{aligned} \tag{2.7}$$

Solving for $\tilde{p}_{0,\,1}\,(s)$, we obtain:

$$\tilde{f}_{1,1}(s) = \frac{k_{ini}\,(s + k^{+}_{sw})}{s\left(s + k^{+}_{sw} + k^{-}_{sw} + k_{ini}\right) + k^{+}_{sw}k_{ini}} \tag{2.8}$$

where $\tilde{f}_{1,1}(s)$ is the Laplace transform of the the first-passage time probability density. The inverse transform of 2.8 allows to obtain an explicit expression in the time domain, i.e. $f_{1,1}(t)$.

Before we spell the final result we need to realize again how the first-passage time *pdf*, we are about to calculate, relates to the waiting time *pdf* we are interested in. The former gives the normalized frequency histogram of times t of the first occurrence of the state $p_{1,1}(t)$ given that initially the system was in state $p_{0,1}(t=0)$. It is exactly the same histogram when intervals between successive arrivals are registered.

The waiting time *pdf* $f_X(t) \equiv f_{1,1}(t)$ for the duration of the interarrival time X to be within $(t, t+\Delta t)$ is a weighted sum of two exponential functions. Its CDF $F_X(t)$ and the density function itself are:

$$F_X(t) = Pr[X \le t] = w_1 \left(1 - e^{-r_1 t}\right) + w_2 \left(1 - e^{-r_2 t}\right) \tag{2.9a}$$

$$f_X(t) = \frac{Pr[t < X < t + \Delta t]}{\Delta t} = \frac{dF_X(t)}{dt} = w_1\, r_1\, e^{-r_1 t} + w_2\, r_2\, e^{-r_2 t} \tag{2.9b}$$

where rates $r_{1,2}$ and weights $w_{1,2}$ are:

$$r_{1,2} = \frac{1}{2}\left(k_{sw}^{+} + k_{sw}^{-} + k_{ini} \pm \sqrt{(k_{sw}^{+} + k_{sw}^{-} + k_{ini})^2 - 4 k_{ini} k_{sw}^{+}}\right), \quad r_1 > r_2 \tag{2.10a}$$

$$w_1 = \frac{k_{ini} - r_2}{r_1 - r_2}, \quad w_1 \in (0,1) \tag{2.10b}$$

$$w_2 = 1 - w_1 \tag{2.10c}$$

2.5.2 The limit of a large time scale separation

It is instructive to calculate the behavior of the coefficients $r_{1,2}$ and w_1 for large time scale separation between the switching rates and the rate of production for the simple model. We assume that the rate constant k_{ini} is much larger than k_{sw}^{+} and k_{sw}^{-}. If we substitute $k_{ini} = 1/\varepsilon$, and expand around small ε, parameters of the interarrival time *pdf* obtained in the previous section amount to:

$$r_1 = k_{sw}^{-} + k_{ini} + \frac{k_{sw}^{-}\, k_{sw}^{+}}{k_{ini}} + \mathcal{O}(\frac{1}{k_{ini}^2}) \tag{2.11a}$$

$$r_2 = k_{sw}^{+} - \frac{k_{sw}^{-}\, k_{sw}^{+}}{k_{ini}} + \mathcal{O}(\frac{1}{k_{ini}^2}) \tag{2.11b}$$

$$w_1 = 1 - \frac{k_{sw}^{-}}{k_{ini}} + \mathcal{O}(\frac{1}{k_{ini}^2}) \tag{2.11c}$$

2.5.3 Moments of the first-passage time *pdf*

The mean, variance and noise in the arrival process can be obtained from the coefficients of Taylor-expanded first passage time *pdf* in Laplace domain, $\tilde{f}_{1,1}(s)$ (Eqs. 2.8 and A.6):

$$\langle t \rangle = \int_0^\infty t\, f_X(t)\, dt = \frac{w_1}{r_1} + \frac{w_2}{r_2} = \frac{k_{sw}^{+} + k_{sw}^{-}}{k_{sw}^{+}} \cdot \frac{1}{k_{ini}}, \tag{2.12a}$$

$$\sigma_t^2 = \langle t \rangle^2 + \frac{2\, k_{sw}^{-}}{{k_{sw}^{+}}^2\, k_{ini}}, \tag{2.12b}$$

$$\eta_t^2 = \frac{\sigma_t^2}{\langle t \rangle^2} = 1 + \frac{1}{\langle t \rangle^2} \cdot \frac{2\, k_{sw}^{-}}{{k_{sw}^{+}}^2\, k_{ini}} = 1 + \frac{2\, k_{sw}^{-}\, k_{ini}}{(k_{sw}^{+} + k_{sw}^{-})^2}. \tag{2.12c}$$

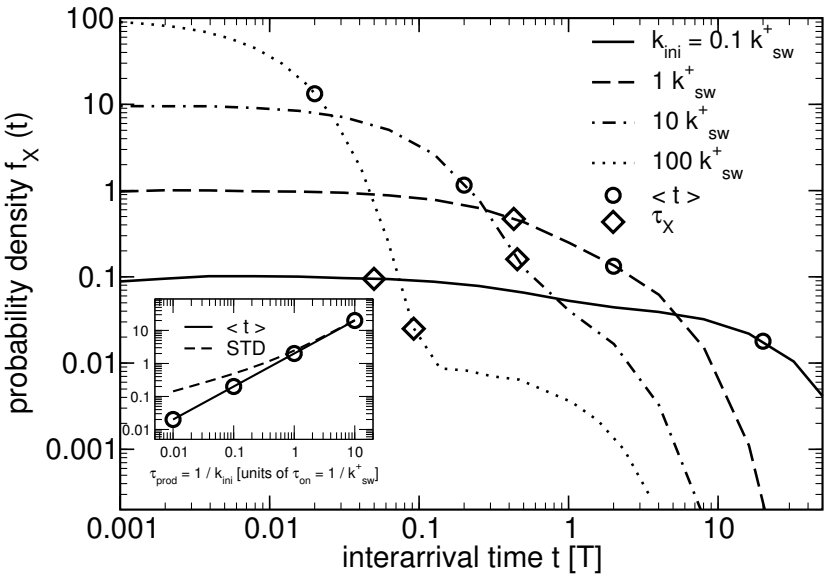


Figure 2.12: Plot of the analytical waiting time distribution $f_X(t)$ (Equation 2.9) for different initiation rate constants k_{ini}. Empty circles denote the average interarrival time, X's denote the point of time scale separation, τ_X. Inset: mean and standard deviation in interarrival times. Note: a power-law function, ax^α, is a straight line in log-log coordinates; if k^+_{sw} equals k^-_{sw}, both mean (Eq. 2.12a) and standard deviation (Eq. 2.12b) depend as $1/k_{ini}$. The nonlinear behavior of variance for $k_{ini} \gg k^-_{sw}$ indicates departure of the IPP from the pure Poisson process.

The inverse of $\langle t \rangle$, the mean arrival rate has a simple interpretation. It is the probability of the *ON* state, $k^+_{sw}/(k^+_{sw} + k^-_{sw})$, times the initiation rate k_{ini}.

For a pure Poisson process the expression for variance (Equation 2.12b) consists of the first term only. The second term exceeds the first one for large values of the initiation rate constant k_{ini}, since it is only linearly proportional to $1/k_{ini}$. In this case the double-exponential character of the interarrival time distribution is very pronounced (Fig. 2.12).

Changing the *ON* and *OFF* switching rate constants such that the average $\langle ON \rangle$ remains the same, does not affect the mean interarrival time. However, higher moments change. If the rate of switching between the two states increases (while the ratio $k^+_{sw}/k^-_{sw} = const$), noise, and similarly variance, approaches values characteristic of a pure Poisson process.

2.5.4 Quantitative characterization of bursts

The *burst size* β is an important quantity from a practical point of view. It also provides information about the *burst duration*, τ_β, if the mean interarrival time within a burst is known. The latter is simply r_1, the inverse of the rate describing the *fast* exponential in the interarrival time distribution (Eq. 2.9). We estimate β by evaluating the *sequence size function* $\Phi(\vartheta)$ at the point of the time scale separation τ_X. Further in this paragraph, we shall derive that for the minimal model this point lies in the middle of the concave region.

This region is precisely where the count of events greater than t suddenly levels off as the threshold interval ϑ increases. The behavior is clearly apparent in the Fig. 2 in the main text, for large time scale separation.

2.5.4.1 Burst size

If the process under study is a pure IPP, relatively simple analytical expressions can be found. The intersection of the two exponentials describes where the separation between two characteristic time scales occurs. For the minimal burst-generating model it equals:

$$\tau_X = \frac{1}{r_1 - r_2} \log\left(\frac{r_1}{r_2} \cdot \frac{w_1}{w_2}\right). \tag{2.13}$$

The value of the sequence size function evaluated at threshold interval equal to τ_X is what we consider the *burst size* β:

$$\beta \equiv \Phi(\tau_X) = \left[w_1 \left(\frac{r_2}{r_1} \cdot \frac{w_2}{w_1}\right)^{\frac{r_1}{r_1 - r_2}} + w_2 \left(\frac{r_2}{r_1} \cdot \frac{w_2}{w_1}\right)^{\frac{r_2}{r_1 - r_2}} \right]^{-1}. \tag{2.14}$$

In order to obtain the behavior for large k_{ini} we need to substitute parameters $r_{1,2}$, $w_{1,2}$ with Eqs. 2.10, and expand around small $\varepsilon = 1/k_{ini}$. We therefore obtain the leading order of the burst size:

$$\beta \equiv \Phi(\tau_X) = k_{ini}/k_{sw}^{-} + \mathcal{O}(\log k_{ini}) \tag{2.15}$$

An example of the leading order behavior for numerical parameters is given in Fig. 2.13. The term k_{ini}/k_{sw}^{-} has a simple interpretation: it is the average number of initiations per *ON* state assuming that production events occur during every active state. We shall refer to it as the *expected* burst size, β_e.

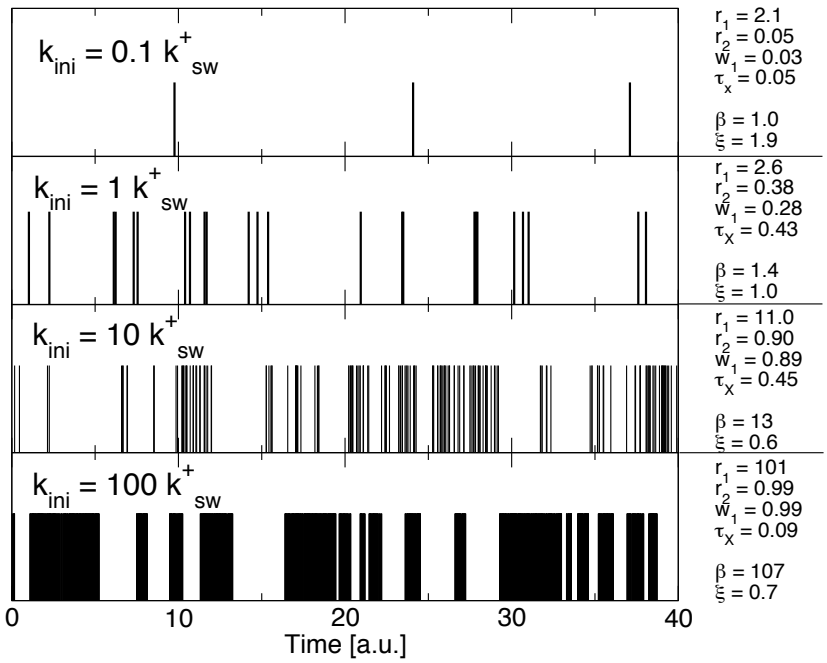


Figure 2.13: Sample time traces of production events obtained from Gillespie simulations. Kinetic parameters are the same as those used in Fig. 2.12 for plotting four interarrival time distributions ($k^+_{sw} = k^-_{sw} = 1\,[1/T]$). Next to the plots we provide numerical values of: rates r_1, r_2 and weight w_1 (Eqs. 2.10), the burst significance ξ (Eq. 2.22), and the sequence size function Φ evaluated at the time scale separation point, τ_X (Eq. 2.13), which estimates the burst size β (Eq. 2.14). In the minimal model waiting times for a production event are drawn from the exponential distribution (63% of the intervals are shorter than the mean). Hence, even for a small number of initiations per *ON* state some events are clustered; short intervals are more frequent. Clusters consisting of many arrivals appear as k_{ini} increases. Note that ξ is larger than 1 in the upper panel, which implies that this measure cannot be applied in this case. Negativity of the first of the roots of Φ is the culprit. Section 2.5.4.5 explains this issue in more detail.

2.5.4.2 Burst duration

The *burst duration*, τ_β, is the burst size multiplied by the mean interarrival time within a burst. A measure describing the cutoff between intervals belonging to a burst and those which are assumed to be interruptions, is given by τ_X (Equation 2.13). In order to compute the mean of the fast time scale one needs to average over those intervals which are shorter than τ_X. For the simple model it comes down to averaging over part of the interarrival time *pdf* and normalizing it to 1:

$$\langle \tau_{ini} \rangle = \frac{1}{F_X(\tau_X)} \int_0^{\tau_X} t \cdot f_X(t) dt. \tag{2.16}$$

In the first order expansion, the above amounts to $1/r_1$, the inverse of the fast time scale in Eq. 2.9. For large k_{ini} the above reduces to $1/k_{ini}$, which is the mean time between

production events during the *ON* state. Therefore, in this limit burst duration amounts to:

$$\tau_\beta = \beta \cdot \langle \tau_{ini} \rangle \to \frac{k_{ini}}{k_{sw}^-} \cdot \frac{1}{k_{ini}} = \tau_{on}. \tag{2.17}$$

2.5.4.3 The existence of time scale separation

If we substitute coefficients $r_{1,2}$, $w_{1,2}$ in the equation for τ_X with expressions given in Eq. 2.10, we obtain a simple condition for $\tau_X > 0$ expressed in terms of kinetic parameters:

$$k_{sw}^+ - k_{sw}^- < k_{ini}. \tag{2.18}$$

For $\tau_X \leq 0$, the interarrival time distribution $f_X(t)$ is still double-exponential on a physically unrealistic domain where time is negative. Intervals that occur in reality are drawn from the *slow* single-exponential waiting times distribution, and the production (arrival) process is purely Poissonian. For $\tau_X = 0$, the sequence size function equals 1.

2.5.4.4 Concave region of the sequence size function

In the next step, we show the relation between τ_X and roots of the second derivative of $\Phi(\vartheta)$, τ_1 and τ_2. The roots equal:

$$\begin{aligned} \tau_{1,2} &= \frac{1}{r_1 - r_2} \log \left[\frac{1}{2\, r_2^2} \frac{w_1}{w_2} \left(\underbrace{r_1^2 - 4 r_1 r_2 + r_2^2}_{\mathcal{A}} \mp \underbrace{(r_1 - r_2) \sqrt{r_1^2 - 6 r_1 r_2 + r_2^2}}_{\mathcal{B}} \right) \right] \\ &= \frac{1}{r_1 - r_2} \log \left[\frac{1}{2\, r_2^2} \frac{w_1}{w_2} \left(\mathcal{A} \mp \mathcal{B} \right) \right]. \end{aligned} \tag{2.19}$$

The mean of the two roots:

$$\frac{\tau_1 + \tau_2}{2} = \frac{1}{2} \frac{1}{r_1 - r_2} \log \left[\frac{1}{4\, r_2^4} \frac{w_1^2}{w_2^2} \left(\mathcal{A}^2 - \mathcal{B}^2 \right) \right], \tag{2.20}$$

where $\mathcal{A}^2 - \mathcal{B}^2 = 4\, r_1^2\, r_2^2$. A straightforward algebraic calculation recovers the Eq. 2.13. Thus, we showed that the separation between time scales in the interarrival time distribution given by τ_X is analogical to the localization of the plateau in the sequence size function, i.e.:

$$\tau_X = \frac{\tau_1 + \tau_2}{2}. \tag{2.21}$$

The estimation of τ_X from the sequence size function is of practical use when the arrival process is not purely an IPP. In such cases the analytical form of function $f_X(t)$ (and likewise $F_X(t)$) is difficult to determine. Nevertheless, one can always perform numerical analysis of $\Phi(\vartheta)$, i.e. compute derivatives and find its roots.

2.5.4.5 Burst significance

In order to fully characterize a burst we need an additional measure which could quantify how the period of activity distinguishes itself from the inactive state. The scaled distance between the two roots $\tau_{1,2}$ calculated in the previous section provides such information:

$$\xi = \frac{\tau_2 - \tau_1}{\tau_2}, \quad \xi \in (0, 1). \tag{2.22}$$

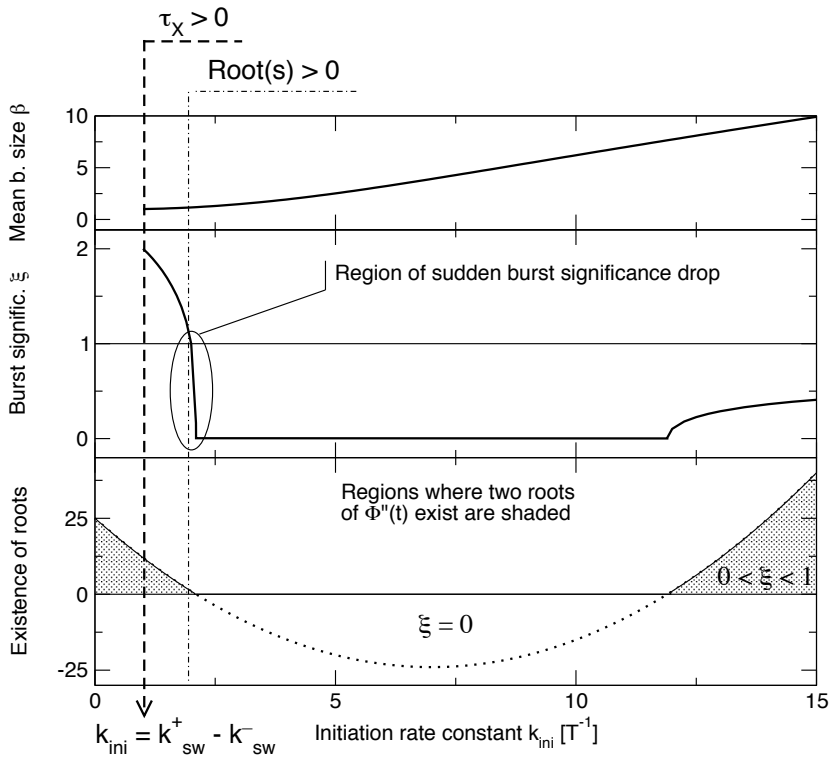


Figure 2.14: Plots of the burst size β (upper), burst significance ξ (middle), and the inequality for existence of two roots $\tau_{1,2}$ (lower) as function of the initiation rate constant k_{ini}. Other parameters: $k^+_{sw} = 3$, $k^-_{sw} = 2$.

Burst significance as defined here is greater than 0 only if the two roots exist. A relatively simple inequality in terms of kinetic parameters of the simple model can be obtained:

$$\left(k^+_{sw} + k^-_{sw} + k_{ini}\right)^2 - 8\, k_{ini}\, k^+_{sw} > 0. \tag{2.23}$$

However, this condition for existence of $\tau_{1,2}$ does not assure that both roots are positive. Even if condition 2.18 for $\tau_X > 0$ is fulfilled, there is a possibility that the smaller root, τ_1, is still below zero. In that case burst significance is larger than 1 which violates our definition of ξ.

An example behavior of all of the quantities we have defined so far is shown in Fig. 2.14. For very mixed time scales (k_{ini} comparable to the switching rates) burst significance displays a non-robust behavior. It decreases very rapidly from 1 to 0 on a small range of k_{ini}, while the burst size increases only slightly. Once the k_{ini} rises above the non-concave region of the $\Phi''(\vartheta)$, two positive roots exists, and the significance remains below 1.

2.5.4.6 How to analyze bursts?

We propose the following procedure for analyzing the stochastic waiting time sequence:

1. create a normalized histogram, preferably using logarithmic binning, in order to obtain the interarrival time *pdf*; a sample MATLAB code can be downloaded from `http://projects.cwi.nl/sic/bursts2008/`,

2. compute the sequence size function $\Phi(\vartheta)$; a sample MATLAB code therein,

3. compute the second derivative of $\Phi(\vartheta)$ and localize its zeros to obtain roots $\tau_{1,2}$,

4. if two positive roots exist, compute the burst significance ξ, and find the value of the sequence size function at $\tau_X = (\tau_1 + \tau_2)/2$,

5. finally, in order to obtain burst duration, τ_β, compute the mean interarrival time within a burst by averaging over intervals lower than τ_X.

2.5.5 Non-exponential waiting time distribution for the switch

In the simple burst model analyzed in Section 2.5.1 we assumed the exponential distribution of waiting times for switching between *ON* and *OFF* states. In principle, such transitions result from the sum of many elementary processes, each of them having the exponential waiting time distribution. Thus, the resulting waiting times for the total process are convoluted, peaked gamma-like distributions (it is a gamma distribution if each of the elementary processes has the same mean). The assumption that waiting times for the whole transition are exponentially distributed is justified only if one of the elementary processes occurs on a much longer time scale than the rest of them. In that case, this particular process accounts for the majority of the area under the peaked distribution.

In Fig. 2.15 we investigate the effect of the number of elementary (irreversible) steps in the *OFF* to *ON* transition on the waiting time distribution of P production and the sequence size function. We compare it to the original IPP process. In all cases we fix the total mean time to complete this transition, i.e. if a transition consists of two steps, the mean waiting time for each of them is half of the time for the total transition. A close inspection of the waiting time *pdf* in panel A shows a more pronounced time scale separation between short and long intervals for the increasing number of elementary steps. An intuitive explanation of this effect is the following. A sharp distribution of waiting times for the *OFF* to *ON* transition results in a peaked distribution of the duration of the inactive state (where no P production takes place). This causes a contraction of the rightmost part of the waiting time *pdf* in the panel A.

The shape of the sequence size function confirms a more pronounced time scale separation for the increasing number of elementary transitions; the plateau becomes flatter (panel B). The burst size decreases slightly and, more importantly, the significance of bursts increases. Both effects are direct consequences of the narrower distribution of inactivity periods. Compared to the minimal IPP model, duration of inactivity periods is centered around the mean value. It is much less probable that two *ON* states are separated by a very short inactive period and then cluster to form one larger burst (the effect likely to occur in case of exponentially distributed waiting times and weak time scale separation). Hence, the burst size decreases to the *expected* value of k_{ini}/k_{sw}^-. Finally, since bursts are separated by inactivity periods drawn from a sharper distribution, their significance increases.

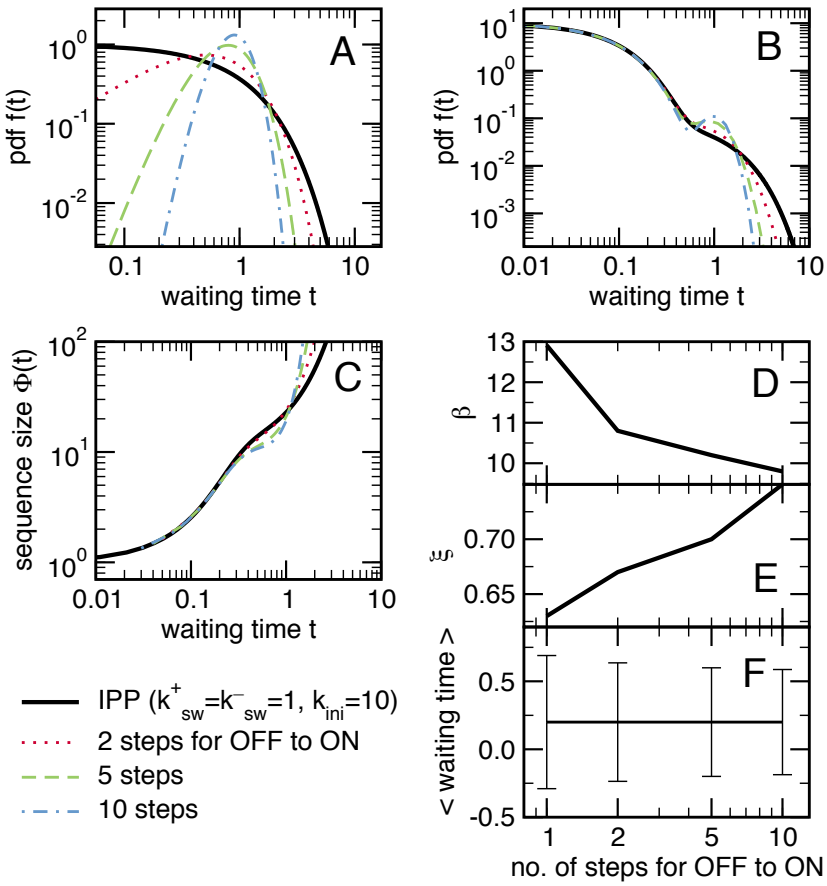


Figure 2.15: The effect of a sharp waiting time distribution for the *OFF* to *ON* transition in the minimal burst-generating model. We vary the number of component steps in the transition, while the total mean waiting time to switch from the *OFF* to *ON* state is fixed to 1[T]. (A) Gamma functions used for the waiting time distribution in the transition. All functions have the same mean, equal 1. (B) The results for the product waiting time *pdf* were obtained from the analytical expression (curve for the IPP), the numerical inverse of the Laplace transform (curve for 2 steps) and from the Gillespie simulation of 1e6 P production events (curve for 5 and 10 steps). The *pdf* is plotted using logarithmic binning. (C) The sequence size function was obtained by a straightforward data analysis in MATLAB. The mean burst size (D), burst significance (E), and the analytical mean and variance (depicted as error bars) (F) as function of the number of steps in the *OFF* to *ON* transition. The mean and variance were obtained by calculating moments of the Laplace transform of the respective expressions.

It can be checked by a straightforward calculation, that the mean waiting time in such a process is independent of the number of elementary steps in the *OFF* to *ON* transition (panel C). At the same time, the variance in the waiting time, and hence the noise (defined as variance over the mean squared), decreases.

2.5.6 Interarrival time CDF in a pool of unsynchronized IPPs

The explicit expression for the interarrival time CDF in the superposition of N independent IPPs [Kamoun & All 1994] is obtained by substituting Equations 2.9 into Equation A.27:

$$F_X^{(S)}(t) = 1 - \left(w_1 e^{-r_1 t} + w_2 e^{-r_2 t}\right) \left(\frac{w_1 r_2 e^{-r_1 t} + w_2 r_1 e^{-r_2 t}}{w_1 r_2 + w_2 r_1}\right)^{N-1} \tag{2.24}$$

The increase in the number of sources causes the overall rate of product arrival in the pooled process to be proportional to N. Therefore, taking the limit $N \to \infty$ of function $F_X^{(S)}$ requires scaling of time by N [Kamoun & All 1994]:

$$\lim_{N\to\infty} F_X^{(S)}\left(\frac{t}{N}\right) = 1 - e^{-\Lambda t} \tag{2.25}$$

As a result, the waiting time probability density function $f_X^{(S)}$ for large number of sources becomes exponential:

$$\lim_{N\to\infty} f_X^{(S)}(t) = \frac{d}{dt}\left(1 - e^{-\Lambda t}\right) = \Lambda e^{-\Lambda t} \tag{2.26}$$

Where Λ is the average arrival rate in the pooled process, i.e. $\langle t \rangle^{(S)}$ (cf. Eq. A.26).

The burst size in the pooled process becomes:

$$\beta = \Phi(\tau_X) = \frac{1}{1 - F_X^{(S)}(\tau_X)} = e^{\Lambda \cdot \tau_X} \tag{2.27}$$

An important conclusion may be drawn based on the above results. The burst significance approaches zero, because Φ becomes a single-exponential without two inflection points. At the same time, the burst size goes to infinity. The overall interarrival time distribution becomes that of a Poisson process.

2.5.7 Progression of motor proteins along the polymer

2.5.7.1 The effect of collisions on $f_X(t)$

Exclusive elongation The simple model of bursts exhibits a double-exponential interarrival time distribution (Fig. 2.16, empty circles) as described earlier in Section 2.5.1. As described in the main article, elongation steps are modeled as forward reactions. Only one motor can occupy a single node which is equivalent to having either concentration 0 or 1 on each node. The addition of elongation has essentially no effect on the product waiting time statistics, if collisions of motor proteins are very infrequent (Fig. 2.16, green dotted line). Due to the elongation traffic, the maximum throughput of the chain reduces as compared to the case without bumping. New proteins cannot begin progression because the sites at the beginning are still not cleared due to collisions (Fig. 2.17, upper plot). As a result, high frequencies (short interarrival times) modulated by k_{ini} disappear from the distribution (Fig. 2.16, blue dashed and cyan dotted-dashed lines). The effect is weaker if the elongation rate k_{el} is faster than the rate of initiation k_{ini}. This corresponds to infrequent collisions as the progression is so fast that the motion of newly added motors

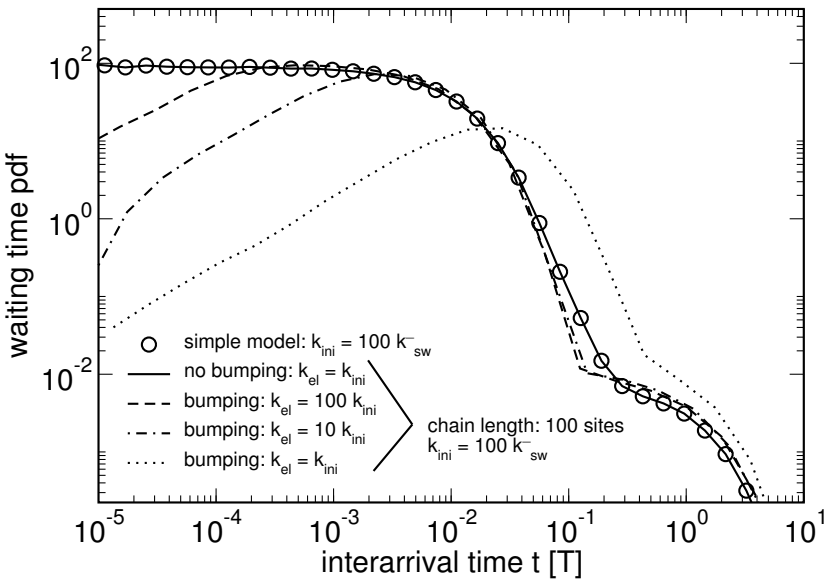


Figure 2.16: The effect of bumping on the interarrival time distribution for a fixed length of the polymer. Results of Gillespie simulations (1e6 events) for the elongation model with bumping and the initiation switch. Chain is 100 sites long, switching rate constants are $k^+_{sw} = k^-_{sw} = 1\,[1/T]$. Initiation rate constant for all elongation models is $k_{ini} = 100k^-_{sw}$, which implies a bursty switch. The simple model of bursts (without elongation) exhibits a double-exponential distribution (empty circles). Elongation with multiple motors allowed to occupy a single node, and thus progression without bumping, has no effect on the distribution, if k_{el} is of the order of k_{ini} (red solid curve). If node occupancy is exclusive (motors collide during progression) high frequencies, i.e. short interarrival times are removed from the distribution. The effect is weak if the elongation (k_{el}) is much faster than the initiation (k_{ini}).

is not inhibited by the molecules ahead of them (Fig. 2.16, green dotted line). The occupancy along the polymer as function of the elongation rate constant, k_{el}, for a bursty switch is shown in Fig. 2.17. A familiar phase-transition-like behavior as function of the order parameter k_{el} is noticeable at the first, most congested node. A sudden drop of the occupancy coinciding with increase of dispersion occurs in the middle of the parameter range [Roussel & Zhu 2006].

Non-exclusive elongation It is also instructive to analyze the behavior of the interarrival time distribution density in case of non-exclusive elongation, i.e. nodes can be occupied by multiple motors and no collisions can occur. In case of absence of the switch, the initiation occurs at a constant rate. The macroscopic flux at the end of the chain is exactly the same as the input flux at the beginning (progression can be freely initiated independent of the occupation of the first site). Similarly, variance does not depend on the chain length. The interarrival time is just a single-exponential and the waiting time to travel the chain of L nodes is given by the Erlang distribution (mean: L/k_{ini}; variance:

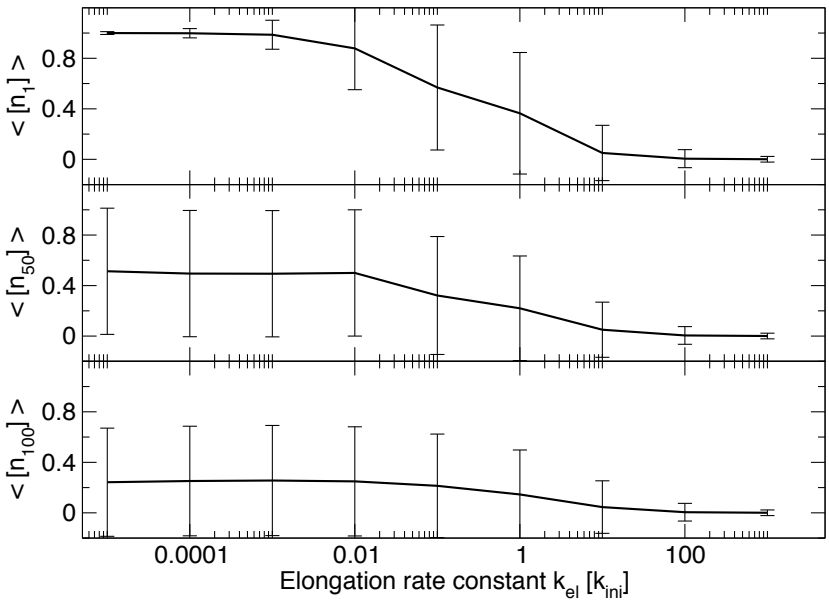


Figure 2.17: The steady-state average occupation at the beginning, in the middle, and at the end of the chain as function of the elongation rate constant k_{el}. Error bars indicate standard deviation. Results were obtained from Gillespie simulations (1e6 events) for the elongation model with bumping and the initiation switch. Chain is 100 sites long, switching rate constants are $k_{sw}^{+} = k_{sw}^{-} = 1\,[1/T]$. The initiation rate constant, $k_{ini} = 100k_{sw}^{-}$, establishes a very bursty switch. The elongation rate constant k_{el} is given in units of the initiation rate constant k_{ini}.

L/k_{ini}^2).

Adding the switch (as in the simple burst model) at the beginning of the chain changes the statistics significantly. There is no concentration gradient as in the case of exclusive progression, and the macroscopic flux is the same throughout the polymer. Although the mean interarrival time is not affected by the chain length, higher moments are. The explanation for the latter is the following. Since nodes are allowed to be occupied by multiple motors, a significant concentration may build up along the chain if the elongation rate constant is smaller than that of the initiation. The elongation rate is in fact the rate at which the concentration of the node is degraded. This remains constant for every node. If the lifetime of the *OFF* state is shorter than the time to reduce the site's concentration significantly, the effect of the silence period is barely noticeable at the end of the polymer. The resulting interarrival distribution approximates single-exponential waiting times. Conversely, if the elongation is of the order of the initiation (no significant node occupation arises), and the *OFF* period is comparable to the elongation (node's occupation can be degraded efficiently), two time scales in the interarrival times due to the switch modulation are preserved despite elongation steps. Red solid curve in Fig. 2.16 represents the simulation with such parameters and no accountability for collisions. It overlaps with results obtained for the the simple burst model alone (empty circles, therein).

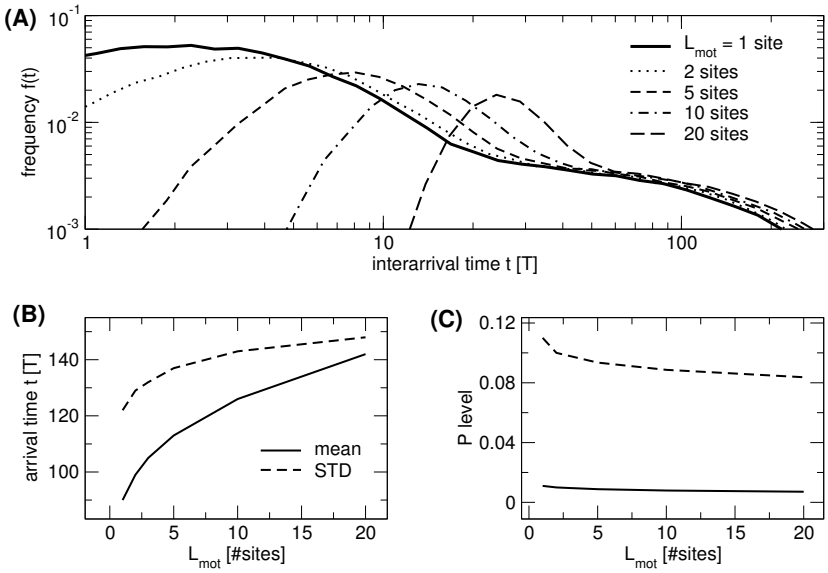


Figure 2.18: The effect of the size L_{mot} of the motor protein on the interarrival time frequency, average and standard deviation in the interarrival time and in the product level. Results of Gillespie simulations for polymer chain of length 100 sites ($k_{sw}^{+} = k_{sw}^{-} = 1\,[1/T]$, $k_{ini} = 100k_{sw}^{-}$, $k_{el} = k_{ini}/100$, $k_{p}^{+} = k_{p}^{-} = k_{el}/100$).

2.5.7.2 The effect of the size of the motor protein on $f_X(t)$

So far we considered the progression of motor proteins which occupy only one site at a time. A more realistic case involves motors of larger size which can block few sites. The effect of sizes $L_{mot} = 1 \ldots 20$ on the interarrival times is shown in Fig. 2.18A. The arrivals cannot occur more frequently than the time required for a single motor to travel its own length. Therefore, as the number of sites occupied by a motor increases, fast frequencies vanish from the interarrival time distribution. The duration of intervals within a burst approaches the duration of the interruptions between bursts. The time scale separation weakens compared to the case where $L_{mot} = 1$. The effect is visible in Fig. 2.18B. Standard deviation in the waiting times approaches mean for larger L_{mot}, which indicates an approach to a pure Poisson with a single-exponential waiting times distribution. As a result of a smaller flux at the end of the chain, caused by less frequent arrivals, the mean and noise in the product level decrease (Fig. 2.18C).

Chapter 3

Two-component signaling networks respond swiftly and robust at low molecule numbers

Contents

3.1 Abstract 54
3.2 Introduction 54
3.3 Results 55
3.3.1 Approximating the search of a single sensor by a first-order process 55
3.3.2 The promoter search can be approximated by a first-order process too 57
3.3.3 Clustering of sensors affects the search time for sensors 58
3.3.4 Two-component signaling-induced gene activation proves swift, robust and efficient 59
3.3.5 Contribution of diffusion and molecule copy number noise to noise in response time 60
3.3.6 Demand for fast and robust signaling can constrain operon organization 62
3.4 Discussion 65
3.5 Materials and Methods 67
3.5.1 Evaluation for 3D: searching the inner sphere 68
3.5.2 Evaluation for 3D: searching the target on the membrane 69
3.5.3 Round-trip: convolution 71
3.5.4 Moments of the first order statistics of a convoluted *pdf* 72
3.5.5 Bias of the optimal search time 72
3.5.6 Time to (de-)activate half of response regulators 73
3.5.7 Diffusive flux at the promoter site 73
3.5.8 Relation between the number of regulators and the number of DNA-binding sites 74
3.5.9 Two sources of stochasticity 76
3.5.10 Processivity 77
3.5.11 Numerical simulations 78
3.5.12 Bioinformatic analysis 78

3.1 Abstract

Two-component signal transduction is the prevailing extracellular sensing mechanism of prokaryotes. Remarkably, two-component networks mediate versatile responses to environmental changes, despite their simple design. We address how two-component systems give rise to fast and robust responses at the level of a single cell. Using detailed random-walk simulations we study how signaling proteins find their molecular targets inside single cells. We derive intuitive analytical approximations for the mean and noise of the response time. We demonstrate that tens of signaling molecules per cell reduce the response time such that additional molecules hardly improve signaling speed. Our analysis indicates that optimal response times are realized when the number of response regulators exceeds the number of sensors. This suggests an advantage for operon designs where the regulator is transcribed earlier on the polycistronic transcript than the sensor. We test those observations using a large bioinformatic screen of prokaryotic genomes. Model calculations indicate that experimentally determined values for RNA polymerase and ribosome processivities are sufficient to bring about the optimal ratio of response regulator over sensor abundance. We discuss the drawbacks of the two-component design in cells larger than prokaryotes, which may explain their ubiquitous use in the microorganisms and much less so in eukaryotes. The theoretical approach we propose relies on the fact that diffusion of molecules towards small targets within a cell can be approximated by a first order reaction with an exponential waiting time. This demonstrates how similar spatially resolved problems can be addressed without performing costly simulations.

3.2 Introduction

Unicellulars respond by adaptive mechanisms to continuous random fluctuations in the environment. Environmental sensing relies often on two-component (2C) signaling networks [Hoch & Silhavy 1995]; *E.coli* has 21 of such systems [Keseler *et al.* 2009]. Many stresses have instantaneous negative influences on cellular physiology suggesting that the response time of homeostatic signaling is minimized during evolution. Then, swift mutants would be favored by natural selection as sluggish responses reduce physiological performance. Whereas slowly responding cells may degenerate into states they cannot readily escape from due to cascading effects. This causes a cell to shift a large part of its resources to a stress response, too many of its other vital functions may become affected. Heat shock proteins can for instance already comprise 25% of all proteins in *E.coli*. Reduction of response times can be achieved by constitutive response systems. Indeed, free radical removal systems are on stand-by [Imlay 2003] as are many others [Fischer & Sauer 2005]. Likewise, the ability to quickly turn off the stress response is often beneficial to cells [Shalem *et al.* 2008]. The osmotic stress response exemplifies this: a sudden removal of excess salt would cause adapted cells to explode. Thus, many potential hazardous consequences of stress are reduced if cells are able to rapidly regulate their compensatory functions.

Many constraints apply to homeostatic adaptive systems. Stress adaptation may suffer from molecular noise distorting the signaling mechanism. Isogenic populations of cells have been shown to display a large heterogeneity in protein and mRNA levels with coefficients of variation as high as 60% [Elowitz *et al.* 2002, Rosenfeld *et al.* 2005, Newman *et al.* 2006].

Part of it derives from external noise and fluctuating reaction rates (thermal noise) [van Kampen 1997].

Noise in sensing and information transfer is manifested in the response time and the magnitude of the molecular response. Response time dispersion is further magnified by the stochasticity inherent to the diffusive motion of signaling proteins. This randomness affects the target search times of signaling molecules. We will use such search times as proxies for the response times of signaling processes. There have been many reports describing the diffusive search of a macromolecule for its target [Wunderlich & Mirny 2008, Halford & Marko 2004, Berg & von Hippel 1985]. Many of those theoretical studies can presently be experimentally scrutinized with single-cell monitoring techniques [Elowitz *et al.* 2002, Elf *et al.* 2007, Cai *et al.* 2006, Ozbudak *et al.* 2002, Yu *et al.* 2006]. Therefore, not only fast signaling is an essential prerequisite for cellular function but also the precision and reliability of the signaling network.

The prokaryotic strategies for stress responses turn out not to be very diverse. Across the Bacteria domain, extracellular signal detection is performed by two-component (2C) signaling networks composed of receptors (RE) and response regulators (RR) [Ulrich *et al.* 2005]. Their design and function is remarkably simple [Hoch & Silhavy 1995]. Diffusing response regulators have to find two molecular targets by way of a diffusive search. They need to find membrane-embedded receptors to become phosphorylated and subsequently the regulatory sequences on the chromosome to regulate transcription (Fig. 3.1). The number of target promoters for a specific regulator may vary between a few and several hundreds. The ubiquitous occurrence of 2C signaling systems suggests that their design is very favorable for dealing with information transfer and that they can cope with numerous performance-reducing factors, such as noise [Batchelor & Goulian 2003, Shinar *et al.* 2007]. Here we focus on the design of 2C signaling systems and their stochastic response characteristics.

Throughout this work we will assume that the actual phosphorylation and dephosphorylation events occur much faster than diffusive search processes. We assumed diffusion-limited kinetics to be able to draw general conclusions as all two-component signaling systems are confronted with the same search problems. As a consequence, the response time we predict will always underestimate the signaling time, especially if any of the two component signaling reactions are reaction-limited (i.e. sensor autophosphorylation, sensor-regulation phosphotransfer, regulator dephosphorylation, and DNA-regulatory site binding). The (de-) phosphorylation times can be obtained from the mean first-passage time of the catalytic scheme of the corresponding reactions [Qian 2008]. These reaction times can simply be added to the diffusive search times [van Zon & ten Wolde 2005a]. Therefore, this simplification does not limit the applicability of our results.

3.3 Results

3.3.1 Approximating the search of a single sensor by a first-order process

A single response regulator starting randomly in the cytosol finds any point on the membrane in $0.013\,s$, which derives from $R_{cell}^2/(15D_{RR})$ (Eq. 3.8, Materials and Methods). For this estimation, we took realistic parameters for a prokaryote, e.g. *E. coli*, with a cell

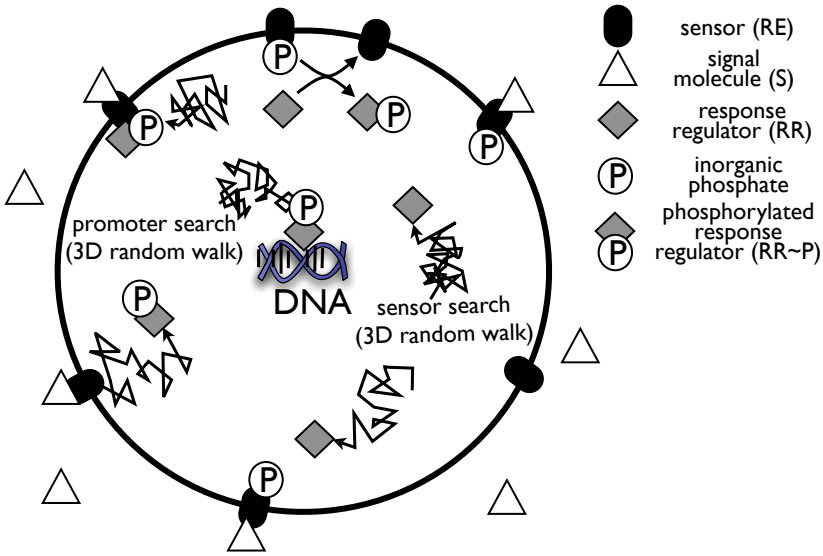


Figure 3.1: **Two-component signaling system.** Initially, response regulators (RRs) are non-phosphorylated and randomly distributed in the cytosol where they search for active sensors. The sensor or membrane receptor (RE) has a signal-recognition domain linked to an autokinase domain. The majority of such systems have membrane-embedded sensors with cytoplasmic DNA-binding response regulators (76%, Tab. S1). Upon signal (S) binding, the sensor autophosphorylates itself on a dedicated histidine residue at the expense of ATP hydrolysis. Subsequently, when a RR finds a RE, the RE-phosphoryl group can be transferred onto an aspartate group of the regulatory domain of the RR giving rise to a phosphorylated RR. Generally, unphoshorylated regulator domains inhibit the output domain of the response regulator. This inhibition is relieved upon aspartyl-phosphate formation in the RR to allow its output domain to carry out its function, which is typically DNA binding and transcription regulation [Stock *et al.* 2000].

radius $R_{cell} = 1\,\mu m$ and a diffusion coefficient $D_{RR} = 5\,\mu m^2/s$. This time only captures the time to reach the membrane. In reality, the membrane coverage for a single sensor type rarely exceeds 0.02 %; taking into account 50 sensors of $2.5\,nm$ radius each and a cell radius of $1\,\mu m$.

The mean diffusion time of a regulator to a single sensor in the membrane is given by the narrow escape limit [Schuss *et al.* 2007, Berg & Purcell 1977],

$$\tau_{RE} = \frac{V_{cell}}{4 D_{RR} R_r}. \tag{3.1}$$

(The reaction radius R_r is the sum of the radii of a RE and RR.) The reciprocal of τ is a rate constant. For *E.coli* parameters this search takes ≈ 42 seconds, which is about 3000 times longer than finding any point on the membrane! During that time the RR traverses the entire cell many times before it hits the sensor.

Due to the inherent stochasticity of diffusion the search duration follows a probability distribution rather than a single number. In Fig. 3.2A we compare the probability density

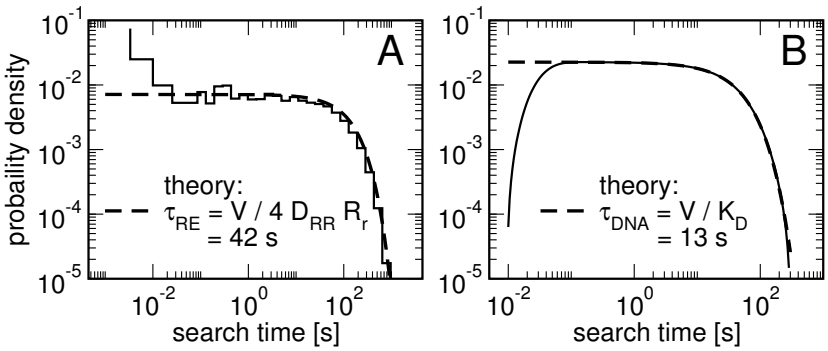


Figure 3.2: **Normalized log-log histograms of search times.** First-passage time probability density functions to reach the target by a single response regulator (RR). (A) RR diffuses from a random position in the cytosol until it hits a membrane-bound receptor. Reaction radius is twice the radius of a protein, $R_r = 5\,nm$. Dashed line: exponential approximation parameterized with the search time in a narrow escape limit (Eq. 3.1). Solid line: explicit 3D random walk simulation ($\Delta t = 10^{-8}\,s$, $D_{RR} = 5\,\mu m^2/s$, 1000 runs). This distribution has two regions. The left-most part scales as a power-law due to initial configurations for which the RR is localized near the target. The region on the right follows an exponential. (B) RR initiates at the membrane sensor and binds to a DNA site in the center. The site is modeled as a sphere; the reaction radius is $5\,nm$. Solid line, an exact analytical solution of Smoluchowski diffusion equation (Eq. 2.24 in SI); dashed curve, an exponential approximation parameterized by K_D (Eq. 3.2). There is a minimal time required to cover the distance from the membrane to DNA, hence the left region of the exact result where it starts at zero.

of simulated search times with the exponential approximation parameterized with the mean from Eq. 3.1. In this parameter regime – targets much smaller than the cell radius – the discrepancy occurs at very short times for regulators starting close to the sensor, and covers a negligible area of the probability density (less than 1%). Since binding to a small target is a rare event [Grigoriev *et al.* 2002], the search process is effectively memoryless: the search time is independent of the regulator's initial position. Hence the validity of the approximation of the waiting-time distribution by an exponential function (a Poisson process). As the dimensions of the cell no longer matter, stochastic simulations that only consider reactions suffice. We shall see further that this approximation holds also for membrane coverage larger than typically found in prokaryotes, i.e. larger than 0.05%.

3.3.2 The promoter search can be approximated by a first-order process too

We consider a single activated response regulator starting at the membrane-bound receptor searching for a single gene regulatory site. The search time distribution is peaked as the radius of the cell always needs to be traveled (Fig. 3.2B, solid line). As with the sensor search, the promoter search also operates in a regime where diffusive trajectories are independent of the initial position. Hence, the search time distribution can be approx-

imated by an exponential (dashed curve) parameterized by a mean search time (Eq. 2.25 in SI, [Redner 2001]),

$$\tau_{DNA} = \frac{V_{cell}}{4\pi D_{RR} R_r}\left(1 - \frac{R_r}{R_{cell}}\right) + \frac{R_r^2 - R_{cell}^2}{6D_{RR}}. \tag{3.2}$$

The Smoluchowski rate constant for a diffusion-limited association, $4\pi R_r D_{RR} \equiv K_D$ (units of volume per time) appears here. In the limit of a small target ($R_r \ll R_{cell}$) the first term, $V/K_D \equiv \tau_D$ dominates (Fig. S2). Taking parameters for *E.coli* we calculate ≈ 13 seconds for a single RR to find its DNA target.

In summary, both searches have a waiting-time probability density that can be satisfactorily approximated by an exponential. This is due to the fact that the size of the target sites is much smaller than the cell radius - compare the $2.5\,nm$ radius of the receptor or the promoter to the $1\,\mu m$ radius of the entire cell. This property allows for a substantial simplification of the theoretical analysis and simulation of diffusion in 2C signaling.

It takes the response regulator on average ≈ 42 seconds to reach the membrane-bound receptor from a random position in the cytosol and ≈ 13 seconds to find a DNA promoter site from the membrane (Fig. 3.2). The signaling time can be approximated as the sum of these two assuming the process is diffusion-limited. This response time of 55 seconds is comparable to the time required to transcribe an average mRNA and to translate it into a protein ($\approx 45\,s$ for an average 350-amino-acid-long protein transcribed at 50 nt/s and translated at 15 aa/s). As transcription and translation are continuously being optimized and involve many different genes, each with different selection pressures, it is likely that signaling speed cannot be enhanced by optimization of these two processes. Instead, natural selection will sift out those 2C signaling networks that are organized such that the search process is fast and reliable.

3.3.3 Clustering of sensors affects the search time for sensors

The response time can be reduced by increasing the number of sensors. The distribution of the sensors on the membrane affects the sensor search as well. Placing 100 sensors in a single cluster reduces the mean first-passage time by a factor of ≈ 5 relative to the search for a single sensor (Fig. 3.3A). An additional reduction is achieved by scattering those 100 sensors uniformly on the membrane (no clusters). Compared to a single sensor, a total 100-fold reduction follows. The search time is now about 30 times shorter than the time to reach the membrane, which was $0.013\,s$ (see Eq. 3.8 in Materials and Methods). This occurs already at 0.02% coverage of the membrane with receptors and this is still considering only a single RR that engages in the sensor search! In the regime where a few sensors ($N_{RE} \approx 100$ and $R_{RE} = 2.5\,nm$, $R_r = 5\,nm$) occupy the membrane, the search time scales as the inverse of the number of receptors (Fig. 3.3A and Eq. 2.31 in SI), which illustrates that the search process remains memoryless.

In Fig. 3.3B, we vary the number of response regulators and keep the number of sensors fixed at 100 per cell. Another ≈ 100-fold reduction in search time can be obtained when the number of receptors is increased to 100 per cell and all 100 sensors occupy a single cluster. If the cluster size decreases to 1 (i.e. cluster size 1 corresponds to single receptors), the search time reduces by another factor of 100 to yield the search time of $\approx 0.004\,s$. Thus, N_{RE} isolated receptors and N_{RR} response regulators give rise to a reduction of the

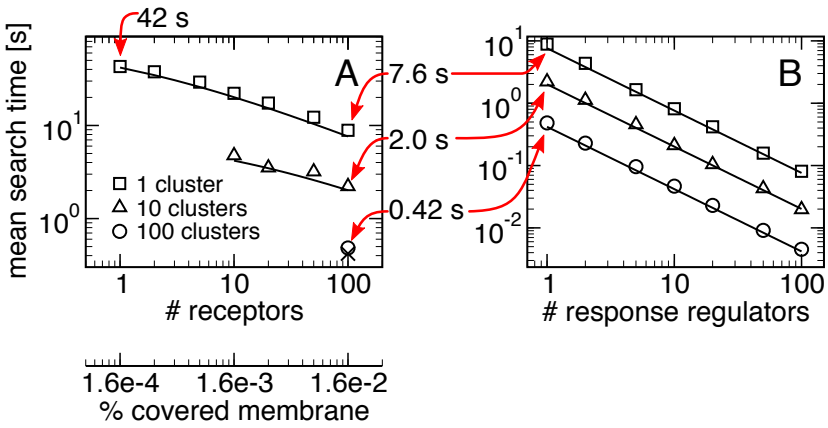


Figure 3.3: **Clustering and regulatoprs affect the mean search time.** The mean first-passage time for RRs to find the sensors for different clustering and total number of RR molecules. Symbols: numerical simulations, $D = 5\,\mu m^2/s$, at least 1000 runs. Lines: analytical approximations. (A) The search time for a single RR to sensor molecules (1 to 100) placed in clusters of different sizes. E.g. a triangle for 100 receptors in 10 clusters denotes a simulation where 10 uniformly distributed membrane clusters contain 10 receptors each. (B) The time for RRs (1 to 100) to find 100 sensors placed in uniformly distributed clusters of size 1 to 100.

response time of $N_{RE} \times N_{RR}$ relative to the time for a single receptor and regulator. This dependency holds true as long as the search trajectory is memoryless.

The observation that the search time is shorter for receptors that are scattered over the membrane agrees with earlier analyses of the influence of receptor partitioning on the membrane in the context of *external* signal reception [Berg & Purcell 1977, Shoup & Szabo 1982, Zwanzig 1990]. In case of bacterial chemotaxis, the membrane receptor molecules are found in large clusters. This may benefit output signal amplification through cooperative interactions between receptors. There are no indications that this holds for 2C systems in general. Many of them may have their sensors randomly distributed over the membrane, which may indicate optimization for a fast response rather than signal sensitization and amplification.

3.3.4 Two-component signaling-induced gene activation proves swift, robust and efficient

The overall mean search time equals the sum of the two mean search durations; the overall waiting-time distribution is the convolution of two exponentials describing the two sequential searches. The resulting distribution will necessarily become peaked (Fig. S5); regulators first have to find sensors before they can regulate genes and therefore instantaneous gene activation is ruled out. The noise in the total search time, its variance divided by the squared mean, $\langle \delta^2 \tau_s \rangle / \langle \tau_s \rangle^2$, is smaller than the noise in either of the two first-order search process (they each have a noise of 1 as they can be approximated by an exponential). The minimal noise in the search time equals 1/2 when both searches have

the same mean. Thus, the response in a two-step design is timed more precisely despite strong fluctuations due to diffusion-limited searches.

The response time distribution becomes even narrower when more response regulators are considered (Fig. S6). Since the overall search time distribution is not an exponential, the mean signaling time is no longer inversely proportional to the number of regulators. The analytical expression for the mean and the noise of the overall signaling times can still be obtained (Ch. 3 in SI) [Yuste *et al.* 2001]. To provide more intuitive insight, we shall discuss a simplified version of the expression of the mean response time, which retains the major qualities of the explicit solution (Eq. 3.7 in SI). We assume that the waiting-time distribution of total search with N_{RR} regulators and N_{RE} sensors, τ_s, is approximated by an exponential. Therefore,

$$\tau_s = \frac{1}{N_{RR}} \left(\frac{V_{cell}}{4 D_{RR} R_r} \cdot \frac{1}{N_{RE}} + \frac{V_{cell}}{4\pi D_{RR} R_r} \right). \tag{3.3}$$

A reduction in τ_s can either result from increasing the number of receptors, the number of response regulators or both.

A cell has limited resources and is packed with proteins (the average distance between proteins in the cytoplasm is the size of a protein) [Ellis 2001]. The membrane hosts numerous macromolecules as well and the space for a receptor species of a specific signaling pathway is limited. In addition, there exist constraints on energy expenditure for protein production [Dekel & Alon 2005]. The following questions then become relevant. How many signaling molecules are sufficient for swift transduction of the external signal to a gene regulatory sequence? Should an optimal cell invest in an equal number of sensors and response regulators?

Fig. 3.4 addresses these issues. It shows that the search time can take up to a minute if very few molecules are involved. A performance below 2 seconds is achieved already with ≈ 40 molecules (Fig. 3.4A and C). More molecules will lead to a faster target search, but the energetic burden of producing additional proteins (linear with protein number) will start to exceed the benefit of search time reduction (scales with the inverse of the total number of molecules). We shall come back to this point later when we discuss signaling designs in eukaryotes. As some 2C networks have cytosolic sensors ($\approx 11\%$, Tab. S1) we also consider such a signaling design. As expected, it takes less time for the first phosphorylated RR to reach the promoter site as compared to the 2C system with membrane-binding REs (Fig. 3.4B).

Intuitively one might expect that the shortest search times are achieved when the number of diffusing and membrane-embedded molecules (i.e. RRs and REs) is approximately equal. However, Fig. 3.4A reveals a clear bias; both for membrane-embedded and cytosolic sensors. The shortest response time is achieved with $\approx 10\%$ more diffusive regulators than sensors. This bias remains fixed even if the total number of molecules changes (Fig. 3.4D). Can cells achieve this ratio, and can it be controlled given the inevitable noise in gene expression? This we will address below.

3.3.5 Contribution of diffusion and molecule copy number noise to noise in response time

The biased ratio of regulators over receptors to achieve the shortest search time will inevitably suffer from fluctuations. They may persist longer than the time scale of envi-

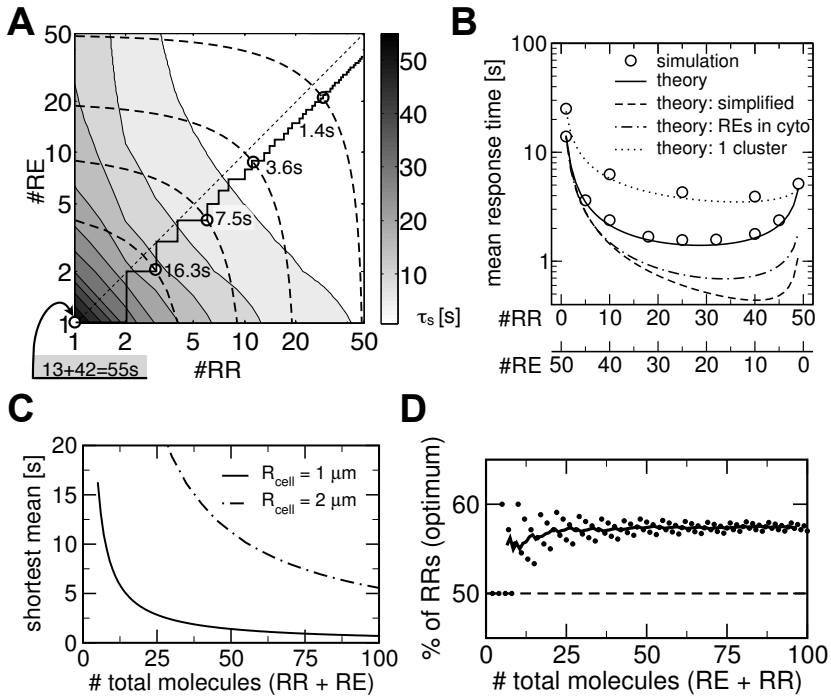


Figure 3.4: **The mean search time as a function of the number of receptors and response regulators.** Computed with the explicit analytical equation 3.7 SI. (A) Phase plot. Dashed lines: isoclines for the total number of $RR+RE = 5$, 10, 20, 50. The minimal search times follow the staircase-shaped curve. It lies below the diagonal, which reflects a bias favoring RRs. (B) The mean search time for 50 signaling molecules in total evaluated for receptors scattered randomly on the membrane (solid line), grouped in one membrane cluster (dotted) or placed randomly in the cytosol (dash-dotted). Numerical simulations (circles) obtained from 1000 runs with $D_{RR} = 5\,\mu m^2/s$, $R_{cell} = 1\,\mu m$, $R_r = 5\,nm$. Fig. S7 illustrates the effect of additional reaction times on the mean search time. (C) The shortest search time as a function of the total number of 2C proteins for two cell sizes. (D) Bias in the ratio of RRs and REs as a function of the total number 2C proteins. For small numbers, the percentage fluctuates strongly due to discreteness of molecular count. The running average is plotted to indicate the trend.

ronmental changes if the proteins are stable [Sigal *et al.* 2006]. Hence, signaling speed may be significantly affected by fluctuations. Note, that the flatness of the curves in Fig. 3.4B and C already hints at robustness of the mean search time to changes in component levels once the total number of molecules exceeds ≈ 25.

On top of diffusion noise copy number fluctuations contribute a second (independent) determinant to the variance in overall signaling response time, $\sigma_s^2 = \sigma_D^2 + \sigma_N^2$. The first term denotes fluctuations in search time resulting from diffusion for a fixed number of signaling components (Fig. S11 and Eq. 3.7 in SI). The second term reflects the fluctuations in the mean search time due to noise in the molecular copy number. This can be

expressed in terms of the noise in the levels of sensors and response regulators as,

$$\begin{aligned}\sigma_N^2 &= \left(\partial\tau_s/\partial N_{RR}\right)^2 \sigma_{RR}^2 + \left(\partial\tau_s/\partial N_{RE}\right)^2 \sigma_{RE}^2 \\ &+ 2\cdot\partial\tau_s/\partial N_{RR}\cdot\partial\tau_s/\partial N_{RE}\cdot COV_{RR,RE}. \end{aligned} \tag{3.4}$$

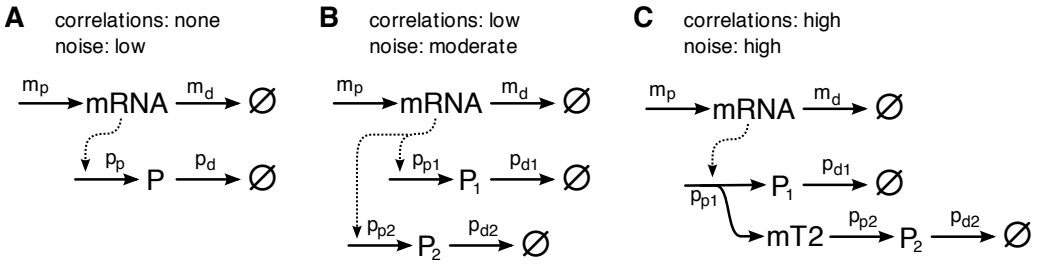
The terms σ_{RR}^2 and σ_{RE}^2 represent variance in the level of response regulators and sensors, respectively; both stem from noise in gene expression. The covariance term, $COV_{RR,RE}$, quantifies the correlation between fluctuations of REs and RRs. Copy number fluctuations result from noisy transcription, translation and degradation of mRNA and protein. Correlations may arise when the genes for RR and RE are regulated by the same operon or else through some global regulatory mechanism. For a given gene expression scheme, all three terms can be obtained by linear noise approximation (LNA) [Elf & Ehrenberg 2003, Paulsson 2004]. In the following section, we shall investigate Eq. 3.4. Our aim will be to find constraints on protein levels, the sensor/regulator ratio and their fluctuations depending on the operon organization of cognate 2C signaling components.

3.3.6 Demand for fast and robust signaling can constrain operon organization

The fluctuations in the copy numbers of sensors and response regulators are partially determined by their gene organization on the genome. Both genes may lie at distant stretches of DNA and regulated independently from separate promoters. Then fluctuations in the level of response regulators (σ_{RR}) do not have a direct effect on the level of sensors and the covariance $COV_{RR,RE}$ from Eq. 3.4 equals zero. Coordinated co-regulation occurs when the genes are on one operon. In addition, correlations depend on whether sensor and regulator genes share a single ribosome binding site (RBS) or not and, in case of separate RBS's, their relative affinity for the translation machinery. Uncorrelated protein fluctuations resulting from disruptions of such operon ordering have been shown to be detrimental to the sensitivity of the chemotaxis network [Løvdok *et al.* 2007] and sporulation efficiency in *B. subtilis* [Iber 2006].

Analysis of 934 microbial genomes, using the MiST database [Ulrich & Zhulin 2007], revealed that out of all receptors containing a trans-membrane domain, 66% of them is chromosomally adjacent to DNA-binding regulators (Tab. S1). One-fourth of such pairs is transcribed onto a single dicistronic mRNA transcript, which may include one or two ribosome binding sites (RBSs), the latter being almost ten times more frequent (Tab. S4). Single or two-RBS designs differ in the ability to constrain the ratio RR/RE. In the case of two RBSs, the level of both proteins may vary more strongly depending on the difference between the strength of the RBSs (assuming similar protein lifetimes). The ratio RR/RE is to a large extent fixed if ribosomes bind only to one RBS in front of the first cistron.

We consider three different operon designs of 2C networks (also analyzed by Swain [Swain 2004] for another purpose). They differ in copy number fluctuations. 2C proteins regulated by independent promoters (Fig. 3.5A) do not exhibit correlated fluctuations other than those induced by extrinsic noise, hence the covariance in Eq. 3.4 equals zero. The other two configurations, a dicistronic mRNA with 1 or 2 RBSs, result in an increased σ_N^2 due to correlated fluctuations (Fig. 3.5B and C). The highest values occur for a dicistronic mRNA with a single RBS.



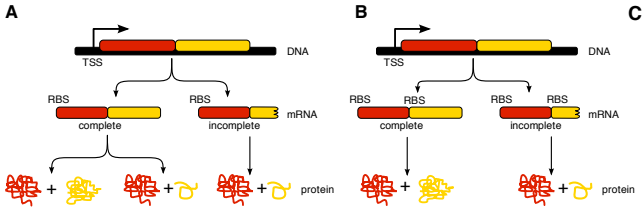
	A: RR & RE separate	**B:** isolated, 2 RBS	**C:** isolated, 1 RBS
occurrence in genome	34%[(a)]	53%	6%
number of P_1 / cell	23	23	20
number of P_2 / cell	17	17	20
response time, τ_s [s]	1.75	1.75	1.78
$\eta_D{}^2 = (\sigma_D / \tau_s)^2$	0.35	0.35	0.37
$\eta_N{}^2 = (\sigma_N / \tau_s)^2$	0.009	0.010	0.015
correlation (P_1, P_2)	0	0.24	0.60

Figure 3.5: Models of gene expression. (A) expression of both genes is controlled by different promoters, (B) both genes are transcribed onto a single dicistronic mRNA with one ribosome binding site (RBS) for each cistron, (C) a dicistronic mRNA has only one RBS. The table includes sample numerical results for all three gene expression schemes. A non-zero correlation, $COV_{P_1,P_2}/(\sigma_{P1} \cdot \sigma_{P2})$, between P_1 and P_2 arise when both proteins are produced from a single dicistronic mRNA. Parameters used: $m_p = m_d = 1\,min^{-1}$, $p_d = p_{d1} = p_{d2} = 1/60\,min^{-1}$; (A and B) $p_{p1} = 23/60\,min^{-1}$, $p_{p2} = 17/60\,min^{-1}$; (C) $p_{p1} = p_{p2} = 20/60\,min^{-1}$.

How much does the search time suffer from these noise strengths? In the regime where the search time achieves a 1-2 second performance, noise in copy numbers accounts for $\approx 10\%$ of the total variability through σ_N^2/τ_s^2; the majority still derives from diffusion (Table in Fig. 3.5). Thus, even for the strongest correlations in protein number noise, the reliability of the search process is compromised only slightly. Remarkably, as correlations constrain fluctuations of 2C proteins to the diagonal region of the phase plot, Fig. 3.4, they are beneficial for maintaining robust short search times.

How can genes transcribed onto a polycistronic mRNA establish a bias in the protein numbers they code for? This can occur when either transcription or translation terminates prematurely giving rise to either an unfinished mRNA transcript or protein, respectively. Depending on the probability for a polymerase or a ribosome to dissociate before reaching the end of the gene (this is termed processivity), genes placed close to the transcription start site or RBS tend to accumulate more product (known as polarity) [Lengeler *et al.* 1999, Schäferjohann *et al.* 1996]. If the mRNA transcript hosts two RBSs, processivity affects the protein ratio only at the level of transcription. A single-RBS ar-

rangement suffers from processivity also during translation; premature termination may occur not only during mRNA synthesis, but during protein production as well (Fig. 3.6A and B).



C

fraction in genome	1 RBS at gene 1		2 RBSs: at gene 1 & 2	
	9%		80%	
gene order	RR → RE	RE → RR	RR → RE	RE → RR
fraction in genome	84%	16%	59%	41%
length of gene 1 [nt]	711 ± 117	1674 ± 561	723 ± 141	1653 ± 513
length of gene 2 [nt]	1410 ± 219	792 ± 285	1401 ± 240	849 ± 318
$mRNA_1$ / $mRNA_2$	2.02 ± 0.22	1.49 ± 0.21	2.02 ± 0.24	1.53 ± 0.24
P_1 / P_2	4.08	2.22	2.02	1.53
P_{RR} / P_{RE}	2.63 (72% P_{RR})		1.26 (56% P_{RR})	
overall P_{RR} / P_{RE}	1.33 (57% P_{RR})			

Figure 3.6: **Incomplete transcription affects mRNA and protein level.** mRNA with a single (A) and two RBSs (B). (C) Bias in steady-state mRNA and protein ratios for response regulators and receptors in 2C networks due to incomplete transcription and translation. Lengths of adjacent response regulators and receptors with trans-membrane domains are obtained from the MiST database [Ulrich & Zhulin 2007] (Table S4). RBS were predicted using GeneMarkS [Besemer *et al.* 2001] and RBS finder [Suzek *et al.* 2001]. $mRNA_1/mRNA_2 = (1-p)^{N_2}$ is the ratio of the number of finished mRNA transcripts of the first and the second gene; it depends only on N_2, length of the second gene, and on p, processivity of RNA polymerase - probability of falling off per nucleotide (Section 3.5.10 in Materials and Methods). Similarly, P_1/P_2 is the ratio of the first over the second protein. We assume p independent of the sequence and equal to 0.0005 [Jülicher & Bruinsma 1998] for transcription and translation. Following the ± sign is the standard deviation.

Our theoretical results suggest a bias in the number of molecules in favor of response regulators in order to attain the shortest search time. This bias can result from unproductive transcription or translation events (RNA polymerase or ribosome falling off the template). This mechanism favors RRs, as predicted, if they are first on the polycistronic transcript. Bioinformatic analysis of 8488 pairs of adjacent genes of 2C proteins mildly supports this claim. In 60% of the cases the response regulators with DNA-binding domains are located before trans-membrane receptors on the transcript (Tab. S2). Using gene lengths for both RR-RE orderings in 1- and 2-RBS configurations, and taking into account their occurrence in genome, we computed mRNA and protein ratios resulting only from incomplete transcription and translation (Fig. 3.6C). The overall RR:RE ratio equals 1.33 (57% of RRs) and is consistent with our prediction of the optimal bias of signaling proteins yielding the shortest search time (Fig. 3.4). It would be erroneous, however, to attribute this bias to the requirement of short search times only as there are certainly more factors at play. Compared to response regulators, the longevity of receptors is likely higher because it is more difficult for proteases to access proteins embedded in the membrane. On the other hand, a reverse gene ordering favoring receptors would enhance signal transduction when not all receptors are activated instantly upon signal appearance, which requires more REs. Additionally, placing RR genes downstream of those coding for receptors offsets a net 5'-3' directionality of mRNA degradation [Selinger *et al.* 2003] which arguably results in a longer time window for translation of the second cistron. Interplay of these effects is likely to be the reason behind the bias found in the bioinformatics data.

3.4 Discussion

Many microorganisms experience their environment through 2C signaling [Stock *et al.* 2000]. These systems act as sensors and induce measures to cope with environmental changes and stresses. *E.coli* houses at least twenty of such systems [Keseler *et al.* 2009, Ulrich & Zhulin 2007]. We determined the response time of 2C signaling by considering the time for the first molecule to reach the DNA regulatory site.

A more detailed description, however, needs to account for the lifetime of the RR-DNA complex compared to the transcription initiation rate. As a result, gene activation (or inhibition) may require multiple attempts of RR binding. For ≈ 40 molecules, the time to phosphorylate half of the inactive RR pool by membrane receptors is below 2 seconds and only weakly affected by fluctuations in the number of components (Fig. S8; the same holds for the dephosphorylation half-time – the time to switch the signaling off). The equilibrium concentration of phosphorylated RRs ensues almost instantaneously, for the time to traverse the cell of the *E.coli* size is as short as $0.2s$. The lifetime of their phosphorylated form may range from minutes to hours [Stock *et al.* 2000]. During that time, many association events will take place before the change in transcription follows. The probability of transcription initiation, $P_{ini} = k_{ini}/(k_{diss} + k_{ini})$, depends on the initiation rate, k_{ini}, and the dissociation rate of the RR, k_{diss}. The average number of RR associations before transcription starts is $1/P_{ini}$. This number corresponds roughly to the k^{th} binding of the response regulator. Given the high probability of the RR in the vicinity of the target site to rebind before the second RR arrives, transcription initiation may still be induced by the first regulator that reached the promoter site.

The description of the promoter search omits any contribution of one-dimensional sliding of response regulators along the DNA. Sliding has been addressed by a number of theoretical studies [Wunderlich & Mirny 2008, Halford & Marko 2004, Berg *et al.* 1981] and recently also experimentally [Elf *et al.* 2007]. Its occurrence does not affect the memoryless character of the signaling process, but it does influence the mean. According to Eq. 3.2, a single *lac* repressor dimer ($D = 3\,\mu m^2 s^{-1}$, $R_{rep} = 2.5\,nm$) reaches the promoter ($R_{DNA} = 2.5\,nm$) in ≈ 20 seconds ($R_{E.coli} = 1\,\mu m$) by means of free 3D diffusion only. The mean first-passage time reported by Elf and colleagues [Elf *et al.* 2007] ranges from 65 to 360 seconds. They measure an effective diffusion coefficient (including the effect of non-specific binding to DNA and subsequent 1D sliding along DNA segments) of $0.4\,\mu m^2 s^{-1}$. Substituting this to Eq. 3.2, yields the search time of 165 seconds, which is consistent with experimental findings. This time corresponds to the time for a single molecule to find the site. For N diffusing molecules this time is divided by N. Although excluded in our exposition, the model can be straightforwardly extended to incorporate sliding by adjusting the average.

We predict that tens of 2C signaling molecules are sufficient to give rise to a fast and robust response; adding more proteins does not reduce the time significantly and the influence of fluctuations in those molecule numbers is only minor (Fig. 3.4C). Only for a few systems are the concentrations of 2C signal transduction proteins known. Sensors typically fall in the nanomolar regime, which means for *E. coli* tens of molecules per cell ($1\,nM \approx 1$ molecule cell^{-1})[1] [Weiss *et al.* 1992, Cai & Inouye 2002, Shen *et al.* 2008].

[1] Personal communication with Prof. Dr. K Hellingwerf: UhpB ($25\,nM$; RE), UhpA ($2500\,nM$; RR), ArcA $25\,\mu M$.

Our analysis considered a single target gene and the complete absence of low affinity binding on the DNA that compete for binding response regulator. Taking this into account, we would predict a higher response regulator than sensor concentration in the presence of low-affinity sites and multiple gene targets. In fact, this appears to be the case.[2] When we assume that the residence time of response regulators on competing sites is long compared to the duration of target finding, we predict a linear dependence between the number of extra sites and response regulators required to find the target gene within a given time (Fig. S10).

The analysis of models of gene expression indicates that diffusion accounts for $\approx 85\%$ of the noise in search time (Fig. 3.5). Copy number noise is responsible for the remainder. However, transcription and translation dynamics that we have assumed does not display bursts. Burst-like transcription rates have been measured for a large number of genes [Golding *et al.* 2005, Bar-Even *et al.* 2006, Cai *et al.* 2006]. In prokaryotes, such bursts have been most convincingly shown for repressed genes. Bursty expression of two-component proteins might severely impair signaling performance especially if both components are regulated on separate operons and fluctuations due to bursts extend over cellular generation time [Rosenfeld *et al.* 2005, Sigal *et al.* 2006] In the extreme case a single cell might receive either a receptor or a regulator during its lifetime, thus inhibiting the signaling function entirely. One solution to avoid such uncorrelated fluctuations is to express the receptor and the regulator from a single operon. In 66% of cases transmembrane receptors are neighboring DNA-binding regulators (Tab. S1). This might indicate evolutionary pressure to minimize detrimental copy number noise in two-component transduction systems by grouping functionally related proteins.

We discussed the effect of diffusion on signaling for a cell of the size of a typical prokaryote ($R_{E.coli} \approx 1\mu m$). A few dozens of molecules suffice to sense the extracellular stimulus, diffuse, and affect gene regulation in little more than 1 second. How many molecules would be required to retain the same efficiency using a 2C design for a larger volume, say a typical eukaryotic cell? Signaling networks that involve SMADs [Clarke & Liu 2008], STATs [Swameye *et al.* 2003] and nuclear receptors [Carlberg & Dunlop 2006] have a network design similar to 2C networks; they are much more linear and involve much less components than for instance MAPK signaling. Our theory predicts that in order to attain search times $< 2\,s$ in a cell of radius $10\mu m$ at least 10'000 RR and REs are necessary (Fig. 3.7). The abundance of receptors in some eukaryotic signaling pathways is of that order [Clarke & Liu 2008, Kholodenko *et al.* 1999]. More importantly, the vast majority of eukaryotic signaling pathways utilize more elaborate schemes than a simplistic two-step architecture. Eukaryotic signal transduction is usually arranged in networks allowing for more regulatory checkpoints and a possibility of crosstalk between other signaling pathways.

Obtaining molecular copy number distributions across a spatial domain usually involves intensive computations. Such simulations are only necessary when the waiting-time distributions for the search processes involved are non-exponential. In the regime of small targets waiting-time distributions become exponential. Such memoryless processes have a number of interesting properties. Their duration does not depend on the cell's shape and internal organization. This decoupling makes search times less dependent on physiological

[2] Personal communication with Prof. Dr. K Hellingwerf: RR concentration in the presence of low-affinity binding sites, $1 - 10\,\mu M$.

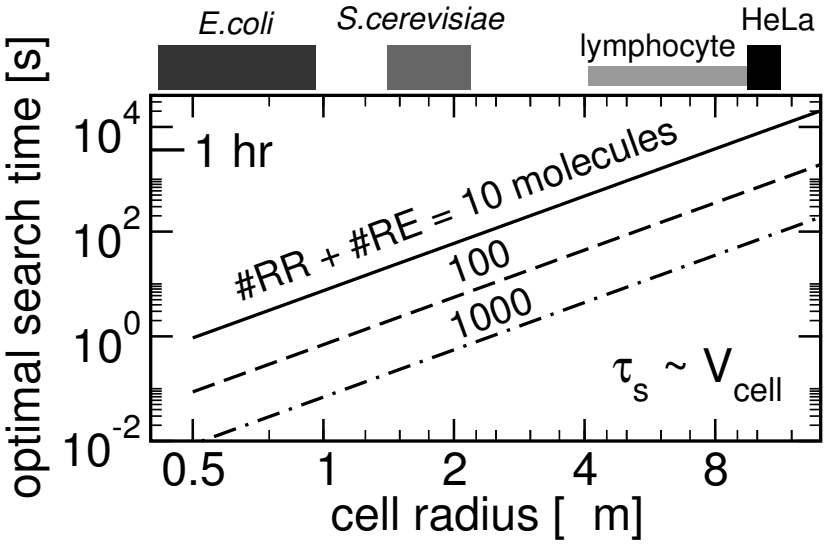


Figure 3.7: The optimal search time as a function of the cell radius. The number of 2C proteins required to maintain short signaling times increases for larger cells. The time scales as r^3 with the radius, hence is linear with the volume.

conditions and allows for more evolvable signaling systems. Moreover, the system can be considered as effectively ideally stirred and the stochasticity can be described by analytical theories. This is, however, rarely done in current modeling efforts in computational systems biology.

Acknowledgments

Authors thanks Hans Westerhoff, Rob van Spanning, Klaas Hellingwerf and Joke Blom for insightful discussions. FJB acknowledges the Netherlands Institute for Systems Biology and the Netherlands Organization for Scientific Research (NWO) for funding. MD and JVR thank the Netherlands Organization for Scientific Research (NWO), project NWO-CLS-635.100.007, for funding.

3.5 Materials and Methods

This section includes an extensive discussion of analytical tools and computational methods. Analysis of three models of gene expression (Fig. 3.5) can be found in the Mathematica file, under the address `http://projects.cwi.nl/sic/2csignet/`. Therein, the reader can find steady-state solutions, computation of variance and covariance from linear noise approximation, and the effect of diffusion and copy number noise on the total search time.

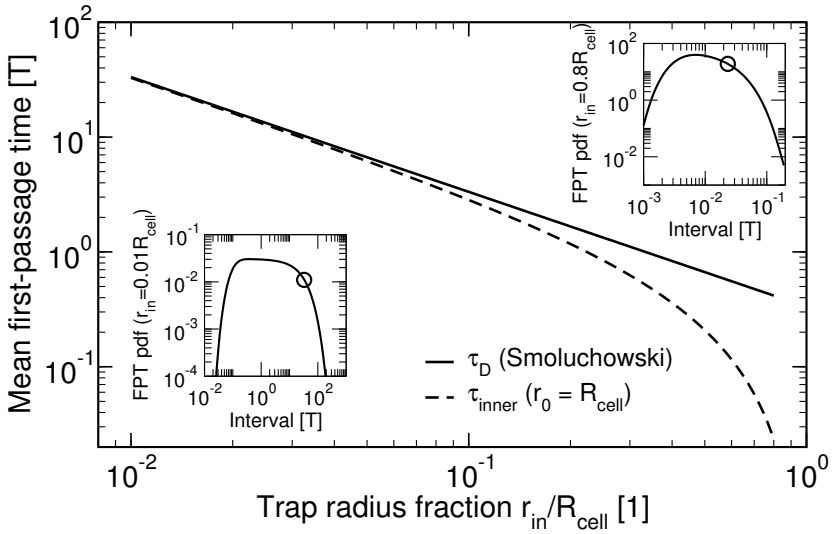


Figure 3.8: Two estimates of the MFPT to the absorbing inner sphere as a function of its radius r_{in}: inverse of the Smoluchowski diffusion-limited rate constant $\tau_D = V_{cell}/4\pi D r_{in}$ (solid line), exact solution for a molecule starting at the outer boundary, τ_{inner}, Eq. 3.6 (dashed). A single molecule diffuses in a sphere of radius R (reflective boundary) with diffusion constant $D = 1\,L^2/T$. Insets: FPT *pdf* for two extreme trap sizes, $r_{in} = 0.01$ and $0.8\,R_{cell}$. Curves were obtained by numerical inversion of the Laplace transform of the analytical flux at the absorbing target.

3.5.1 Evaluation for 3D: searching the inner sphere

The same procedure applied for $d = 3$ yields the FPT *pdf* to the inner sphere for a single molecule starting at the outer membrane:

$$\tilde{f}(r_0 = R_{cell}, s) = \frac{r_{in}\sqrt{\frac{s}{D}}}{R_{cell}\sqrt{\frac{s}{D}}\cosh\left[(r_{in} - R_{cell})\sqrt{\frac{s}{D}}\right] + \sinh\left[(r_{in} - R_{cell})\sqrt{\frac{s}{D}}\right]}, \quad (3.5)$$

where r_{in} is the radius of the inner sphere.[3]

The MFPT obtained by procedure given in Eq. A.6 reads:

$$\tau_{inner} = \frac{\frac{4}{3}\pi R_{cell}^3}{4\pi D r_{in}} + \frac{r_{in}^2 - 3R_{cell}^2}{6D} = \underbrace{\frac{V_{cell}}{K_D}}_{\tau_D} + \underbrace{\frac{r_{in}^2 - 3R_{cell}^2}{6D}}_{\tau_{corr}}. \quad (3.6)$$

The first term, K_D, is a Smoluchowski diffusion-limited reaction rate (V_{cell} is volume of the system). In case of targets much smaller than the radius of the cell, this term dominates the mean (Fig. 3.8). It is the regime where trajectories become so long that they become insensitive to the initial conditions. The FPT *pdf* for a molecule searching a small

[3]In this derivation we implicitly assume that the diffusing molecule is a point object. The reaction at the absorbing inner sphere takes place if the distance between the molecule and the target is smaller than or equal r_{in}. This distance can be perceived as a reaction radius. In that case it is a sum of the molecule's and the target's radii.

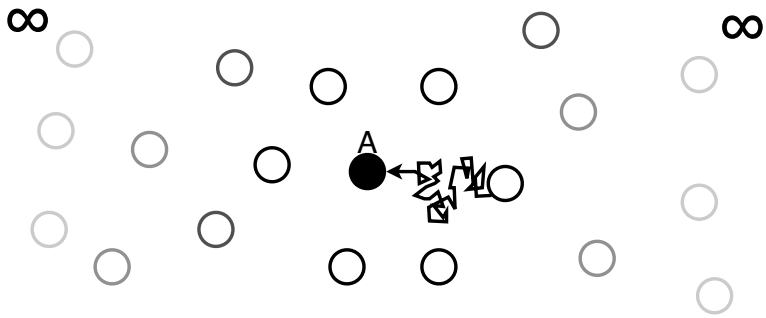


Figure 3.9: Illustration of the single-species annihilation, $A + B \to 0$, as a first-passage problem. Molecule A, solid black, is fixed as a "trap". A fixed concentration of molecules B, empty circles, extend to infinity. They diffuse and annihilate upon collision. Reaction rate k is then the flux to the target, i.e. the frequency of arrivals of molecules at the absorbing molecule. In 3D this rate equals $K_D = 4\pi a D$, where D is diffusion constant of molecules B, a is the reaction radius, typically the sum of radii A and B.

target is approximately exponential. The memoryless property of diffusive trajectories is also present in the reasoning behind the rate constant K_D. It is the steady-state flux of molecules to the spherical target. The target is immersed in an infinite sea of diffusing molecules (Fig. 3.9). Hence, intervals between binding events are independent of initial spatial configuration of molecules.

Note that it is incorrect to estimate the time needed to reach the target for a molecule starting at the outer boundary similar to Eq. B.15, i.e. $\langle t \rangle = R_{cell}^2/2d\,D$. Such procedure gives an average distance covered by a diffusing molecule, which in fact models a situation where we look for the time to cross a (hyper-)sphere at distance R_{cell}.

3.5.2 Evaluation for 3D: searching the target on the membrane

Diffusive search times can be estimated using first passage time theory [Redner 2001, Weiss 1967]. For instance, a macromolecule with a diffusion coefficient D (in $\mu m^2/s$) that starts in the middle of a cell (with radius $R\,\mu m$) will reach any point on the membrane in $R^2/(6D)$ seconds on average. The corresponding distribution of search times (a normalized histogram, or a probability density function) is peaked and asymmetrical. The noise, η_τ^2, (variance divided by the mean first passage time squared) of such a distribution equals $2/5$. Such a process is therefore less noisy than a pure Poisson process ($\eta_\tau^2 = 1$). This time reduces to $R^2/(15D)$ when they start randomly in the cytosol.

3.5.2.1 Fully absorbing sphere

In this section we will evaluate first-passage properties for a domain where a particle diffuses inside of a sphere. In order to calculate the first-passage time to any point on that sphere, we shall solve a spherically symmetric Eq. B.14 with absorbing boundary at $R_+ = R$, and no inner sphere, i.e. $R_- \to 0$. The MFPT to the outer boundary for a

particle initiating at any radial distance r_0 equals:

$$\langle t \rangle_{r_0} = \frac{R^2 - r_0^2}{6D}. \tag{3.7}$$

If the particle starts from a random r_0, then the MFPT decreases:

$$\langle t \rangle_{rnd} = \frac{3}{R^3} \int_0^R \langle t \rangle_{r_0} r_0^2 \, dr_0 = \frac{R^2}{15D} = \frac{2}{5} \langle t \rangle_{r_0=0}. \tag{3.8}$$

3.5.2.2 Partially absorbing sphere

If one is interested in the MFPT to a part of the sphere (reflecting boundary except for an absorbing part of interest), radial symmetry is lost and problem becomes much more mathematically involved. Singer and colleagues [Singer *et al.* 2006b, Singer *et al.* 2006a] obtained a number of results for mean exit times from the center to a small opening on a disk and a sphere :

$$2D: \qquad \langle t \rangle_{r_0=0} = \frac{R^2}{D} \left(\log \frac{1}{\varepsilon} + \log 2 + \frac{1}{4} + \mathcal{O}(\varepsilon) \right), \tag{3.9}$$

$$3D: \qquad \langle t \rangle_{r_0=0} = \frac{\frac{4}{3}\pi R^3}{4 r_{abs} D} \left(1 + \frac{r_{abs}}{R} \log \frac{R}{r_{abs}} + \mathcal{O}\left(\frac{a}{R} \right) \right), \tag{3.10}$$

where ε is a fraction of disk's perimeter occupied by the absorbing part, r_{abs} is the radius of a circular absorbing patch on a sphere. These results hold if the absorbing opening is small compared to total perimeter or surface.

3.5.2.3 Many absorbing patches

In the main text we argue that the MFPT to many receptors (absorbing patches) decreases if receptors are scattered on the sphere. The MFPT to any of N_{RE} receptors scales with the inverse of the number of receptors if (i) receptors are small compared to the total surface, i.e. a diffusive trajectory to a single target is memoryless, (ii) distance between them is large compared to their size, i.e. diffusive flux around a single target is not affected by other receptors.

In order to better illustrate this effect we resort to a solution of the *adjoint equation* for the mean first-passage time t [Redner 2001]:

$$D \nabla^2 t(\vec{r}) = -1 \tag{3.11}$$

subject to the absorbing boundary condition $t(\vec{r}) = 0$, for $\vec{r} \in \Omega$. Solution of this Poisson equation at point $\vec{r}$ is the mean first-passage time from that point to any of the absorbing boundary points. A numerical integration for different configurations of absorbing areas is shown in Fig. 3.10.

We assume that a cluster of receptors is simply a larger receptor of surface equal to the sum of single receptors' area, i.e. $R_{CL} = R_{RE}\sqrt{N_{RE/CL}}$, where R_{CL} and R_{RE} are radii of a cluster and a receptor, and $N_{RE/CL}$ is the number of receptors per cluster. A full approximate expression for the mean first-passage time of a *single* response regulator of radius R_{RR} to N_{CL} clusters containing $N_{RE/CL}$ receptors each, reads:

$$\tau_{RE} = \frac{V_{cell}}{4D \left(R_{RE}\sqrt{N_{RE/CL}} + R_{RR} \right) \cdot N_{CL}} \tag{3.12}$$

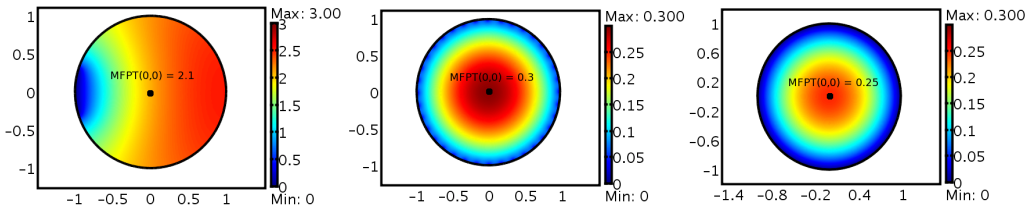


Figure 3.10: The mean first-passage time to a disk with absorbing outer boundary. Plots obtained by numerical integration of adjoint equation for the mean first-passage time, Eq. 3.11. Color coding corresponds to the MFPT from a point on a disk to the absorbing area on a perimeter; blue - short, red - long MFPT. The MFPT for a molecule initiating in the center is marked with a dot. (Left) A continuous absorbing region takes 10% of the perimeter. (Middle) The same absorbing fraction of 10% is scattered uniformly into 36 parts. (Right) The absorbing boundary takes the whole perimeter. The MFPT from the center is only slightly lower compared to the case in the middle panel.

The scaling with the number of clusters, $1/N_{CL}$, reflects the exponential approximation which is only valid for a small cluster size. Once a waiting time distribution for binding of a single response regulator to any of the clusters is assumed to be exponential, the mean first-passage time for many receptors follows, τ_{RE}/N_{RR}.

3.5.3 Round-trip: convolution

We concluded that due to the small size of the binding sites, the FPT *pdf* for searching the membrane-bound sensor(s) and the subsequent search of the DNA site can be approximated by the memoryless exponential distribution:

$$f(t) = \frac{e^{-t/\tau}}{\tau}, \tag{3.13}$$

where τ is the mean of this distribution: the mean first-passage time.

Eventually, we are interested in the mean and variance of the multi-particle FPT *pdf* for the whole process: cytoplasm → sensor → DNA. By *multi-particle* we mean the distribution for reaching DNA by the first out of N_{RR} independently diffusing response regulators. We assume that the first binding of a transcription factor is the one effecting gene regulation.

As we shall see further in this section, the single-particle *pdf* for the two-step search is not an exponential function. Hence, the mean for the superposition of many particles does not simply scale as $1/N_{RR}$; it follows from order statistics shown in the next section. This is why we need to obtain the full *pdf* for the full search process. Since the probability to find the DNA site is the probability to bind the sensor *and* then search the DNA, the full

FPT *pdf* follows from the convolution of individual *pdfs* (Fig. 1.10 in the Introduction),

$$\begin{aligned} f_A(t) * f_B(t) &= \int_0^t \frac{e^{-t'/\tau_1}}{\tau_1} \cdot \frac{e^{-(t-t')/\tau_2}}{\tau_2} dt' \\ &= \frac{e^{-t/\tau_1} - e^{-t/\tau_2}}{\tau_1 - \tau_2}. \end{aligned} \tag{3.14}$$

Moments of such a convolved distribution are:

$$\langle t^k \rangle = \int_0^\infty t^k f_A(t) * f_B(t) dt = \frac{k!}{\tau_1 - \tau_2} \left(\tau_1^{k+1} - \tau_2^{k+1} \right) \tag{3.15}$$

From that we obtain the first moment, which is simply the sum of the means $\tau_1 + \tau_2$. Noise, defined as variance over squared mean is always smaller than 1, i.e. noise of the pure Poisson process:

$$\eta^2 = \frac{\langle \Delta t^2 \rangle}{\langle t \rangle^2} = \frac{\tau_1^2 + \tau_2^2}{\tau_1^2 + \tau_2^2 + 2\tau_1\tau_2}. \tag{3.16}$$

3.5.4 Moments of the first order statistics of a convoluted *pdf*

After substituting Eq. 3.14 into Eq. A.13 we obtain the explicit FPT *pdf* for the superposition of N independent single-particle functions:

$$f^{(1,N)}(t) = N \frac{e^{t/\tau_1} - e^{t/\tau_2}}{\tau_2 e^{t/\tau_1} - \tau_1 e^{t/\tau_2}} \left(\frac{\tau_1 e^{-t/\tau_1} - \tau_2 e^{-t/\tau_2}}{\tau_1 - \tau_2} \right)^N \tag{3.17}$$

The m^{th} moment of the first-passage time out of N particles for a two-step process with exponential waiting times for each of the steps reads:

$$\langle t^m \rangle = m \left(\frac{1}{\tau_1} - \frac{1}{\tau_2} \right)^{-N} \sum_{n=0}^{N} \binom{N}{k} \frac{\left(\frac{1}{\tau_1}\right)^{N-n} \left(-\frac{1}{\tau_2}\right)^n}{\left(\frac{n}{\tau_1} + \frac{N-n}{\tau_2}\right)^m} \tag{3.18}$$

3.5.5 Bias of the optimal search time

We use Eq. 3.18 to study the total search time (cytosol - receptor - promoter) as function of the number of receptors and regulators. To achieve that, we shall substitute τ_1 with Eq. 3.12, and τ_2 with Eq. 3.6. For typical *E.coli* parameters ($R_{cell} = 1\,\mu m$, $D_{RR} = 5\,\mu m^2/s$, $r_{react} = R_{RE} + R_{RR} = R_{DNA} + R_{RR} = 0.005\,\mu m$), $\tau_1 = 42/N_{RE}\,s$ and $\tau_2 = 13.3\,s$. For a fixed number of REs and RRs the minimal search time is achieved when the number of regulators exceeds slightly the number of receptors (Fig. 3.11, solid red line). Here, we assumed that both steps of the search time are diffusion-limited. In reality, these processes might be longer due to additional time required to undergo a chemical reaction once a regulator and a sensor, or a regulator and the DNA-binding site are *close* to each other. In order to account for that, one needs to add the reaction time to τ_1 and τ_2. In Fig. 3.11 we study the effect of this increase on the optimal search time. Incrementing the first search, τ_1, shifts the optimum towards the 50/50 ratio of regulators and receptors. Conversely, if DNA search τ_2 lasts longer, the shortest search time is always achieved with more regulators; their required count increases with τ_2.

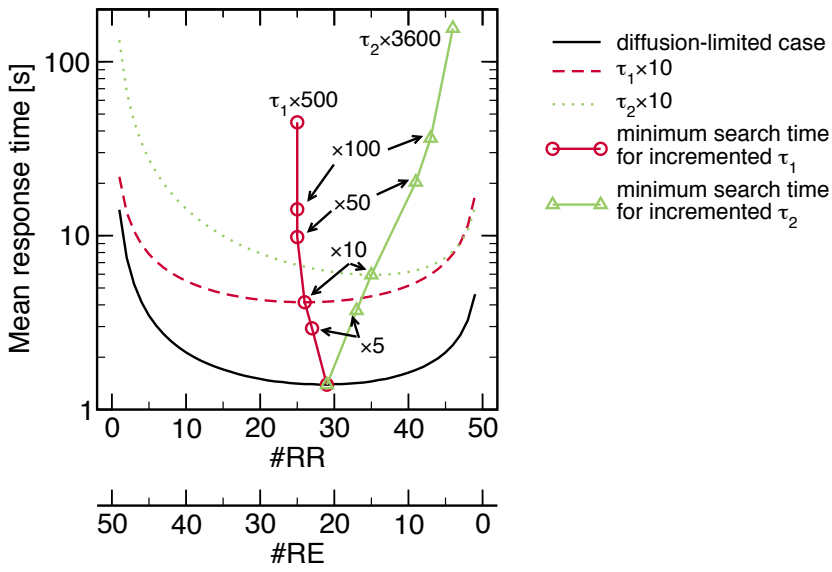


Figure 3.11: Minimal search time and bias in protein ratio depend on search times for single steps. The total number of receptors and regulators equals 50. Red solid line, purely diffusion-limited case, Eq. 3.18 with the time to find receptors, τ_1, equal to Eq. 3.12 and the time to find the DNA site, τ_2, equal to Eq. 3.6. In order to reflect the reaction limitation, we increase τ_1 and τ_2.

3.5.6 Time to (de-)activate half of response regulators

A response regulator becomes activated (or deactivated) when it reaches a membrane bound receptor. We assume that the waiting time *pdf* for a single RR to bind to clusters of receptors is an exponential function with the mean given by Eq. 3.12. The mean time for binding of half of RRs is the k-th passage problem, where k is $N_{RR}/2$. Substituting the exponential *pdf* into Eq. A.14 and computing the first moment yields the mean time to activate first k regulators out of N_{RR}:

$$\tau_{RE}^{(k,N_{RE})} = \tau_{RE}\left(H_{N_{RR}} - H_{(N_{RR}-k)}\right), \tag{3.19}$$

where H_N is the harmonic number. Fig. 3.12 illustrates the solution. Similarly to the mean first-passage time, half-(de-)activation time is only weakly affected by fluctuations in the copy number, once the number of molecules exceeds ≈ 40. Scattering receptors uniformly over the membrane is also decreasing the response time making the system faster.

3.5.7 Diffusive flux at the promoter site

Our estimation of search times in the two-component signaling network assumes that it is the first of the regulators that has an effect on gene expression either by activation or repression. In reality this might not be the case implying that the activation time resulting from our analysis is underestimated. As described in the main text, the number

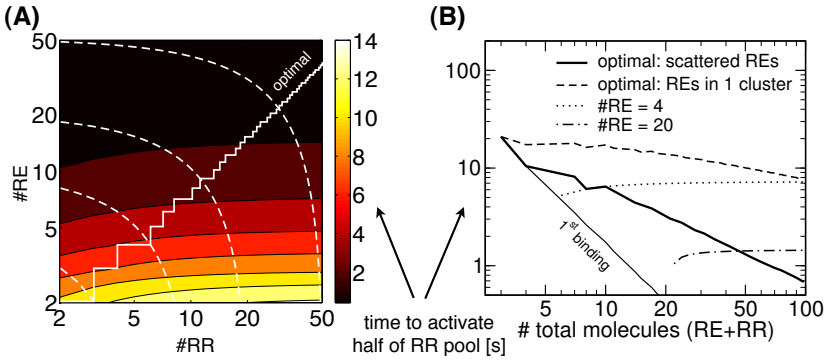


Figure 3.12: Time to (de-)activate half of the response regulator pool (*ON/OFF* time). It equals to the time required for half of freely diffusing RRs to reach the membrane receptors (Eq. 3.19). It is the same for inactive RRs yet to be phosphorylated at the membrane, and for active RRs yet to be dephosphorylated at the membrane. Consistent with biological evidence, the lifetime of an active RR is assumed to be much longer than the search time [West & Stock 2001]. All receptors maintain their kinase (or phosphatase) activity during the RR activation time. (A) Phase plot. The optimal number of RRs and REs yielding the shortest search time is indicated in white. (B) *ON/OFF* time as function of the total number of signaling components for scattered and clustered receptors. Additionally, two curves for fixed number of receptors are shown in red and green. For a comparison, we draw the times for the first RR to bind to any of the receptors (light black line).

of regulator bindings prior to alternation of gene expression can be estimated as $1/P_{ini}$, the inverse of the transcription initiation probability. $P_{ini} = k_{ini}/(k_d + k_{ini})$ depends on the transcription initiation rate constant, k_{ini}, and the life time of the regulator-promoter complex, $1/k_d$. Whether $1/P_{ini}$ indeed corresponds to the number of distinct regulators binding to the promoter site before transcription initiates depends on the following. The regulator which fails to change gene expression by unbinding before the transcription commences, has a probability of rebinding before the next regulator arrives. Even for the number of molecules as high as 100 the time between arrivals of consecutive distinct regulators is as high as 0.2 seconds (3.13. Within that time the probability of rebinding is very high for a diffusion-limited process and equals $\approx 90\%$. Therefore, the first regulator is most likely to perform more binding attempts. Whether it succeeds depends on the forward reaction rate constant, the aspect that we do not investigate here.

3.5.8 Relation between the number of regulators and the number of DNA-binding sites

So far our analysis considered only a single DNA target for regulators. In reality the number of targets may reach even hundreds. How many response regulators is required to activate all of them within given time constraints? Our analytical framework provides a rough estimate. We assume that all DNA targets are located in the center of the spherical cell and have the same radius. A diffusive flux at the center due to phosphorylated regulators will cause subsequent activations of these targets. The first regulator out of

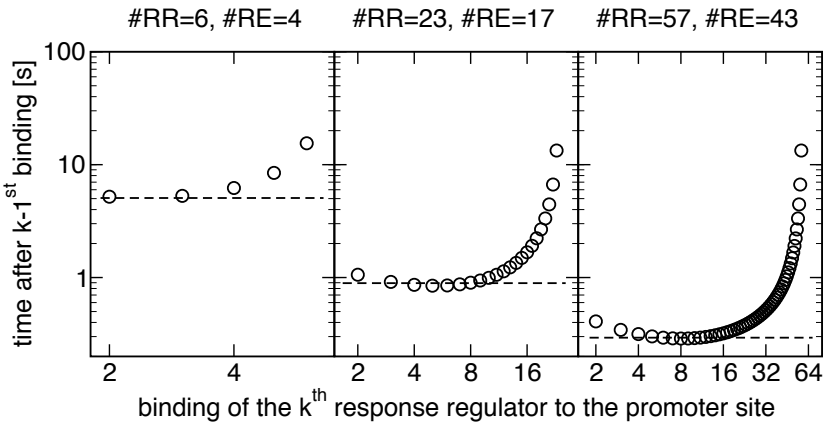


Figure 3.13: Time between bindings of successive regulators to the promoter (both searches included). We substituted Eq. 3.14 into Eq. A.14, computed numerically the mean for the k^{th}-passage, and finally differences between k^{th} and $k-1^{st}$ passage. Straight solid line indicates a regime where intervals between subsequent bindings of phosphorylated regulators are approximately equally timed akin to behavior of a steady state flux.

N_{RR} induces activation of the first out of N_{DNA} targets, the second regulator activates the second target, and so forth. Activation implies binding which effectively removes the regulator from available diffusing pool. It is a k^{th}-passage problem and the question we are addressing is the number of total regulators necessary to activate N_{DNA} targets within time τ_{act}. The number of receptors is fixed in this setting. Fig. 3.14 illustrates this for two cases of receptor count, 10 and 50, and for two τ_{act} thresholds.

Here, we shall derive an approximate solution of results shown in Fig. 3.14. There are N_{RR} regulators which, after reaching the center of the cell, are sequestered on N_{DNA} target sites. The quantity we are looking for is the total number of regulators that allows to bind all target sites within a time τ_{act}. In order to avoid computing the k^{th}-passage of a peaked waiting time distribution (Eq. 3.14) we make the following assumption. The waiting time *pdf* for the whole search process is the exponential function with the mean given by the sum of mean times for individual search processes, $\tau_T = \tau_1 + \tau_2$. Now the model is analogous to the first-order annihilation process of N_{RR} molecules. The corresponding mass-action law for the number of unbound regulators at time t reads,

$$N_{RR} - k = N_{RR} e^{-\frac{t}{\tau_T}}, \tag{3.20}$$

where k is the number of regulators that already reached the targets. We are interested in the situation where $k \equiv N_{DNA}$. Time t is our desired activation time threshold; we shall denote it ϑ. A straightforward transformation shows that the number of regulators is linearly dependent on the number of target sites,

$$N_{RR} = \frac{N_{DNA}}{1 - e^{-\frac{\vartheta}{\tau_T}}}. \tag{3.21}$$

This equation is a good approximation of the results in Fig. 3.14 for small values of τ_{act}.

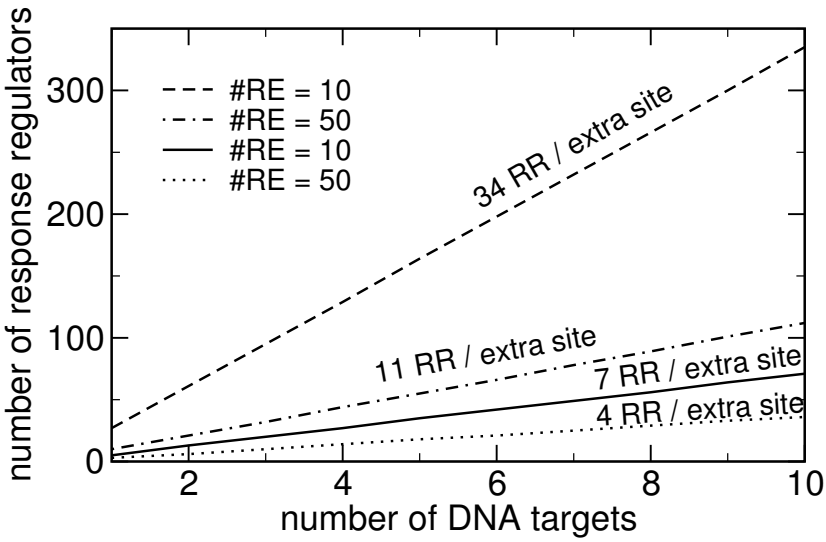


Figure 3.14: The number of regulators required to activate N_{DNA} targets within activation time τ_{act}. We substituted Eq. 3.14 into Eq. A.14, computed numerically the mean for the k^{th}-passage, where k corresponds to N_{DNA}, and solved iteratively for the number of regulators N_{RR} required to obtain prescribed binding time, τ_{act}. The dependence is linear and the slope, indicated along the lines, is the number of additional regulators required to maintain τ_{act} below the threshold if one extra site is added.

The same result can be obtained by substituting the exponential function with the mean τ_T into Eq. A.14. Then, the mean time for k^{th} regulator to be sequestered (the mean k^{th}-passage) reads,

$$\langle t \rangle^{(k,N_{RR})} = \tau_T \left(H_{N_R R} - H_{N_R R-k} \right) \tag{3.22}$$

$$= \tau_T \sum_{i=N_R R-k+1}^{N_R R} \frac{1}{i}. \tag{3.23}$$

In the continuous limit, the above is equivalent to,

$$t = \tau_T \log \left(\frac{N_{RR}}{N_{RR} - k} \right), \tag{3.24}$$

which can be easily transformed into Eq. 3.21 by taking $t = \vartheta$ and $k = N_{DNA}$.

3.5.9 Two sources of stochasticity

The noise in the total search time, $\tau_s = \tau_1 + \tau_2$, consists of two additive terms corresponding to stochastic diffusion and noisy gene expression. For a fixed number of two-component proteins only fluctuations in time due to diffusion are manifested; the search-time probability density function remains the same, so does the mean. Fluctuations in protein production change the shape of the distribution and hence the mean (Fig. 3.15).

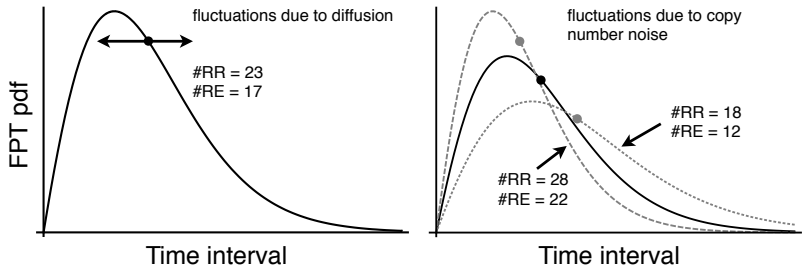


Figure 3.15: Two sources of stochasticity in the mean search time (denoted by full dots) in a two-component network. (Left) For a fixed number of components, search times are described by a peaked probability density function resulting from the convolution of two approximately exponentially-distributed processes. Stochastic nature of the diffusion process introduces fluctuations around the mean. (Right) Protein copy numbers fluctuate due to noise in gene expression. As a result, the speed capacity of the two-component network changes.

3.5.10 Processivity

We shall consider two adjacent genes A and B regulated by a single promoter. Due to incomplete transcription, i.e. RNA polymerase falling off prematurely, a bias in the number of fully completed mRNA transcripts for the two genes may arise. This ratio can be calculated in the following way. Transcript A is finished if polymerase falls off either on or after gene B. mRNA with gene B is obtained only if both genes are transcribed, thus the polymerase may fall off only after gene B. The above can be formalized in the equation:

$$\frac{mRNA_A}{mRNA_B} = \frac{\text{probability to fall off on gene B} + \text{probability to fall off after gene B}}{\text{probability to fall off after gene B}}. \tag{3.25}$$

We assume that the probability to fall off on a particular nucleotide is independent of the sequence and equals p. From the geometric distribution we infer the probability to fall off after transcribing n nucleotides,

$$Pr(N = n) = (1-p)^n p. \tag{3.26}$$

From this we obtain expressions for both probabilities in Eq. 3.25. The first one, the probability to terminate on gene B (produces mRNA with A gene only) is,

$$Pr_A = \sum_{i=N_A}^{N_A+N_B-1} (1-p)^i p = (1-p)^{N_A} - (1-p)^{N_A+N_B}. \tag{3.27}$$

N_A and N_B are lengths of gene A and B, respectively. Note that taking N_A in the lower bound of the sum ensures that transcription terminates on the first nucleotide of gene B (no spacing between genes assumed) thus yielding mRNA with full gene A and incomplete

gene B. Similarly, the upper bound $N_A + N_B - 1$ accounts for termination on the last nucleotide of gene B.

The second term, the probability to terminate after gene B (mRNA with gene A and B is produced) reads,

$$Pr_{AB} = \sum_{i=N_A+N_B}^{\infty} (1-p)^i p = (1-p)^{N_A+N_B}. \tag{3.28}$$

After substituting Pr_A and Pr_B into Eq. 3.25 we obtain,

$$\frac{mRNA_A}{mRNA_B} = (1-p)^{-N_B}. \tag{3.29}$$

The ratio depends solely on the length of gene B!

3.5.11 Numerical simulations

We carried out computer simulations using a custom written C code. Response regulators are modeled as point objects in 3D space confined by a sphere of radius R_{cell}. Membrane receptors are treated as randomly distributed disks of radius R_{RE} on that sphere. Binding between RE and RR takes place if a RR exceeds the cells radius *and* is within a reaction radius $R_{RR} + R_{RE}$ from the center of the RE disk. Similarly, for binding to a DNA site, a regulator needs to be within $R_{RR} + R_{DNA}$ from the DNA site. The site is modeled as a sphere of radius R_{DNA} placed in the center of the cell's volume. The simulation consists of discrete time steps during which regulators are assigned a new position in X, Y and Z direction drawn from a Gaussian distribution of mean 0, and standard deviation $\sqrt{2D\delta t}$, D - diffusion constant, δt - simulation time step. δt is chosen such that on one hand the error due to fixed time step is minimized, and on the other hand the simulations do not run exceedingly long. For instance, for binding targets sized at $R_{RE} = R_{DNA} = 2.5nm$ and for the regulator with diffusion constant $D_{RR} = 5\mu m^2 s^{-1}$ and the radius 2.5 nm we ran the simulation with $\delta t = 1e^{-8}s$. During that time, with a 3-sigma confidence a RR can displace a maximum of 0.9 nm, which is a safe margin given the reaction radius equals 5 nm. Decreasing δt further did not exhibit any appreciable improvement in results.

3.5.12 Bioinformatic analysis

We utilized the MiST database [Ulrich & Zhulin 2007], which contains 934 completely sequenced genomes and their associated signal transduction repertoires, to perform the bioinformatic analysis of two-component systems. To mitigate bias due to overrepresented genomes, we grouped them by their genus and species classification and only included the largest genome from each group. At the species level, 623 unique genomes were considered in this study. We queried the MiST database for histidine kinases with at least one predicted trans-membrane region or signal peptide and response regulators that contain a DNA-binding domain. Hybrid histidine kinases (possess both transmitter and receiver domains on the same polypeptide molecule) and chemotaxis proteins were excluded.

	constraints	count	% of all:	% of all HK-TMs
global	all HKs	14322	100%	N/A
	HK-TM	10133	71%	100%
isolated pair	all pairs	8488	100%	N/A
	RR-DNA	7597	90%	N/A
	HK-TM	7187	85%	71%
	HK-TM & RR-DNA	6670	76%	66%
dicistronic	all pairs	2306	100%	N/A
	RR-DNA	2007	87%	N/A
	HK-TM	1937	84%	19%
	HK-TM & RR-DNA	1765	77%	17%

Table 3.1: Abundance of histidine kinases in 623 unique genomes (out of 934) from MiST database. An isolated pair was defined as a single HK and RR on the same strand, with no intervening genes, with any amount of intergenic space, bounded on both sides by a gene in opposite direction or a non-two-component protein or a 250-bp sequence. HK-TM denotes a histidine kinase with at least one trans-membrane domain. RR-DNA denotes a response regulator with at least one DNA-binding domain.

3.5.12.1 Isolated vs dicistronic pairs

We operationally define an isolated, classical two-component system as a single histidine kinase that putatively interacts with a single cognate response regulator. To identify these at a genomic scale, we scanned each genome for operons containing one histidine kinase gene that is immediately adjacent to a response regulator gene and this pair is bounded by non two-component gene(s) or distinct operon(s). Distinct operons are operationally defined as one or more genes encoded on the same DNA strand each of which is no further than 250 base pairs from its immediate neighbor. Dicistronic operons are specialized operons that comprise only two genes.

3.5.12.2 Identification of RBS sites

We identified ribosome binding sites (RBS) for each genome using a local copy of GeneMarkS version 4.6e [Besemer *et al.* 2001] in conjunction with a modified version of the RBS-finder tool [Suzek *et al.* 2001]. First, a RBS consensus probability matrix is derived by running GeneMarkS against the entire genomic DNA (prokaryotic sequence model). All RBS are then predicted with RBS-finder, which we modified to train its internal Markov model based on this externally derived consensus probability matrix.

5' gene	3' gene	934 genomes	623 genomes		
		all pairings	all pairings	HK-TM & RR-DNA	HK-TM & RR-DNA (dicistronic)
HK	RR	5894 (45%)	4094 (48%)	2690 (40%)	705 (40%)
RR	HK	7074 (55%)	4394 (52%)	3980 (60%)	1060 (60%)
sum		12968	8488	6670	1765

Table 3.2: Count of isolated HK and RR pairings for different gene orders. An isolated pair is defined as in Table 3.1. Bias favoring RR → HK order is higher if only HKs with trans-membrane domain and RRs with DNA-binding domains are considered.

	isolated pair: HK-TM and RR-DNA				all HK-TM
count	6670 (66%)				10133 (100%)
	RBS at 5' gene		RBS at 5' and 3' gene		
count	610 (9%)		5334 (80%)		
	HK → RR	RR → HK	HK → RR	RR → HK	
count	98 (16%)	512 (84%)	2193 (41%)	3141 (59%)	
length of HK [aa]	558 ± 187	470 ± 73	551 ± 171	467 ± 80	
length of RR [aa]	264 ± 95	237 ± 39	283 ± 106	241 ± 47	

Table 3.3: Number of ribosome binding sites (RBS) for pairs of isolated HK and RR. 623 genomes from MiST database were considered. Note that the numbers for 1 and 2 RBSs do not sum to 100%. The remaining cases include an RBS at 3' gene which we omit here. Gene lengths include the mean and the standard deviation.

	dicistronic: HK-TM and RR-DNA				all HK-TM
count	1765 (17%)				10133 (100%)
	RBS at 5' gene		RBS at 5' and 3' gene		
count	146 (8.3%)		1433 (81%)		
	HK → RR	RR → HK	HK → RR	RR → HK	
count	27 (18%)	119 (82%)	584 (41%)	849 (59%)	
length of HK [aa]	577 ± 186	477 ± 79	557 ± 186	473 ± 83	
length of RR [aa]	259 ± 89	238 ± 43	272 ± 101	247 ± 56	

Table 3.4: Number of ribosome binding sites (RBS) for pairs of isolated HK and RR transcribed on a dicistronic mRNA. 623 genomes from MiST database were considered. Note that the numbers for 1 and 2 RBSs do not sum to 100%. The remaining cases include an RBS at 3' gene which we omit here. Gene lengths include the mean and the standard deviation.

Chapter 4

Computational methods for diffusion-influenced biochemical reactions

Contents

4.1 Abstract 81
4.2 Introduction 82
4.3 Regimes and models in biochemistry 83
4.4 Test cases 85
4.4.1 Gene expression 85
4.4.2 Signal transduction 87
4.5 Computational methods 88
4.5.1 BD-level 88
4.5.2 RDME-level 89
4.5.3 CME-based 90
4.6 Results 90
4.6.1 Gene expression 90
4.6.2 Dynamics of gene expression 90
4.6.3 Reversible reaction of an isolated pair 91
4.6.4 CheY diffusion 93
4.6.5 Dynamics of CheY diffusion 94
4.7 Discussion 96

4.1 Abstract

We compare stochastic computational methods accounting for space and discrete nature of reactants in biochemical systems. Implementations based on Brownian dynamics and the reaction-diffusion master equation are applied to a simplified gene expression model and to a signal transduction pathway in *E. coli*. In the regime where the number of molecules is small and reactions are diffusion-limited predicted fluctuations in the product number vary between the methods, while the average is the same. Computational approaches at the level of the reaction-diffusion master equation compute the same fluctuations as the reference result obtained from the particle-based method if the size of the sub-volumes is

comparable to the diameter of reactants. Using numerical simulations of reversible binding of a pair of molecules we argue that the disagreement in predicted fluctuations is due to different modeling of inter-arrival times between reaction events. Simulations for a more complex biological study show that the different approaches lead to different results due to modeling issues. Finally, we present the physical assumptions behind the mesoscopic models for the reaction-diffusion systems.

Input files for the simulations and the source code of GMP can be found under the following address: `http://projects.cwi.nl/sic/bioinformatics2007/`.

4.2 Introduction

There are many examples of biochemical reactions where spatial effects play an important role. In case of gene expression, transcription of a gene involves an encounter of RNA polymerase and transcription factors with a specific place on a DNA strand. The inclusion of diffusive effects is also important in the description of signaling pathways where additional noise due to sub-cellular compartmentalization can cause the signal weakening [Bhalla 2004]. Especially if the reactions are fast, diffusion can be a limiting factor in these processes since the environment is crowded and the dimensions of a cell are large compared to the size of the molecules. Besides, the number of molecules involved can be low which is an additional source of stochasticity. The presence of the stochastic effects in biological systems has numerous consequences. One of them is the appearance of redundancy in regulatory pathways in order to obtain deterministic behavior [McAdams & Arkin 1999]. Fluctuations may also increase the phenotypic heterogeneity which in turn improves the organism's environmental adaptation [Kaern *et al.* 2005, Perkins & Swain 2009, Eldar & Elowitz 2010].

The need for discrete-spatial-stochastic computational methods is apparent when confronted with theoretical studies of biochemical networks. Spatial coupling between chemically reacting systems is known to stabilize the autocatalytic reaction kinetics [Marion *et al.* 2002]. Numerical analyses of a population model [Shnerb *et al.* 2000], a 4-component autocatalytic loop [Togashi & Kaneko 2001], or a simple reaction-diffusion system [Togashi & Kaneko 2004] show the emergence of a new behavior induced by the discrete nature of reactants. A behavior that could not be captured by the continuum approaches, let alone methods without space. The models of calcium wave propagation [Stundzia & Lumsden 1996] and intracellular Ca^{+2} oscillations [Zhdanov 2002], the study of Soj protein relocation in *Bacillus subtilis* [Doubrovinski & Howard 2005] or MinD/MinE protein oscillations in *E. coli* [Fange & Elf 2006] are another illustration where the stochastic effects due to space and discreteness need to be accounted for to explain the experimental results.

In this chapter we focus on some computational approaches that have been published recently and applied to biological systems. Green's Function Reaction Dynamics, a Brownian dynamics-based method has been used to study fluctuations in gene expression [van Zon & ten Wolde 2005a, van Zon *et al.* 2006] and to analyze the gain in a network of two antagonistic enzymes that modify a substrate covalently, a so called "push-pull" network [Morelli & ten Wolde 2008]. Smoldyn, the second of Brownian dynamics-based methods, has been applied in a number of biological settings. Most notably it has been used to simulate signal transduction in *E. coli* chemo-

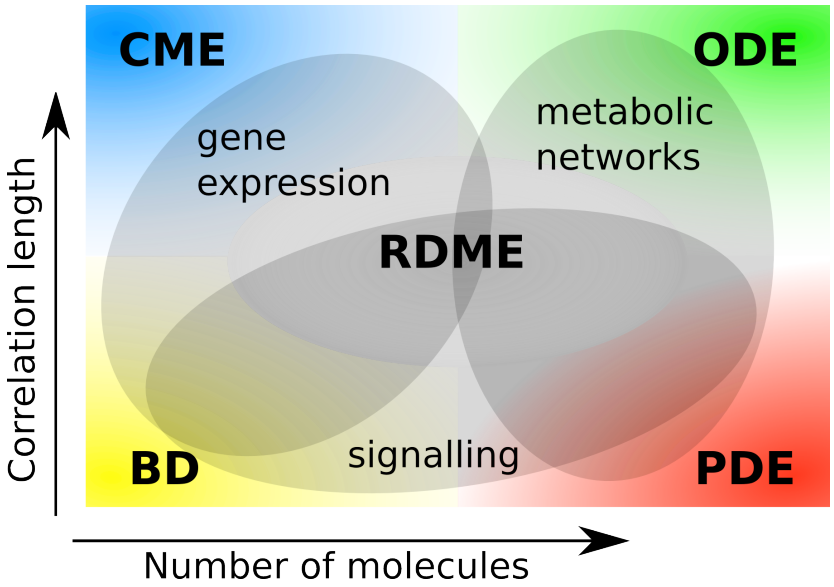


Figure 4.1: Biological problems and relevant models placed in correlation-length versus number-of-molecules space. Abbreviations for (1) models with space: BD – Brownian dynamics, PDE – partial differential equation, RDME – reaction-diffusion master equation, and (2) models without spatial detail: CME – chemical master equation, ODE – ordinary differential equation. ODE and PDE are deterministic models; CME, RDME and BD are stochastic.

taxis [Lipkow *et al.* 2005, Lipkow 2006], generation of cAMP microdomains in synapses [Oliveira *et al.* 2010], and pheromone response system signaling among yeast cells [Andrews *et al.* 2010]. MesoRD and Gillespie Multi-Particle, two implementations of the reaction-diffusion master equation, allowed to study spatio-temporal dynamics of the cellular processes [Elf & Ehrenberg 2004, Fange & Elf 2006, Rodríguez *et al.* 2006]. Finally, the Stochastic Simulation Algorithm by Gillespie has been frequently applied to investigate the influence of noise on biochemical networks [Arkin *et al.* 1998, Kierzek *et al.* 2001, Krishna *et al.* 2005]. A general overview of the main features of the methods can be found in Takahashi et al. [Takahashi *et al.* 2005]. Here we compare the various assumptions in the different models (Tab. 4.1) and the computational results. For clarity we choose a simplified model of gene regulation as a case study. Using this model we make a detailed comparison between the methods. In particular we are looking at fluctuations of the product protein in gene expression. For a more realistic and complex biological system we discuss the influence of the necessary modeling choices.

4.3 Regimes and models in biochemistry

Biological phenomena in a single living cell span over a wide range of spatial and temporal scales. Also the number of molecular species involved can vary significantly. Concentrations of agents in reactions involved in gene expression reach nanomoles, while molecules

Model	Space	Discrete	Extra assumptions
BD	Yes	Yes	–
RDME	Yes	Yes	WM locally
CME	No	Yes	WM
PDE	Yes	No	C, LMA
ODE	No	No	WM, C, LMA

Table 4.1: Models, their main features and assumptions with respect to BD for modeling biochemical systems. Abbreviations: WM – well-mixed system, C – continuum hypothesis for reactants, LMA – Law of Mass Action.

are highly abundant in metabolic pathways (Fig. 4.1).

Current silicon cell platforms can often make reliable predictions for metabolic networks based on ordinary differential equations (ODEs). There it is assumed that concentrations are high and space is not important. Only the rates of the processes determine changes in concentration of the metabolites. When spatial effects come into play, and the correlation length (CL)[1] decreases, indicating that the volume can no longer be treated homogeneous, methods based on partial differential equations (PDEs) are an appropriate approach. PDEs describe the change of continuous concentrations in time and also in the spatial dimension. This can be a good model for biochemical networks where some of the biomolecules are bound to the membrane like in signaling pathways or in eukaryotic cells in general because molecules are localized in compartments such as nucleus, mitochondria, endoplasmic reticulum, etc. In all of these instances possibly significant concentration gradients appear, and a simulation may require spatial methods [Francke *et al.* 2003].

It is known that the process at the very origin of the whole cellular machinery, gene expression, gives rise to fluctuations in the concentration of the final protein products. One of the sources of stochasticity in gene regulation is the low number of DNA-binding proteins which have to find their specific target in order to initiate translation [Halford & Marko 2004]. A low copy number of regulators and the positioning of genes on the chromosomes result in decreasing frequency of gene activation, thus increasing the fluctuations of mRNA [Becskei *et al.* 2005].

The discrete nature of matter as expressed in low-molecule-number conditions violates the continuum hypothesis used in ODEs and PDEs. A model accounting for this is based on the chemical master equation (CME), a deterministic linear ODE for the evolution of the probability density function for a Markov process [van Kampen 1997]. The Markov process models the stochastic transitions between discrete states of the system. In this case stochasticity reflects the fluctuation in the number of reactants' collisions, and hence the fluctuation in the number of molecules participating in a chemical reaction. The CME approach remains valid as long as the system is well-mixed or, equivalently, has a large correlation length.

The question is whether this is a correct assumption when dealing e.g. with gene expression. Since there is a specific binding site which needs to be found by a relatively small number of competing transcription factors, diffusion might limit the process thus

[1]The correlation length (CL) is a measure of the typical length scale at which a system retains its spatial homogeneity. Estimating CL can be a difficult task, since the different subprocesses involved can all have a different CL.

Reaction			Rate
DNA+A	$\overset{k_a}{\rightarrow}$	DNA·A	$3 \cdot 10^9\,\mathrm{M^{-1}s^{-1}}$
DNA+A	$\overset{k_d}{\leftarrow}$	DNA·A	$21.5\,\mathrm{s^{-1}}$
DNA·A	$\overset{k_{prod}}{\rightarrow}$	P+DNA+A	$89.55\,\mathrm{s^{-1}}$
P	$\overset{k_{dec}}{\rightarrow}$	$\emptyset$	$0.04\,\mathrm{s^{-1}}$

Table 4.2: Reaction scheme and parameters associated with the gene expression model. Initially there is one free DNA site fixed in the center and 18 molecules A (corresponds to a 30 nM concentration) diffusing with $D = 1\,\mu\mathrm{m^2s^{-1}}$.

giving rise to larger fluctuations [Metzler 2001]. The probability of a reaction becomes inherently dependent on the distance from the target site. As a result the frequency of diffusion-limited binding events, for times smaller than the typical time needed to cross the volume, has a power-law distribution [Redner 2001] instead of the exponential one used in mean-field approaches as CME. In order to resolve single diffusive encounters between biomolecules a more detailed approach such as Brownian dynamics (BD) is needed. In this approach the solvent is treated as a continuum medium while solute molecules are modeled explicitly in space [Allen & Tildesley 2002]. Their trajectory is described by a random walk due to collisions with the much smaller solvent molecules. Since the majority of degrees of freedom is characterized by the fluctuating force, the computational cost is much smaller than that of molecular dynamics (MD) where the positions and velocities of all atoms or groups of atoms are traced.

Unfortunately brute-force BD is too expensive for whole-cell simulations. Much more promising candidates for a versatile multi-scale framework are methods based on the reaction-diffusion master equation (RDME) – an extension of CME for spatially distributed systems. Space is incorporated by dividing the volume into smaller sub-volumes, which allows to tackle inhomogeneities due to diffusion [Gardiner 1983]. Tracking a single molecule is not possible in this model; unlike in BD, apart from the occupancy of the sub-volumes no exact positions of molecules are stored. Diffusive effects are treated correctly with RDME if the size of a sub-volume is of the order of the correlation length. Small sub-volumes are important if we want to account for fluctuations not only due to the probabilistic nature of chemical reactions but also resulting from rare binding events in diffusion-limited sparse (i.e. low concentration) systems. Obviously such detailed simulations are computationally expensive. Faster computations with large sub-volumes will give only a crude estimation of higher moments, but also the average will not be correct if the sub-volumes' size is larger than the CL and if the reactions are nonlinear.

4.4 Test cases

4.4.1 Gene expression

In order to study fluctuations due to low number of molecules and spatial effects van Zon and colleagues [van Zon & ten Wolde 2005b] used a very simplified model to focus on the first step of gene regulation, reversible binding of polymerase to the operator site. Only

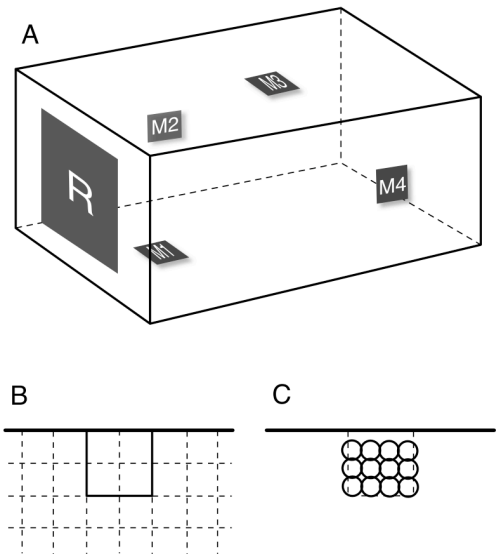


Figure 4.2: Geometry for diffusion of CheYp. (A) Length of the cell is 2.5 μm for Smoldyn simulations. In order to make the number of sub-volume divisions even the length is increased to 2.52 μm for MesoRD and GMP simulations with $L_{sv} = 20$ and 40 nm, and to 2.56 μm at $L_{sv} = 80$ nm. Width and height are 0.88 μm in all of the simulations. Motors are placed on every long sidewall at 0.5, 1.0, 1.5 and 2 μm from the anterior cell wall. Again the positioning of the motors was slightly adjusted in case of simulations with MesoRD and GMP. (B) Scheme of motor discretization for MesoRD and GMP simulations. For $L_{sv} = 20$ nm motors consist of 8 sub-volumes (2x2x2). For 40 and 80 nm motors occupy only one sub-volume. (C) Geometry of the motor for simulations with Smoldyn. Here, the reaction radii have been drawn around the molecules. For the time step equal to 0.2 ms the binding radius for the CheYp association to FliM proteins is 6 nm, hence the overlap of the radii. A total of 34 proteins is placed on the walls of the cuboid of approximate size 40x40x30 nm.

this step is modeled explicitly in space.

The system under consideration is a closed volume V with a DNA binding site fixed in the center surrounded by molecules A diffusing freely with diffusion coefficient D. Once the DNA·A complex is formed with association rate k_a it can either dissociate back to separate DNA and A (with rate k_d) or a protein P can be produced with a production rate k_{prod} with subsequent complex dissociation. In both cases dissociation of DNA·A results in two separate molecules, DNA and A, at contact. The protein further decays at rate k_{dec}. Obviously the single protein production step in this model encompasses both transcription and translation which, as a matter of fact, consist of many biochemical reactions. Protein degradation is also simplified and treated as a first-order reaction. Tab. 4.2 includes the chemical reactions in the model.

The assumption that after protein production molecules are placed at contact is not fully correct if we treat A as a RNA polymerase like in the original study. In fact polymerase travels a certain distance along DNA and unbinds at a position further than the

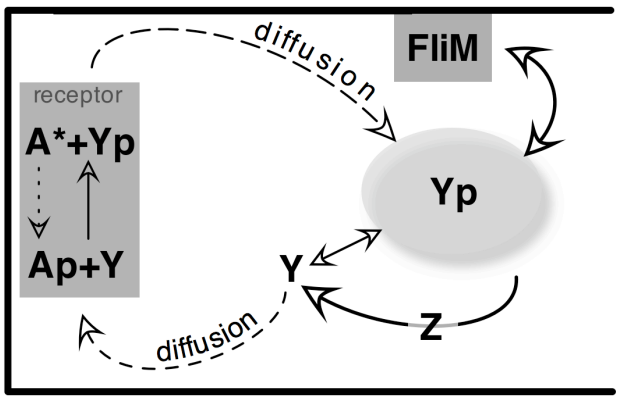


Figure 4.3: Reaction scheme for the diffusion of CheY through the cytoplasm. Prefix *Che* is omitted in the names of the components. CheA* kinase dimers form an array of 1260 chemotaxis receptors inside the anterior cell wall. Dotted arrow in the receptor denotes an autophosphorylation of active dimers, CheA* $\rightarrow$ CheAp. Solid arrow in the receptor is a phosphorylation of CheY. Only one flagellar motor (34 FliM proteins) is depicted here. 8200 CheY signaling molecules (both non- and phosphorylated) and 1600 CheZ dimers diffuse freely in the cytoplasm. The biochemical network is described in the text, reactions and rate constants are listed in Table 4.4. Diffusion coefficients and spatial locations of chemical species are listed in Table 4.3.

initial one. Hence, we would like to remark that the freely diffusing agent A could be a transcription factor or an activator, which are also reported to occur in small quantities, instead of the RNA polymerase. Protein P could be seen as mRNA. In fact the idea of this model is to demonstrate product fluctuations due to rare events where the frequency is diffusion-limited.

We also assume that the molecules perform a pure random walk where the mean square displacement of the molecules is linear with time. The diffusion coefficient and the reaction rates are taken constant. Therefore we do not consider anomalous diffusion due to molecular crowding or hydrodynamic effects.

4.4.2 Signal transduction

In our comparison we include also a model for diffusion of phosphorylated CheY in the *E. coli* chemotaxis pathway as reported by [Lipkow *et al.* 2005]. The cell is modeled as a rectangular box of length 2.52 μm, and width and height 0.88 μm (see the detailed scheme of the geometry in Figure 4.2). Chemotaxis receptors are positioned inside the cell at the anterior wall. They form an array of 35 by 36 CheA dimers, which amounts to a total size of the receptor of 510 by 520 nm. Four motors are placed on the long sidewalls of the cell at 500 nm distance from each other. Each motor consists of 34 FliM molecules positioned on the walls of a cube (empty inside) of 40 nm. The cytoplasm contains 8200 CheY signaling molecules (partially in phosphorylated form), and 1600 CheZ dimers, both diffusing freely.

The reaction network is schematically depicted in Fig. 4.3. CheY monomers are phos-

Species	Location	Diffusion constant
CheA, CheA*	Receptor at the anterior end of the cell	Fixed
CheY, CheYp	Cytoplasm	$10\,\mu m^2\,s^{-1}$
CheZ	Cytoplasm	$6\,\mu m^2\,s^{-1}$
FliM	Four motors on sidewalls	Fixed

Table 4.3: Location and diffusion constants for species in the CheY diffusion model.

Reaction	Rate constant	$\Delta t = 0.1\,\text{ms}$	$\Delta t = 0.2\,\text{ms}$
Unimolecular		Reaction probability per time step	
CheA* → CheAp	$3.4 \times 10^{1}\,s^{-1}$	3.4×10^{-3}	6.8×10^{-3}
CheY → CheYp	$5.0 \times 10^{-5}\,s^{-1}$	5.0×10^{-9}	1.0×10^{-8}
CheYp → CheY	$8.5 \times 10^{-2}\,s^{-1}$	8.5×10^{-6}	1.7×10^{-5}
FliM·CheYp → FliM + CheYp	$2.0 \times 10^{1}\,s^{-1}$	2.0×10^{-3}	4.0×10^{-3}
Bimolecular		Reaction radius [nm]	
CheAp + CheY → CheA* + CheYp	$1.0 \times 10^{8}\,M^{-1}\,s^{-1}$	12.8	16.1
CheZ + CheYp → CheZ + CheY	$1.6 \times 10^{6}\,M^{-1}\,s^{-1}$	3.2	4.0
FliM + CheYp → FliM·CheYp	$5.0 \times 10^{6}\,M^{-1}\,s^{-1}$	4.7	5.9

Table 4.4: Details of the reaction network for the CheY diffusion model. Data adapted from Tab. 1 in [Lipkow *et al.* 2005]. The reaction probability per time step for unimolecular reactions and the reaction radius for bimolecular reactions is given for two simulation time steps, 0.1 ms and 0.2 ms.

phorylated at the receptors where the phosphotransfer from CheAp to CheY takes place. Active CheA dimers (CheA*) produced in this reaction are converted back to CheAp in an autophosphorylation reaction. Phosphorylated CheY (CheYp) diffuses in the cytoplasm and binds reversibly to the FliM motor protein. CheYp can be also dephosphorylated by CheZ scavengers diffusing in the cytoplasm or it can autodephosphorylate. Once dephosphorylated, CheY converts to CheYp in a relatively slow reaction or it diffuses back to the receptor to go through the CheA-mediated phosphorylation. Further, CheYp can diffuse and form again the FliM·CheYp complex.

4.5 Computational methods

Here we describe shortly the main features of the algorithms used in the comparison. A more detailed explanation of the models and computational methods can be found in Appendix C. The following versions of computer implementations have been used for the simulations: Smoldyn – version 1.72, MesoRD – version 0.2.1. For GFRD and GMP no version system has been maintained at the time of this publication.

4.5.1 BD-level

Green's Function Reaction Dynamics (GFRD) developed by [van Zon & ten Wolde 2005a] and Smoldyn (Smoluchowski Dynamics) by

Method	Receptor	Motors	Total no. of sv's
Smoldyn	510 x 520 nm, 1 layer at 20 nm from the wall.	40 x 40 x 30 nm, 3 layers at 16, 25 and 35 nm from the wall.	NA
GMP & MesoRD 20 nm	1 x 26 x 26 sv	2 x 2 x 2 sv	243936
GMP & MesoRD 40 nm	1 x 13 x 13 sv	1 x 1 x 1 sv	30492
GMP & MesoRD 80 nm	1 x 7 x 7 sv	1 x 1 x 1 sv	3872

Table 4.5: Geometry and discretization into sub-volumes (sv) of receptor and motors. A total number of 1260 CheA dimers is placed at the receptor. Every motor consists of 34 FliM motor proteins. In case of Smoldyn the reaction rates and the simulation time step (0.2 ms) result in the reaction radius equal to 16 nm for CheA-mediated phosphorylation of CheY, and to 6 nm for binding of CheYp to FliM motor proteins.

[Andrews & Bray 2004], are two particle-based computational methods, which allow to explicitly model the gene expression problem described above. Reacting biomolecules are represented as spheres diffusing freely in a volume, and no excluded volume interactions are assumed.

GFRD uses the analytical solution of the Smoluchowski diffusion equation to resolve the reactive collisions. This allows to increase the simulation time step as compared to the traditional BD approach. This varying time step depending on the nearest neighbor distance is particularly efficient for systems with a low number of molecules. The method is not available as a general tool, and the code has been obtained from the authors.

Smoldyn, on the other hand is a convenient package. It is a more coarse-grained approach to BD simulations of biochemical reactions. The simulation time step is set by the user such that the probability of any reaction event per time step is small. Also the mean square displacement of diffusing molecules must allow for correct treatment of collisions. Here every collision leads to a reaction and the length of the binding and the unbinding radius (larger than the binding distance) for every reaction reproduces the macroscopic reaction rate.

4.5.2 RDME-level

MesoRD [Hattne *et al.* 2005] simulates trajectories of discrete, stochastic systems with space described by the reaction-diffusion master equation. Gillespie Multi-Particle (GMP) [Rodríguez *et al.* 2006] approximates this trajectory by splitting diffusion and reactions into two separate processes. In both cases the simulation volume is divided into sub-volumes, and the number of reactants inside them is recorded. Thus the knowledge of the position of the reacting molecules is limited by the resolution of the space discretization.

MesoRD employs the *next sub-volume method* [Elf & Ehrenberg 2004] in order to identify the region of the domain where the next event triggers. The event can be either a transfer of particles between neighboring cells due to diffusion or a (bio)-chemical reaction.

GMP is based on the Lattice Gas Automata algorithm [Chopard *et al.* 1994] for the diffusion process. The time step in GMP is fixed for every diffusing species and prescribed by the size of the lattice and respective diffusion coefficient. Reactions are executed in every sub-volume between the diffusion steps (different for every species) using Gillespie's

SSA algorithm. Note that the fixed diffusion time step is in fact the average time between diffusion events in the reaction-diffusion master equation. This fact assures that the macroscopic diffusion in GMP is the same as obtained from MesoRD.

4.5.3 CME-based

The widely used Stochastic Simulation Algorithm (SSA) developed by [Gillespie 1976, Gillespie 1977] generates realizations of the Markov process whose probability density function is described by the chemical master equation [Gillespie 1992]. During every step of the simulation two random numbers are drawn from appropriate distributions, and provide time and type of the next chemical reaction. Time assumes continuous values and the state of the system is a discrete number of all components present. Space is not included in the CME model.

4.6 Results

4.6.1 Gene expression

The simplicity of the gene expression model discussed in Section 4.4.1 allows to illustrate how the implementations perform in the regime where spatial fluctuations are important. Additionally it is possible to address the issue of choosing the proper lattice size in RDME methods. The results of the RDME methods are compared to the solution obtained with CME and with the two BD-level simulations, where single molecules are modeled explicitly in space.

We analyze the average of the protein level and its noise η quantified as the ratio of standard deviation over average. Values of the parameters for the simulation are given in Tab. 4.2. Note that protein fluctuations depend on the frequency of the encounters of A and DNA. The influence of space is omitted in computational schemes based on CME such as the SSA algorithm by Gillespie. In that case the distribution of times between successive bindings of A to the promoter site on DNA is exponential because the process is assumed to be dependent on the reactants' concentration and not on their position. For spatially resolved methods the distribution of arrival times changes due to diffusion and results in burst-like behavior of the protein production. Therefore, we consider this problem, despite the great simplification of the gene expression, to be a good setting for the analysis of noise influenced also by spatial effects.

4.6.2 Dynamics of gene expression

The parameters we use in the simulations are such that the production process is limited by diffusion and that new proteins appear in bursts. We anticipate that stochastic effects will be significant under such circumstances. This idea is supported in Fig. 4.4, where the protein level behavior in time is shown. Fluctuations are higher for methods explicitly accounting for space (GFRD and Smoldyn) comparing to the widely used SSA while the averages are the same. The reader will also notice that GFRD yields significantly larger fluctuations than Smoldyn – in principle a method at the same level of detail. The reason for the differences between the two BD methods is explained further in Section

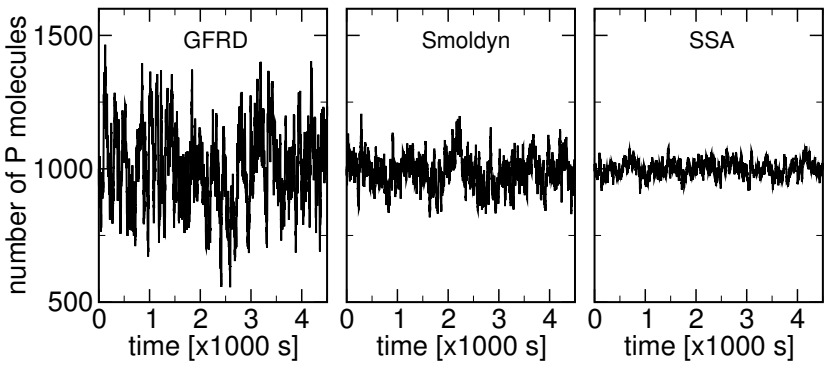


Figure 4.4: Sample simulated time trajectories of the protein level for the gene expression problem. Note that the average is the same for all methods, while the fluctuations are higher for Smoldyn and GFRD, methods that include spatial effects. Parameters as in Tab. 4.2.

4.6.3 and later in the Discussion. For now it is sufficient to say that GFRD produces more trustworthy results since it is an exact method to solve diffusion-limited reversible reactions.

The comparison for the RDME-class methods, MesoRD and Gillespie Multi-Particle, reveals the behavior for an increasing number of space divisions n_{sv} (the number of subvolumes per unit length). We know that for $n_{sv} = 1$, the well-mixed case, RDME methods are equivalent with the chemical master equation which does not include space. In Fig. 4.5 we see that as the spatial detail is increased the predicted fluctuations reach the value given by GFRD, while the average number of proteins is the same, within statistical error, for all lattice sizes.

The fact that MesoRD and GMP are able to reproduce not only the average but also the correct variance of the solution as compared to the reference result obtained from GFRD, shows their capability to give good results in the regime where spatial effects are important (the diffusion-limited regime with small numbers of molecules).

4.6.3 Reversible reaction of an isolated pair

In order to explain the differences in noise level predicted by various methods (Fig. 4.5) we look at a simple example of an isolated pair of molecules undergoing a reversible reaction, the same type of reaction as the first step in the gene expression model. This will allow us to examine the distribution of times between successive reactive events.

The target molecule is fixed in the center of the unit volume V, the second molecule is diffusing freely with diffusion coefficient D. The molecules can undergo a reversible reaction with association and dissociation rates, κ_a and κ_d, respectively. We look at the time between consecutive bindings of the molecules. Simulations with GFRD, Smoldyn, GMP and SSA reveal that the distribution of inter-binding times is different for methods

[2]The reaction mean free path is defined as $\lambda_{\mathtt{RMFP}} = \sqrt{\langle \tau_R \rangle \cdot D_{\mathtt{rel}}}$, where $\langle \tau_R \rangle$ is the mean time between reactions, $D_{\mathtt{rel}}$ is the relative diffusion coefficient [Togashi & Kaneko 2005, Baras & Mansour 1996, Rodríguez *et al.* 2006].

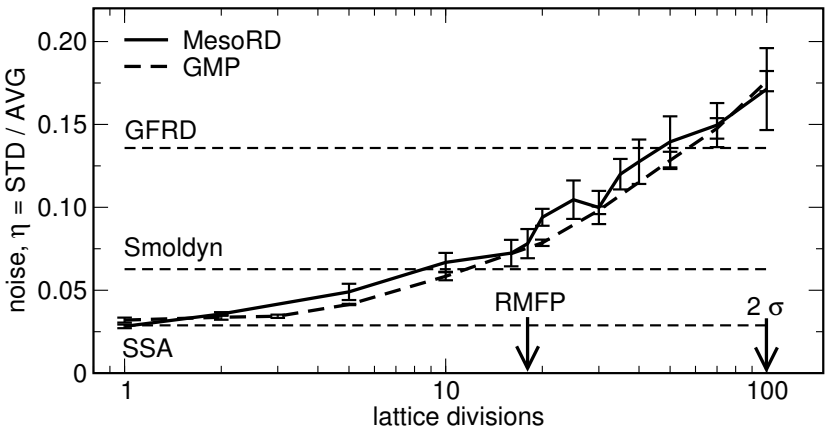


Figure 4.5: Noise in the protein level as a function of n_{sv} – the number of sub-volumes per unit length, for RDME-level methods. The noise for SSA, Smoldyn and GFRD is computed from the experiment of Fig. 4.4. RMFP – the reaction mean free path[2], the estimate of the correlation length; 2σ – size of the sub-volume equal to two reaction distances which corresponds to two molecules fitting in one sub-volume. Noise is proportional to the square root of n_{sv}. Average protein level is the same for all experiments.

with and without space, and also that methods with spatial detail treat the diffusion-limited reactions differently.

It is known [Redner 2001] that for a particle diffusing in an infinite three-dimensional space the probability that it reaches a specific target at a specified time (the first-passage probability) has a power-law distribution. This is depicted in the log-log plot in Fig. 4.6. All spatial methods reproduce the power-law behavior (straight line) for times shorter than the average time needed to approach the boundary; a classical result for diffusion in an infinite 3D space. On such a short time scale the molecule is not influenced by the finite boundary because it simply did not have time to travel that distance. On the other hand, diffusion in a finite volume results in an exponential decay of the first-passage probability for long times. Exponential behavior is characteristic to mean-field approaches like the chemical master equation, where the time of a next reaction is independent of the molecules' position (the well-mixed assumption). Note that, the exponential part of the result obtained with GFRD can be reproduced by changing the forward reaction rate in SSA from the intrinsic rate κ_a to the overall *on* rate coefficient $k_{\mathtt{ON}}$ [Agmon & Szabo 1990, van Zon & ten Wolde 2005a], which "includes" the time needed to reach the target by diffusion and the time to undergo a chemical reaction.[3] It is important to note that the average of the first-passage distribution SSA($k_{\mathtt{ON}}$) in Fig. 4.6 differs from the rest of the experiments. This fact indicates that the simple change of the reaction rates from the intrinsic to the *overall* rates which include the effects of diffusion will not preserve the average time between bindings. The plot also shows that an increase of spatial resolution of an RDME-class method like GMP results in the distribution which can be as close as desired to the exact solution given by GFRD. For clarity we draw only intermediate distributions with small (5) and average (30) number of sub-volumes per unit volume. The simulation with $n_{sv} = 50$ overlaps with the result from GFRD.

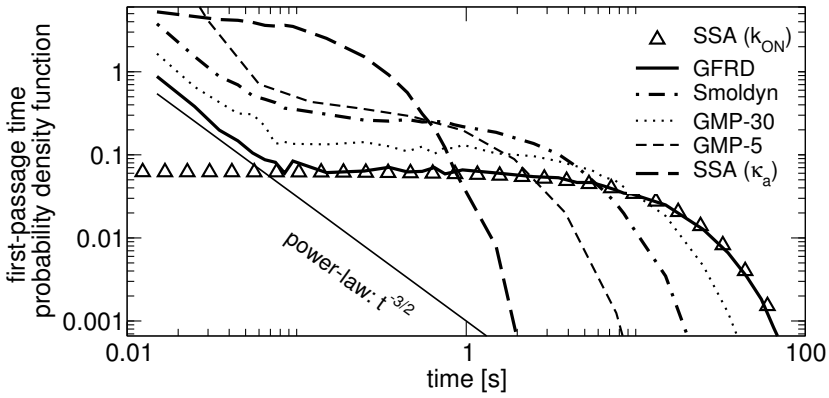


Figure 4.6: Probability density function of time between subsequent bindings of an isolated pair of particles, also known as the first-passage probability. Methods including spatial effects, GFRD, Smoldyn and GMP, reproduce the power-law behavior for short times. For times larger than $\approx 0.1s$, which is the average time needed to reach the boundary, the first-passage probability exhibits an exponential decay. GMP-5 and GMP-30 denote simulations where the whole volume's side is divided into 5 and 30 sub-volumes. The left- and rightmost curves are computed with SSA with the forward reaction rate equal to the *intrinsic* κ_a and to the *overall* $k_{\mathtt{ON}}$ reaction rate, respectively. The averages of all distributions are the same, equal $1/\kappa_a$, except for the $\mathtt{SSA}(k_{\mathtt{ON}})$, where the average is larger and amounts to $1/k_{\mathtt{ON}}$. Note, that the position of the power-law line is chosen arbitrarily; it only compares the slope of data for short times obtained from different methods.

4.6.4 CheY diffusion

We have simulated the chemotaxis pathway in *E. coli* [Lipkow *et al.* 2005] using Smoldyn, MesoRD and GMP. GFRD is omitted in this case study because it is not suited for solving such a complex problem. A detailed description of the geometry is shown in Figure 4.2.

Since Smoldyn allows for placing and tracking every molecule in the system, the implementation of the geometry of the receptor and motors is straightforward: CheA and FliM molecules can be fixed exactly at their positions. However, one needs to make an assumption about the placement of the motor and receptor molecules because Smoldyn does not account for excluded volume interactions and does not have any special treatment of reactions near walls. Following the choice made in [Lipkow *et al.* 2005], we placed the receptor array of CheA dimers inside the cytoplasm, 20 nm from the anterior wall. The FliM motor proteins form a cuboid consisting of three layers at 16, 25 and 35 nm distance from the cell wall. Although molecules are modeled as points in Smoldyn, the macroscopic reaction rates prescribe the microscopic binding radii for every bimolecular reaction channel. For the given parameters and a simulation time step of 0.2 ms the binding distances are 16 nm for CheY phosphorylation at the receptor, 4 nm for CheZ-

[3]The overall *on* reaction rate equals $\frac{1}{k_{\mathtt{ON}}} = \frac{1}{\kappa_a} + \frac{1}{K_D}$, where κ_a is the intrinsic association rate, and K_D is the diffusion-limited reaction rate given by $4\pi\sigma D$, dependent on the reaction distance σ and the relative diffusion coefficient D of two reacting molecules. Note that inverses of rates are equivalent to quantities with a dimension of time.

Method	Comments	T_{M1}	T_{M2}	T_{M3}	T_{M4}
Smoldyn	$\Delta t = 0.2\,\mathrm{ms}$	0.11	0.19	0.22	0.29
MesoRD	$L_{sv} = 20\,\mathrm{nm}$	0.06	0.10	0.15	0.21
MesoRD	$L_{sv} = 40\,\mathrm{nm}$	0.06	0.11	0.15	0.22
MesoRD	$L_{sv} = 80\,\mathrm{nm}$	0.06	0.10	0.14	0.19
GMP	$L_{sv} = 40\,\mathrm{nm}$	0.06	0.08	0.17	0.23

Table 4.6: Average time (in seconds) to reach motor occupancy of 10 CheYp molecules. Results are averaged over 10 runs for every method.

mediated dephosphorylation, and 6 nm for binding to the FliM proteins. Smaller time step of 0.1 ms did not affect the results.

To reach the same level of spatial detail with RDME methods like MesoRD and GMP, a very fine discretization is required, since the exact positions of the molecules are known only up to the size of the sub-volume. We performed simulations where the side of the sub-volume L_{sv} equals 20, 40 and 80 nm. For $L_{sv} = 20\,\mathrm{nm}$ the receptor consists of one boundary layer of 26 by 26 sub-volumes. In this case a cube constructed out of 8 sub-volumes approximates a motor. Four such cubes are positioned on the long sidewalls of the cell's surface. When bigger sub-volumes are used, motors occupy only one sub-volume. The receptor is a one-layer array of 13x13 and 7x7 sub-volumes for $L_{sv} = 40$ and $80\,\mathrm{nm}$, respectively.

4.6.5 Dynamics of CheY diffusion

In the computer simulations we first measure the time required to reach a given level of motor occupancy. Initially all CheA dimers in the receptor are in phosphorylated form, CheY and CheZ are freely diffusing in the cytoplasm. The results in Fig. 4.7 show averages over 10 runs of 1 second for every method. The number of motor-bound CheYp is growing visibly slower for motors placed further along the cell. The time required to reach a threshold of 10 CheYp molecules bound to a FliM motor cluster predicted by MesoRD is the same within statistical error for all three sizes of the sub-volumes (Tab. 4.6). GMP produces slightly higher averages which can be attributed to the splitting error between reaction and diffusion. Molecules diffuse in "bursts" due to the fixed diffusion time step which affects the first-passage properties of the diffusing front while the macroscopic mean square displacement is reproduced correctly. Finally, the average time to reach the given threshold is noticeably higher for Smoldyn. This difference cannot be explained by a wrong treatment in RDME of the nonlinear reactions due to sub-volumes larger than the correlation length. Simulations for different sizes of the sub-volumes yielded the same results within statistical error. We contribute the discrepancy between Smoldyn and other methods to the difference in the modeling of receptor and motors. In MesoRD and GMP a reaction may occur when reactants are in the same sub-volume. For BD-based methods like Smoldyn two reacting molecules need to be within the binding radius in order for an event to occur. This is a stricter constraint because it is possible that two molecules may simply pass each other despite being in a very close vicinity which would result in a reaction in RDME-level methods (unless the discretization is such that each sub-volume contains just one motor molecule).

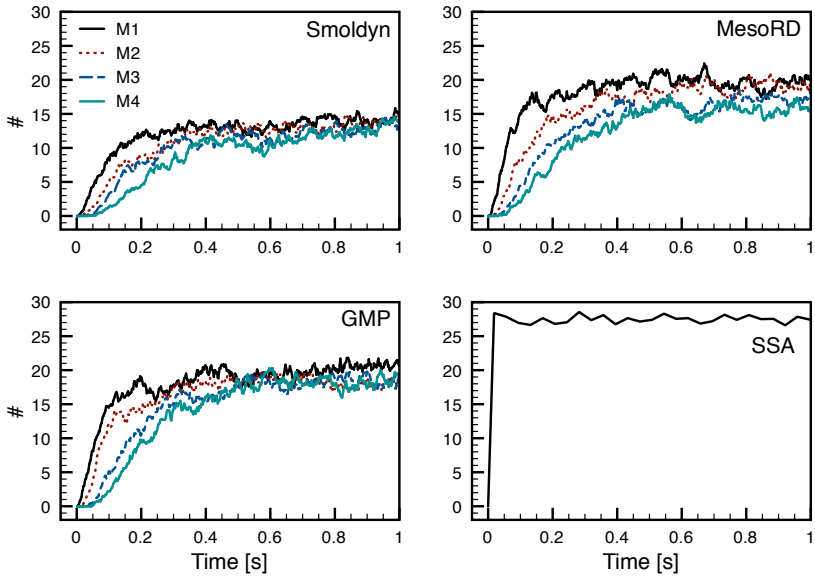


Figure 4.7: Change in motor occupancy in time. Time step used in Smoldyn is 0.2 ms. The side of the sub-volume L_{sv} in MesoRD and GMP is 40 nm. Results do not change when $L_{sv} = 20$ or 80 nm is used.

Another property of the CheYp diffusion model we have studied is the average and the noise in the motor occupancy in steady-state (Tab. 4.7). For MesoRD and GMP we pick the simulations with 40 nm sub-volumes. Both RDME-level methods yield very similar results, although again averages from GMP are slightly higher than those obtained from MesoRD; Smoldyn computes approximately 20% lower averages. Note the interesting effect regarding noise in the motor occupancy, which increases for motors placed further from the receptor. This behavior of noise can be attributed to the concentration gradient of CheYp (high at the receptor and low at the posterior end). A smaller CheYp concentration at motor four compared to motor one results in a drop of the average motor occupation, and causes the fluctuations in binding to the FliM cluster to increase.

Additionally, we provide the SSA result for the motor occupancy (lower-right plot in Fig. 4.7 and Tab. 4.7) with the same reaction rates as in the other simulations. The

	M1		M2		M3		M4	
Method	$\langle N \rangle$	η	$\langle N \rangle$	η	$\langle N \rangle$	η	$\langle N \rangle$	η
Smoldyn	13.9	0.20	13.2	0.21	12.5	0.23	12.4	0.23
MesoRD	19.1	0.17	18.0	0.18	16.5	0.20	16.0	0.22
GMP	20.3	0.15	19.0	0.16	18.6	0.16	18.1	0.16
SSA	27.5	0.08						

Table 4.7: Average and noise in the level of motor occupancy in steady-state. Averages were computed from simulations of length 21 s after the equilibration period of 1 s. Parameters of the simulations are the same as in Fig. 4.7. MesoRD and GMP used $L_{sv} = 40$ nm.

occupancy for only one motor cluster is shown because all of them are equivalent if space is not included. This is clearly a wrong approach to model CheY diffusion, nevertheless it gives an indication of the error one can make when spatial information is omitted either by not accounting for geometry or by lack of correction in the diffusion-limited reaction rates. The average occupation which is higher than in the other methods is a direct consequence of a lack of delay due to the diffusion of CheYp towards the motors. Lower noise, on the other hand, results from the constant, and immediate supply of reactants while in case of simulations accounting for space, the supply is considerably lower due to the appearance of the CheYp concentration gradient.

4.7 Discussion

The computational methods for modeling biochemical systems with single-particle and spatial detail compared in this study are based on Brownian dynamics and reaction-diffusion master equation models. In principle they are all suited for application to systems with a low molecule number and a short correlation length. The need for methods in this regime is steadily increasing as new results of experiments on biochemical reactions in single biological cells appear [Pedraza & van Oudenaarden 2005, Rosenfeld *et al.* 2005, Acar *et al.* 2005, Golding *et al.* 2005]. A theoretical study of simple systems as the gene expression shows that fluctuations arise in diffusion-limited processes not only due to the small number of reactants but also resulting from spatial effects [van Zon *et al.* 2006].

In the comparison for the simple gene expression test case we show that not all methods compute fluctuations correctly, although the average is the same as those given by the mean-field models (CME, ODE, PDE). Note that in general, if reactions are nonlinear, one cannot expect RDME methods to give a correct estimate of the average when the sub-volumes are larger than the correlation length because concentration gradients are not represented within a sub-volume. Smoldyn yields much smaller fluctuations compared to GFRD, even though both methods are BD-based (Fig. 4.4). The reason for the incorrect prediction of the second moment by Smoldyn lies in the way it deals with diffusion-limited reversible reactions. The assumption that every collision is reactive leads to the introduction of the *unbinding radius* such that it reconstructs the macroscopic geminate recombination probability (Section C.2, Appendix C provides more details on Smoldyn). By doing so part of the spatial fluctuations is "averaged" and the resulting first-passage probability lies between the exact solution of the Smoluchowski diffusion equation obtained with GFRD and the mean-field result from SSA for the well-mixed system (Fig. 4.6).

The methods at the level of the reaction-diffusion master equation, MesoRD and GMP, are able to predict correctly the fluctuations. The key issue here is to choose the space discretization (division into sub-volumes) such that the size of a single sub-volume is of the order of the system's correlation length. Otherwise the assumption about the local independence of the reaction probability from the inter-particle distance does not hold. If the requirement of well-mixed sub-volumes is not satisfied, spatial fluctuations are averaged which is clearly visible in Fig. 4.5. For a small number of volume sub-divisions both the noise in the protein level and the distribution of times between bindings approach the prediction from the CME model. On the other hand, if the number of sub-volumes increases up to the limit where two molecules fill completely one sub-volume[4], the first-passage probability gradually recovers the desired characteristics typical for the diffusion

process in a closed volume: power-law behavior for short times and exponential decay for long times (Fig. 4.6).

A word of caution about the notion of *exact prediction of fluctuations* needs to be added here. Although we treat noise computed with GFRD as a reference value, one should bear in mind that this is a result of fluctuations with a rather simple Brownian dynamics model for chemical reactions. We are ignoring here other, possibly important, microscopic effects, like hydrodynamic interactions, electrostatic forces or molecular crowding. These certainly affect the diffusion process [Echeveria *et al.* 2007] but their significance for enhancing noise in biological systems is an open issue. Methods like GFRD which solve numerically the Smoluchowski model for diffusion-limited chemical reactions will provide an upper bound for the magnitude of fluctuations if compared to mesoscopic methods based on the master equation. The latter contain additional physical assumptions (Tab. 4.1) in order to simplify computations at the cost of averaging microscopic phenomena.

In Section 4.6.3 we argue that for a reversible reaction of a pair of particles the methods reproduce the first-passage probability differently, which is the cause of the variation in noise for the gene expression case. The power-law behavior for short departures from the target diminishes as the spatial detail is decreased, which is equivalent to: increasing the size of sub-volumes in MesoRD and GMP, increasing the difference between the binding and the unbinding radius for the reversible reaction in Smoldyn, and obviously the well-mixed postulate in SSA. The power-law region constricts also with increasing number of molecules or with accelerating the diffusion process (not shown in this study). Then the system moves away from the *low-molecule-number* and the *short-correlation-length* regime and the distribution converges to the mean-field exponential behavior. This can be properly approximated either by RDME-level methods with a coarse discretization or simply by the SSA algorithm. If the *overall* forward reaction rate $k_{\mathtt{ON}}$ is taken instead of the intrinsic κ_a, the SSA is also able to reproduce the exponential decay of the first-passage probability equal to the one obtained from GFRD or RDME methods with a large number of sub-volumes. However, for obvious reasons the power-law part is not recovered (in SSA the next event is drawn from the Poisson distribution), and hence the average inter-binding time is larger than that of GFRD.

Whereas the previous examples show the biophysical behavior of the methods for diffusion-limited reactions, we discuss in Sections 4.6.4 and 4.6.5 a more realistic biological problem, the chemotaxis pathway in *E. coli* [Lipkow *et al.* 2005, Lipkow 2006]. There it is shown that modeling aspects and their consequences for the computational approaches can result in different predictions of averages and noise.

Qualitative estimates addressing the computational cost of the mesoscopic methods considered here are given in Tab. 4.8. GFRD and Smoldyn scale in a manner typical for BD-based methods. Their main computational cost lies in computing the next position for every molecule (involves drawing few random numbers) and computing distances between reacting molecules, typically a N^2 operation if no neighbor list technique is applied. Differences in computational time may arise, however, because GFRD is an event-driven

[4] In the physical picture, if the size of the sub-volume is further decreased, the probability of finding two molecules, which in reality have a certain diameter, is zero. This can be done in RDME-level methods since they do not model single molecules but their population in sub-volumes. Such an operation is not physical, however (see section Materials and Methods in [Fange & Elf 2006]).

Method	Computational cost	Comments
GFRD	$\sim \sum_S N_S$	Diffusive movements.
(event-driven)	$\sim \sum_{N_R} \prod_{S \in R} N_S$	Reactive distances.
Smoldyn	As GFRD	As GFRD.
(fixed time step)		
MesoRD	$\sim \log N_R$	[Gibson & Bruck 2000].
(event-driven)	$\sim \log N_{sv}$	A sub-volume adds
		a diffusive reaction.
	$\sim \langle \tau_D \rangle^{-1}$	
	$\sim \langle \tau_R \rangle^{-1}$	
GMP*	$\sim \sum_S N_S$ **	Diffusive movements.
	$\sim N_R$	As in the SSA.
	$\sim \tau_D^{-1}$	Fixed diffusion time step.
	$\sim \langle \tau_R \rangle^{-1}$	

Where:
$\langle \tau_D \rangle \propto L_{sv}^2/D \cdot N_{sv}/N_S, \quad \langle \tau_R \rangle \propto L_{sv}^3/k_R \cdot \prod_{S \in R} N_{sv}/N_S,$
$\tau_D \propto L^2/D \cdot N_{sv}^{-2/3}.$

* The scheme is event-driven for reactions but the diffusion time step τ_D is fixed. The diffusion time step is assigned for every diffusing species.
** If $N_S/N_{sv} > 90$ molecules are moved in bulk, otherwise one-by-one in τ_D.
Note that for event-driven schemes the cost of diffusive movements or of computing reactive distances is given per iteration time step.

Table 4.8: Scaling of the computational cost for the spatial discrete methods presented in this comparison. Abbreviations: N_S – number of molecules of a given species, N_R – number of reaction channels, N_{sv} – the total number of sub-volumes, $\langle \tau_R \rangle$ – average time between reactions, $\langle \tau_D \rangle$ – average time between diffusive movements, D – diffusion coefficient, k_R – rate of reaction R, L_{sv} – length of the sub-volume, L – length of the total volume.

scheme while Smoldyn uses a fixed time step. Choosing the right Δt in the latter is not a completely arbitrary procedure since one has to assure that the probability of events per time step is *small*. In GFRD the maximum simulation time step during an iteration depends on the distance of the molecules to the target. If the total number of molecules decreases, the inter-particle distances increase, thus making a larger time step possible. Using similar arguments one can explain differences in the cost of performing diffusion in MesoRD and GMP (the first is an event-driven scheme, the latter uses a fixed Δt). Obviously the number of molecules of a given species in the sub-volume has to be used instead of the inter-particle distance. Then, the average time between diffusive jumps, $\langle \tau_D \rangle$, in MesoRD is inversely proportional to that quantity (see the caption of Tab. 4.8). Additionally, thanks to the next sub-volume method, MesoRD finds sub-volumes where the next event occurs instead of looping over the whole volume. On the other hand, GMP favors higher densities because, contrary to all other methods, particles can be diffused

in bulk rather than one-by-one. The computational cost of the two RDME-level methods differs also in scaling with the number of reaction channels N_R. The usage of the SSA scheme in GMP results in linear scaling with N_R; MesoRD achieves approximately $\log N_R$ scaling. Note that a diffusion event in the latter method is treated similarly as a reaction, and is also entered into an event queue.

Results for the test cases give an additional indication of performance. In case of the gene expression model general tools like Smoldyn, MesoRD or GMP will not outperform significantly the supposedly more expensive GFRD since they are not optimized for this very specific problem. On the other hand GFRD needs tailoring to every new problem and in general is implementation-wise difficult to adjust in order to tackle bigger problems. Computations of the CheY model showed a similar performance of all the methods. Smoldyn was only approximately twice slower than MesoRD and GMP, and appeared to be the most flexible method from the modeling point of view. For example, it allows to construct a more realistic, disc shape of the FliM motor cluster without any additional efficiency penalty. Such geometry of the motor implemented in RDME-level methods requires a much finer spatial resolution, which adds a significant computational cost.

Acknowledgments

The authors would like to thank Jeroen van Zon and Marco Morelli for the code of GFRD, Karen Lipkow for the Smoldyn input files for the CheY diffusion model. This work was supported by the Netherlands Organization for Scientific Research, project NWO-CLS-635.100.007, and by the Dutch BSIK/BRICKS project.

Chapter 5

Discussion

The following topics were discussed in the thesis: (i) the effect of stochastic motion of motor proteins along biopolymer on gene expression bursts (Chapter 2), (ii) the effect of protein abundance, variability and localization on response time in bacterial signaling (Chapter 3), (iii) the emergence of the exponential regime in stochastic processes with non-exponential waiting times (Chapters 1 and 3), (iv) the comparison of theoretical and computational models for stochastic, spatially-resolved biological problems (Chapter 4). Below we shall summarize and discuss the main findings.

Bursty gene expression. We suggest that bursts in mRNA (or protein) production may result from stochastic changes in promoter occupancy as well as from stalling of RNA polymerase (or ribosomes) during their stochastic motion along biopolymers, i.e. transcription and translation. This stochastic motion disturbs or enhances temporal patterns such as transcription initiation bursts. Stochastic bursting, a source of uncertainty in establishing intracellular protein levels, is a potent mechanism of inducing phenotypic heterogeneity in a population of cells [Yu *et al.* 2006, Sigal *et al.* 2006] (Fig. 5.1). Such noise-induced variability in properties of individual cells may confer fitness advantage on the population level in the face of antibiotic exposure [Blake *et al.* 2006], chemotherapy treatment [Cohen *et al.* 2008, Spencer *et al.* 2009, Sharma *et al.* 2010] or general changes in extracellular conditions [Acar *et al.* 2008].

Experimentally verified stalling of RNA polymerase during transcription [Artsimovitch & Landick 2000, Core & Lis 2008, Greive & von Hippel 2005] and ribosome pausing during translation [Hayes & Sauer 2003, Sunohara *et al.* 2004, Buchan & Stansfield 2007] is a target for altering gene expression. Poised polymerases have been argued to play a decisive role in influencing regulation of genes in embryo development [Hayes & Sauer 2003]. Our theoretical findings imply that motor protein motion during transcriptional (or translational) elongation is a potentially additional level of regulating the extent of cellular heterogeneity resulting from bursts in gene expression. Due to collisions of motor proteins during transcription and translation, genes under the control of an activating transcription factor are more likely to induce bursts by means of polymerase or ribosome stalling. Proteins synthesized from strongly suppressed genes, on the other hand, will suffer from stochastic bursts generated at the initiation phase. Whether stalling and pausing during elongation indeed can produce coherent and significant bursts of protein synthesis, which would further translate to physiologically relevant phenotypic differences among individual cells in a population remains to be verified *in vivo*. Even more intriguing is the ability of transcriptional elongation to modulate temporal activity of events generated upstream, e.g. at the transcription initiation stage. Experiments have demonstrated that the frequency of bursts (rather than concentration) of transcription factor activity might also encode

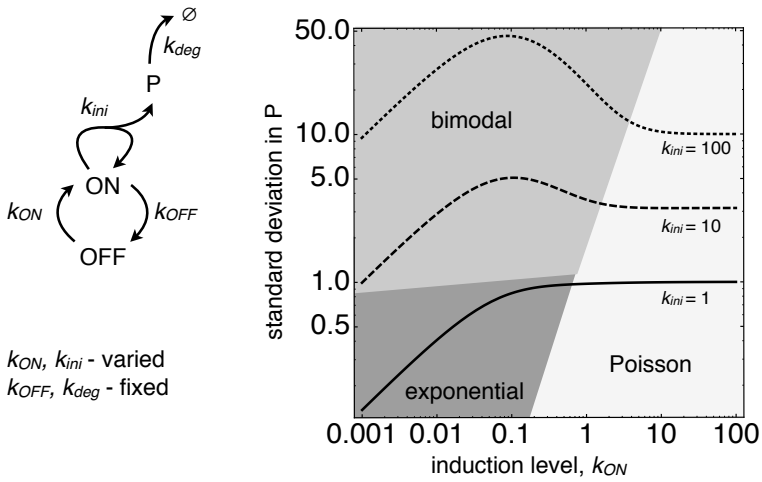


Figure 5.1: The relationship between the standard deviation and the induction level, k_{ON} - the on switching rate of the burster. Indicated are three regimes of the shape of the steady-state P distribution, which correspond to distributions plotted in Fig. 1.7 in the Introduction. Parameter k_{ini} approximates the size of a burst. Bursty protein production resulting in bimodal distribution has been measured in yeast cells and attributed a biological function: an increased cell-cell variability due to bursty gene expression enables cells to respond quickly when challenged with antibiotic stress [Blake *et al.* 2006]. Bimodality in protein distribution emerges only due to stochasticity, in a system without feedbacks and associated bistability, a condition traditionally regarded as necessary to achieve multiple states (compare the two mechanisms in Fig. 1.8 in the Introduction).

information about the extracellular conditions [Cai *et al.* 2008, Ashall *et al.* 2009]. In these cases, physiologically relevant patterns of gene expression corresponding to different cell fates depend heavily on frequency-modulation. To ensure robust information relay, transcription elongation should be in the regime where collisions of motor proteins are negligible and hence unable to impair timescale separation introduced upstream. Disruption of frequency patterns in transcription factor activity such as NF-κB by means of polymerase pausing could be a possible therapeutic target.

The analysis of bursts in the context of gene expression is frequently marred by the inability to elucidate the mechanism responsible for generating such variability in times of mRNA or protein synthesis. Problems originating in other fields, such as telecommunication networks or generic analysis of time series, are affected too. This complication becomes even more serious when the stochastic events in question are very infrequent. Without a prior knowledge of the mechanism generating bursts, e.g. the length of the periods of synthetic activity, it is impossible to group individual occurrences of product arrival into a single burst unambiguously. The same time trace of synthetic activity can be characterized by various burst sizes and durations (Fig. 5.2). Therefore, comparison and quantification of different stochastic bursts is impossible without objective measures.

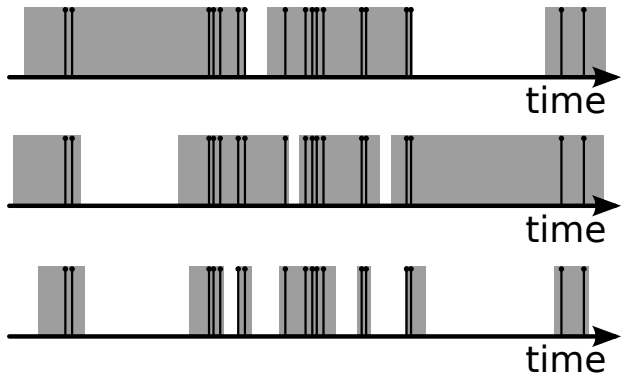


Figure 5.2: Grouping occurrences of synthetic events is ambiguous. If synthetic events are infrequent (vertical lines), apparently close events may result from different continuous periods of activity (dark grey regions).

In Chapter 2, we introduce three indices: *burst size*, *significance*, and *duration*, which are independent of the underlying mechanism of bursts. For the simplest case, bursts generated by an ON/OFF switch (Fig. 5.1), the indices have a clear analytical interpretation. For general cases we provide a straightforward framework to calculate our measures from the analysis of a time trace of product arrivals.

Objective indices allow for comparison of measurements and identification of the underlying burst-generating mechanism. Recent experiments revealed some contradicting results possibly indicating a fundamental difference in regulation of stochastic bursts between yeast and mammalian cells. A study of a synthetic construct implemented in yeast showed that the concentration of the inducer affected the frequency of bursts, i.e. the dynamics of the OFF-to-ON state transition, rather than their size [Raser & O'Shea 2004]. In mammalian cells, on the other hand, it was the burst size that was affected by the concentration of an inducing transcription factor [Raj *et al.* 2006]. Burst size and duration are both properties of the active (ON) state and could relate to such mechanisms as exposing new binding sites due to chromatin remodeling or engaging additional transcriptional regulators localized in the vicinity of the active site [Skupsky *et al.* 2010]. Globally controlled changes in chromatin configuration are more likely to affect burst significance, a quantity related to the frequency of bursts [Skupsky *et al.* 2010]. Therefore, further experiments are required to reveal how much of the stochasticity in gene expression is regulated, which properties of the noise spectrum are modulated and what parts of the regulatory machinery are actively controlling the expression dynamics.

Prokaryotic signaling. In order to quantify the effect of protein abundance, their fluctuations and diffusion on the speed of signaling response we analyzed a generic two-component signaling network in Chapter 3. Other studies have focused on input-output performance of signaling networks at steady-state and have shown that the two-component design is very favorable for dealing with information transfer [Batchelor & Goulian 2003,

Shinar *et al.* 2007]. We are concerned with the temporal response of this network, in particular with the time to initiate gene expression after the extracellular signal appears. The response time is likely optimized during evolution as many environmental stresses have an instantaneous negative influence on cellular physiology. Our theoretical results backed with analysis of bioinformatics data suggest how such an optimization could arise in this remarkably simple system. Distances between the response regulator, the membrane and the DNA site that have to be covered by means of diffusion are physical limitations around which the structure and the composition of the network has to evolve in order to increase its performance.

We find that already a few dozen molecules that constitute the two-component system suffice to achieve a sub-second performance. Scattering receptors on the membrane instead of grouping them in one cluster confers additional two to three-fold reduction. A single response regulator and a single membrane receptor molecule can only achieve this task in about a minute mainly due to the time required to search cognate binding targets on the membrane and on the DNA. This estimate assumes a kinetic scheme without dephosphorylation, free diffusion of proteins and assumes that the first binding to the DNA site of the active response regulator is already successful. Additional time spent on (de-)phoshorylation events, temporal changes in receptor activity (not all receptors remain active during signal relay, which effectively decreases the rate of the regulators' phosphorylation), and the presence of low-affinity binding sites on the DNA (regulators bind to "wrong" sites thus requiring additional attempts to search the relevant promoter), all of these effects will slow down the initial phase of signal response even more. Our theoretical framework uses a single-molecule perspective as a basis for further analysis. Additional biochemical phenomena as those mentioned above can be simply added as transitions that a single molecule undergoes in a sequence. Since many molecules are eventually engaged in signaling response, the time to initiate response is calculated as the first-passage time in an ensemble of molecules (Section A.4).

If the time of the initial stage of the signaling response is indeed under selective pressure (an issue awaiting experimental verification), the results of our study indicate directions this optimization may take. A higher abundance of network constituents evidently decreases the mean and the variance of the response time. However, the benefit of this reduction scales only as the square root of the number of molecules and further synthesis of signaling proteins is likely offset by the cost of their production. For a cell of the size of an *E.coli* bacterium tens of thousands of molecules will not bring about a physiologically significant reduction in the response time compared to a hundred molecules. Scattered receptors do reduce the response time but the change is only few-fold. Receptor clusters as a way to amplify the signal may therefore evolve without hindering the signal relay [Berg & Purcell 1977, Shoup & Szabo 1982, Zwanzig 1990]. The presence of many low-affinity binding sites and many target genes may require a much higher level of response regulators being synthesized than the estimation for the simple model (many response regulators would be arrested at these sites and would slow down the search process). In Figure 3.14 we provide an approximation to account for this effect. Indeed, such a strong bias in favor of a higher response regulator concentration has been observed experimentally (for a UhpA/UhpB network: thousands of response regulators, a few dozen of receptors). This ratio may be further facilitated by the bias in gene order of the two components. We back this claim by the analysis of 623 unique genomes from

MiST database [Ulrich & Zhulin 2007] (Section 3.5.12.2) where we find that in 60% of the cases the gene coding for response regulator is placed before the receptor gene.

Even though the behavior of the generic two-component system that we study will likely be affected by numerous additional effects mentioned earlier, the major findings regarding the effects shaping the design of this network should remain valid. Kinetic parameters, protein abundance and spatial configuration of membrane receptors are currently known for only a handful of systems. Therefore, our numerical estimations may only serve as a diffusion limit for the efficiency of a two-component signal transmission.

How non-exponential waiting times can converge to exponential? All of the problems discussed in this thesis involved processes with times between chemical events described by a waiting time distribution differing from the exponential function. Bursty production analyzed in the Introduction and in Chapter 2 occurred at times drawn from a double-exponential function. Sequential processes are another rich class of problems where the shape of the waiting time distribution differs from the exponential. The time to complete a chemical transition taking place in a series of small steps has a smaller variance than a single-step process. We have seen examples of such a reduction in variance when discussing a sequence of events such as motion of a polymerase (or a ribosome) during transcriptional (or translational) elongation (Chapter 2), a response in a two-component network (Chapter 3), diffusive search for a cognate target (Chapter 3), or threshold activation (Section 1.3.1 of Introduction). Kinetics of a single enzyme have this property too [Qian 2008]. Subsequent steps of the enzymatic transition form a cycle where the time to complete it (the waiting time) is a stochastic variable from a peaked, non-exponential distribution – a fact also confirmed by experiments [English *et al.* 2006] (Fig. 5.3B). Due to lower noise in the time to complete a sequence, it is conceivable that serial chemical transitions gave organisms not only the ability to checkpoint the cell's regulation but also to synchronize biochemical events in time.

Brownian motion of biomolecules and their diffusion-limited searches for cognate targets also belong to that category. A chemical interaction like gene activation, receptor binding or a generic enzyme-substrate binding requires an encounter of two molecular species in cellular space. Typically the distance is covered by means of diffusion (another possibility is active transport which we do not cover here). The shape of the distribution of times required to cover that distance or the distribution of times required to undergo a sequence of chemical transitions (the waiting time or the first-passage time) is a peaked distribution as shown in Fig. 1.13 or Fig. 1.10 in the Introduction. The characteristic shape is a reflection of the simple fact that it is physically impossible for two molecules to meet or to complete the sequence instantaneously, hence the waiting-time distribution assumes zero at the origin.

Non-exponential waiting times may strongly affect steady-state product distributions – a quantity frequently measured in studies of cellular stochasticity by means of experimental technique known as flow cytometry [Blake *et al.* 2003, Newman *et al.* 2006, Friedman *et al.* 2006]. Bursty gene expression may yield a bimodal protein distribution (Fig. 1.7) [Shahrezaei & Swain 2008] inducing cell-cell heterogeneity [Perkins & Swain 2009, Eldar & Elowitz 2010]; narrow, gamma-like waiting time distributions in sequential production and degradation processes will give rise to a narrow steady-state product distribution with noise smaller than the noise of the Poisson dis-

tribution (Fig. 1.3, Introduction). Transients in processes with non-exponential waiting times change too. The time to reach the steady-state in a sequential degradation process discussed in the Introduction differs entirely from the analogous process but modeled as a single step with exponential timing (Fig. 1.11).

Since much of the intuition about biochemical processes in systems biology is based on the assumption that waiting times in chemical reactions are distributed according to the exponential function, it is worthwhile to understand the regimes when non-exponential biochemical processes converge to a much more familiar description. The theoretical framework dealing with exponential processes is well established: a Markov chain can be formed as an abstraction of the stochastic sequence of events. The master equation which describes the temporal evolution of such a system can be solved using various analytical approaches or numerical schemes. Numerous approximations are available too. For instance, moments can be obtained by a variation of the fluctuation-dissipation theorem - the so called linear noise approximation, which assumes a Gaussian distribution of fluctuations around the mean obtained with ordinary differential equations following the law of mass action. The macroscopic limit of the master equation, the law of mass action, is also built on the assumption about the exponential distribution of chemical events in time. Below we shall discuss how bursts, enzymatic reactions and diffusion reactions loose their timing properties and become effectively first-order processes described by exponential waiting times.

Stochastic bursts loose their significance when bursty gene expression takes place from multiple copies of the same gene simultaneously. The double-exponential waiting time in mRNA synthesis converges to an exponential function already for a few gene copies (Fig. 2.9 and Fig. 5.3A). In Section 3.3.5 we discussed the effect of fluctuations in protein concentration on the performance of the signaling network. Infrequent bursty protein synthesis results from slow fluctuations extending cellular generation time [Sigal *et al.* 2006]. This in turn, may impair signal relay even further as some cells may end up with a highly asymmetrical distribution of proteins constituting the network. Positioning genes under control of the same promotor may avoid the effects of such strong fluctuations [Løvdok *et al.* 2009] – variability in concentration will still exist but all protein species will fluctuate in a correlated fashion. The existence of multiple gene copies may significantly reduce the magnitude of fluctuations altogether.

Time precision in single enzyme activity can be also diminished by increasing the copy number of the enzyme (Fig. 5.3B) – the overall waiting time for product release quickly converges to the exponential with the amount of the enzyme rising [Qian 2008]. An ODE description becomes an appropriate level of abstraction; the Michealis-Menten formalism derived from the mass-action law, which intrinsically assumes exponential waiting times, correctly describes the biochemistry of *many* molecules (cf. Eqs. (1.1)–(1.3) in the Introduction).

Exponential approximation in diffusion-limited systems Exponential approximation can be applied to diffusion-limited reactions too. The discrepancy is small if diffusion involves very long trajectories, for instance, when the signaling molecule searches for a small target on the DNA site (Fig. 5.4). The reason behind the exponential character of the distribution becomes clear: intervals between stochastic events without memory of past events are described by the exponential distribution (such a process forms a Markov

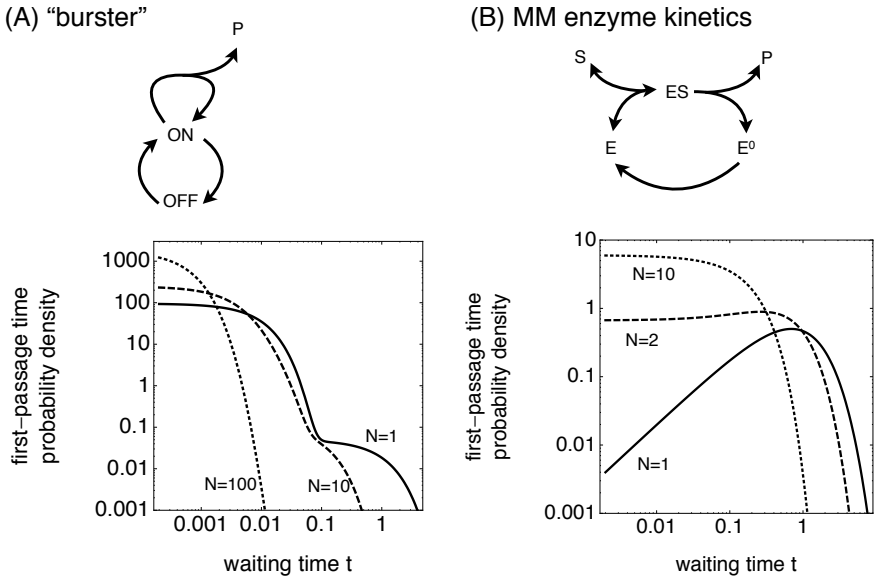


Figure 5.3: In the limit of many independent stochastic processes (N) running in parallel, the waiting time distribution converges to a single exponential function - a macroscopic limit. (A) Waiting time distribution for the synthesis of product P in the generic burst model. Production is very “bursty” - the waiting time distribution has two distinctive regimes (the timescale separation is large). (B) Waiting time distribution for a single enzyme cycle following Michealis-Menten kinetics.

chain). For a diffusive encounter, the trajectory becomes independent of initial conditions (“looses memory”) if the search for the target involves very long trajectories where the searching molecule “wanders off”. In that case, the peak of the waiting time distribution (its typical value) is much smaller than the mean. The whole trajectory becomes a single memoryless step without memory of the initial spatial configuration.

Diffusion-limited reactions are rarely trackable by analytical methods and usually require costly numerical simulations. Calculation of the behavior of the two-component signaling network in Chapter 3 demonstrated how the exponential approximation can be applied to a spatially resolved system without expensive computation: biochemical reactions were modeled as first order processes with exponential waiting times. The existence of such a regime is also an interesting example of how biological systems could evolve the independence of cellular biochemistry to spatial geometry.

The exponential approximation of waiting times in diffusion-limited reactions conserves the mean but, depending on the spatial configuration of the problem, it introduces deviations in higher moments of the waiting time statistics. Deviations in the waiting times affect the steady-state variance as discussed earlier. A coarse-grained theoretical description of a diffusion-limited reaction may introduce additional discrepancies. Computational methods based on various physical models of a diffusive encounter may yield different waiting time distributions. A study in Chapter 4 demonstrated this issue. The

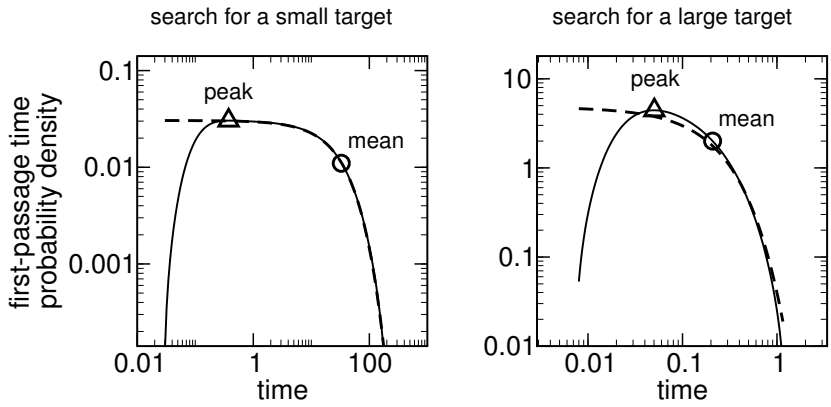


Figure 5.4: Waiting time distributions for small and large target searches. Solid line is the waiting time distribution for a diffusive search process. Dashed curve indicates the exponential approximation with the same mean as the original search process. The search for a small target has the characteristics of the memoryless process. The waiting time distribution is peaked but wide (the mean is much larger than the peak). The exponential approximation of the search for a large target is poor; the distribution is very peaked.

magnitude of steady-state fluctuations computed for a generic gene expression model depended on the underlying model and the level of spatial discretization (Fig. 4.5). Inconsistency in results was a direct consequence of the inability to reproduce a correct waiting time statistics of a reversible bimolecular reaction (Fig. 4.6).

The extent to which the waiting time diverges from the exponential function and the influence of this departure on macroscopic properties of the cell is still largely unrecognized in the field of molecular biology. The topics discussed here will become even more relevant in the light of recent advances in experimental techniques, which make possible the observation of the evolution of single enzyme molecules. Understanding of these phenomena may come only from the type of analysis presented earlier in this thesis. Experiments tackling the physiology of a single cell shall further improve our understanding of timing properties of biochemical reactions and verify theoretical hypotheses, some of which have been presented in this thesis. Recognizing the exponential regime shall be also of great interest to the modeling community as the complexity of the problem can be greatly reduced if stochastic events are assumed to be exponentially distributed in time.

One common hurdle in the quantitative study of living organisms is the immense complexity of biochemical processes. The amount of involved biomolecules and vast interaction networks they engage into, results in phenomena spanning several temporal and spatial scales. A tiny change in the DNA sequence may change the organism's phenotype, promotea a novel anatomical structure by rearranging connectivity between several core processes, modify the social behavior or impair the functionality to the point the organism is no longer viable. The understanding of how complex gene networks define the phenotype still falls short from being complete and will likely occupy systems biologists in the foreseeable future.

Despite the difficulty in analyzing theoretically the multi-scale nature of biochemical

processes and collective phenomena emerging in complex biological networks, the laws of physics behind them are no different from those governing the flight of a projectile or the flow of electric current in a wire. Even though a consistent analytical framework dealing with complex biological systems is still likely to materialize, it is conceivable that no new form of physics will be required in such a theory. Physics as we know it today already provides countless intuitive insights into the way functionality of the living organism is shaped in the course of the evolution. *Restriction* and *facilitation*, are two forces to which organisms owe their design.

The former one is the fairly obvious observation that the building blocks of any living system cannot escape the limitations of the physical world they evolved in. For instance, the process that relays information about the environment to the inside of the cell will be subject to thermal noise affecting a sensory module. The time to process and adjust cellular physiology to that information is also limited by rates of chemical reactions - an outcome of the speed of molecular motion in the intracellular space, the relative position of enzymatic active sites, the effective forces exerted by neighboring molecules and cellular structures. Similarly, anatomical assimilations must conform to the laws of physics: a wing of a bird, a fin of a fish, etc. Numerous mechanisms have evolved in order to push those limits and to maximally optimize the structure such that the reproductive success of the organism increases.

Aside from restricting the structure of the living system, physical interactions facilitate a number of its functionalities. It is that "physical scaffold" that allows a relatively small number of human genes (a little more than 20000) to direct the matter to form several hundred cell types, coordinate them to produce more than 100 trillion cells organized into dozens of organs, one of them the brain: a staggering collection of 100 billion of neurons linked with a million billion of synapses. Due to information capacity constraints neither the exact position of every cell in the human body nor the information about its shape or the precise molecular content is given explicitly in the gene sequence. It is not even necessary. An exploratory process, a purely physical phenomenon, facilitates formation of the vascular and the nervous system, or growth of microtubules which form and maintain the cytoskeleton - the scaffolding of every eukaryotic (and according to recent findings, possibly also prokaryotic) cell. The information contained in the genome merely defines the building blocks, while the physics takes over the assembly process giving the final organization of the organism growing from the embryo, the three-dimensional fold of the protein or the messenger RNA, the lipid bilayer structure of the cell membrane. The much discussed utilization of randomness in gene expression to induce heterogeneity in the microbial population that increases survival rate in fluctuating environment belongs to the same category of processes: the genome merely triggers the usage of a physical phenomenon but does not prescribe the exact outcome. Hence, a lot of information about the final anatomy of the organism or even structure of the population is written in the physical interactions itself. The understanding of these relationships and mimicking them by theoretical models is the holy grail of systems biology.

Appendix A

First-passage time basics

A.1 *pdf* and CDF

This chapter introduces the probability density function (*pdf*), $f(t)$, and a function related to it, the cumulative distribution function (CDF), $F(t)$. In great simplification, the *pdf* describes the following: the probability that a random variable (RV) X falls within an infinitesimal interval $(t, t + \Delta t)$ equals:

$$Pr[X \in (t, t + \Delta t)] = f_X(t)\Delta t \tag{A.1}$$

Throughout the thesis we analyze first-passage times of various biological phenomena. Variable t is a measure of time elapsed till the commencement of the first event. Such an event could be, for instance, reaching a certain area in a 3D space by one or more diffusing molecules (Chapter 3), or the time between two consecutive syntheses of a molecule (Chapter 2).

The area under the $f(t)$ curve is normalized to 1, and in most cases its value at a particular t exceeds 1. The *pdf* is not the probability of an event to occur *per se*. It can be viewed as a continuous version of the histogram. Such a histogram shows frequencies of values taken by the RV X.

A probability distribution has a density function if there exists a continuous CDF:

$$F_X(t) = Pr[X \leq t] = \int_0^t f_X(\tau)d\tau \tag{A.2}$$

The CDF gives the probability that a RV X takes on a value less than or equal to t. Its derivative is the *pdf*:

$$\frac{d}{dt}F_X(t) = f_X(t) \tag{A.3}$$

It is also helpful to introduce the survival probability (SP) $S_X(t)$:

$$S_X(t) = Pr[X > t] = 1 - F_X(t) \tag{A.4}$$

$$f_X(t) = \frac{d}{dt}F_X(t) = -\frac{d}{dt}S_X(t) \tag{A.5}$$

The term *survival* can be easily understood by analyzing the example of binding macromolecules. Take a single molecule that starts diffusing at time t_0. Then, the random variable X denotes the time elapsed before the molecule finds a target, an area on a domain boundary, for instance. In that sense, as long as the molecule is engaged in diffusion it *survives*; that is, the molecule has not reached the target within time t. Similarly, the CDF $F_X(t)$ is often termed the *mortality* function in the literature, as it gives the probability that X is smaller than t, i.e. the molecule finds the target before time t.

A.2 Moments of the first-passage

A number of properties of the first-passage time can be inferred from $\tilde{f}(s)$ even without performing the inverse transform. The behavior of the Laplace transform for $s \to 0$ determines the long-time behavior of the function $f(t)$. Using the definition of the Laplace transform, the small-s expansion generates all positive moments of $f(t)$ [Redner 2001]:

$$\begin{aligned}\tilde{f}(s) &= \int_0^\infty f(t)e^{-st}dt \\ &= \int_0^\infty f(t)\left(1 - st + \frac{s^2t^2}{2!} - \dots\right)dt \\ &= F(\infty)\left(1 - s\langle t\rangle + \frac{s^2}{2!}\langle t^2\rangle - \dots\right) \end{aligned} \tag{A.6}$$

where $F(\infty) \equiv \tilde{f}(s=0)$ is the probability of *eventually* arriving in the state of interest. In case of diffusing molecules, the long-time limit of the cumulant $F(t)$ gives the probability of reaching the absorbing sphere. For a reaction network, this limit refers to the total probability of reaching a discrete molecular state. If the system always reaches this state for any given initial condition, then $F(\infty) = 1$.

A.3 Moments and the survival probability

Alternatively, one can compute the n-th moment directly from the interarrival time probability density $f_X(t)$:

$$\langle t^n\rangle = \int_0^\infty t^n f_X(t)\,dt \tag{A.7}$$

Simple algebra allows to obtain the relation between the survival probability and the mean. Using definition A.4 and integrating Eq. A.7 for the moments by parts yields:

$$\begin{aligned}\langle t^n\rangle &= -\int_0^\infty t^n \frac{d}{dt}S_X(t)\,dt \\ &= -t^n S_X(t)\Big|_0^\infty + n\int_0^\infty t^{n-1}S_X(t)\,dt \end{aligned} \tag{A.8}$$

The first term equals zero at the lower limit by definition. The upper limit can be inferred using the fact, given in Section A.2, that the eventual arrival probability equals 1. This in turn implies that none of the intervals can *survive* the infinite duration. Hence, the SP must approach zero for $t \to \infty$ and the integrated term at the upper limit must be zero.

For $n = 1$, we obtain the mean first-passage time (MFPT):

$$\langle t\rangle = \int_0^\infty S_X(t)dt \tag{A.9}$$

A.4 Superposition of N independent random processes

In this section we derive a general expression for the first-passage time *pdf* for the superposition of N independent and identically distributed (*iid*) random variables (RVs). We

shall begin with the first order statistics. This result allows to obtain the first-passage time *pdf* in an ensemble of N independent diffusing molecules (i.e. the *pdf* for the first molecule that arrives at the target [Weiss *et al.* 1983]) or N independent bursting mechanisms (i.e. for the first molecule synthesized; Chapter 2). The assumption behind this derivation is that all independent processes are initially *synchronized*. In the case of diffusing molecules, this translates to all molecules being in the same position (although the initial position could also be *random*) and *not* at the target. For an ensemble of bursters, synchrony of the initial state implies the same state, ON or OFF, for all bursters. The k-th order statistics provides analogous results for the k-th molecule arrival or the k-th synthetic event.

Further on, we define the excess lifetime and we derive the first-passage time in an ensemble of *unsynchronized* random processes. This result is applicable for a computation of the distribution of interarrival times due to synthesis of molecules by an ensemble of bursters, each being in a random state (Section 2.3.5). We also use this equation to illustrate waiting time distributions for product synthesis in a classical enzymatic reaction (Fig. 5.3).

A.4.1 First order statistics

Let $Y^1, Y^2, \ldots, Y^N$ be a sample of independent and identically distributed (*iid*) continuous RVs. A RV Y^i has a *pdf* $f_Y(t)$, and a CDF $F_Y(t)$ (for brevity we skip the subscript Y in this section). We define $Y^{(1,N)} = min\left\{Y^1, Y^2, \ldots, Y^N\right\}$ as the first order statistics of the *iid* sample of size N. We are interested in the *pdf* and CDF of the first order statistics.

First, we note that:

$$f^{(1,N)}(t) = \frac{d}{dt} F^{(1,N)}(t) \tag{A.10}$$

The CDF of the first order statistics, $F^{(1,N)}(t)$, can be easily related to the CDF of a single RV, $F(t)$:

$$\begin{aligned} F^{(1,N)}(t) &= 1 - Pr[Y^{(1,N)} > t] \\ &= 1 - Pr[Y^1 > t \text{ and } Y^2 > t, \ldots, \text{ and } Y^N > t] \\ &= 1 - (1 - F^{(1)}(t)) \cdot (1 - F^{(2)}(t)) \ldots (1 - F^{(N)}(t)) \\ &= 1 - (1 - F(t))^N \end{aligned} \tag{A.11}$$

The third line results from the fact that all RVs are *iid*.

The survival probability $S(t) = 1 - F(t)$ (cf. Section A.1) for the first order statistics reads:

$$S^{(1,N)}(t) = S(t)^N \tag{A.12}$$

Finally, upon substituting the above into Eq. A.10 we obtain the *pdf* for the first passage time out of N *iid* processes:

$$\begin{aligned} f^{(1,N)}(t) &= N\, f(t)\, (1 - F\,(t))^{N-1} \\ &= N\, f(t) \left(1 - \int_0^t f\,(\tau)\, d\tau\right)^{N-1} \end{aligned} \tag{A.13}$$

A.4.2 k-th order statistics

The result of the previous section can also be achieved by a slightly different approach. Additionally, one can obtain a k-th order statistics, i.e. a k-th smallest passage time $f^{(k,N)}(t)$. A combinatorial term appears:

$$\begin{aligned} f^{(k,N)}(t)\,dt &= Pr(Y^{(k,N)} \in (t,\, t+dt)) \\ &= Pr(\text{one of the } Y^i \in (t,\, t+dt),\ k-1 \text{ of all the others } \leq t) \\ &= N\, Pr(Y^1 \in (t,\, t+dt),\ (k-1) \text{ others } \leq t) \\ &= N\, Pr(Y^1 \in (t,\, t+dt)) \binom{N-1}{k-1} F(t)^{k-1}(1-F(t))^{N-k} \\ &= Nf(t)dt \binom{N-1}{k-1} F(t)^{k-1}(1-F(t))^{N-k} \end{aligned}$$

and finally:

$$f^{(k,N)}(t) = Nf(t) \binom{N-1}{k-1} F(t)^{k-1}(1-F(t))^{N-k} \tag{A.14}$$

A.4.3 Excess lifetime

We set out to obtain a distribution of interarrival times in the pooled (superposed) process, where the states of the independent sources are not synchronized. Therefore, the result of the previous two sections cannot be applied directly here. The Equations A.13 or A.14 would be relevant if one was interested in the *pdf* of the first arrival out of N, given that all sources were in the same initial state at T_0. However, we are looking for the interarrival time at an arbitrary epoch T from the last arrival at T_0. In this case, all other processes, except for the one which fired at T_0, are in some other state which is not necessarily the very moment of firing. Therefore, we introduce an auxiliary RV, which describes the interval between an arbitrary epoch and the event directly preceding *regardless* of the source. A RV with these properties is known in the literature as the *excess lifetime*. We shall define its *pdf* for a single source. This in turn, will allows us to use it in Eq. A.13 for the first order statistics. Although, in our case arrivals are generated by a single source – an interrupted Poisson process (IPP) as in Section 2.5.1, the results presented here are not restricted to any particular distribution function.

We consider a stochastic process where the interarrival times between successive events are *iid* random variables $X = \{X_k,\, k = 1, 2, \dots\} \equiv Y^i$. A process with this property and an arbitrary distribution of the interarrival times is a renewal process (cf. Chapter 1.7 in [Medhi 2003]). RV X is described by a continuous probability density function $f_X(t)$ and the cumulated distribution function $F_X(t)$. For instance, in case of a single-source IPP, the *pdf* and CDF are given by Eqs. 2.9.

Since X denotes the length of the interarrival time of duration t, the epoch of the k-th arrival is:

$$T_k = X_1 + X_2 + \cdots + X_k, \quad k \geq 1, \quad \text{and } T_0 = 0 \tag{A.15}$$

The time of arrival that occurs immediately before an arbitrary epoch T is:

$$T_{K(T)}, \quad \text{where } K(T) = \sup\{k : T_k \leq T\} \tag{A.16}$$

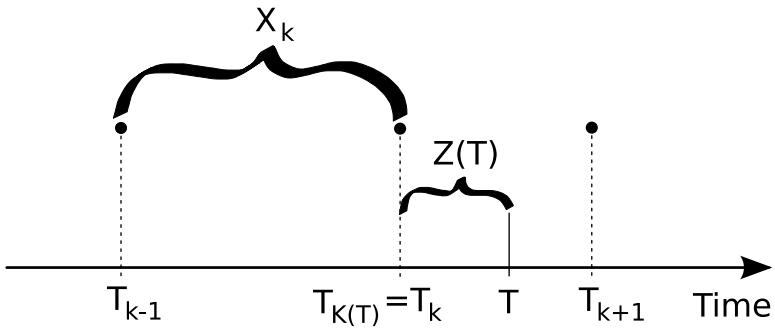


Figure A.1: Random variables for the interarrival time (X) and the excess lifetime (Z) shown for three consecutive random arrivals. The equilibrium *pdf* for RV Z is related to the probability distribution of interarrival times X through relation A.21.

Therefore, if an arbitrary epoch of time T lies between the k-th and the $(k+1)$-th arrival, then the time t between arrivals fulfills:

$$T_{K(T)-1} < t \leq T_{K(T)} \tag{A.17}$$

Further, we define an additional RV, $Z(T) = T - T_{K(T)}$, which is the length of the interval between an arbitrary point at time T and the time of occurrence of the immediately preceding event $T_{K(T)}$ (Fig. A.1). RV $Z(T)$ is known in literature as the *excess* or *spent lifetime*, or *backward-recurrence time* at T [Medhi 2003], or simply the *delay* [Cox & Smith 2006].

The *excess lifetime* function at T is the *pdf* g associated with RV $Z(T)$:

$$g_Z(t,T) = Pr[Z(T) \in (t, t+\Delta t)] \tag{A.18}$$

Note that the value of RV $Z(T)$ is also denoted by the letter t, which is a measure of the interval length. It is safe to do so, as long as the function is properly described by the name of a corresponding RV, in this case subscript Z.

It can be written as the total density $h(T_{K(T)})$ of events at time $T_{K(T)} = T - t$, times the probability that the interarrival time $X_k = T_{K(t)+1} - T_{K(t)}$ exceeds t:

$$g_Z(t,T) = h(T-t) \cdot Pr[X > t] = h(T-t) S_X(t) \tag{A.19}$$

where $S_X(t)$ is the SP for the interarrival time frequency.

In the long-time limit, the total density of events equals simply the expectation value of RV X:

$$\lim_{T\to\infty} h(T-t) = \frac{1}{E(X_k)} = \frac{1}{\langle t \rangle} \tag{A.20}$$

and hence:

$$g_Z(t) = \lim_{T\to\infty} g_Z(t,T) = \frac{S_X(t)}{\langle t \rangle} \tag{A.21}$$

Thereby, we obtained a very useful relation between the excess lifetime function, $g_Z(t)$, and the interarrival time SP. In the next section, we will use it to construct the excess

lifetime *pdf* for the superposed process, and to obtain the interarrival time CDF for N sources.

Similarly, the CDF associated with the equilibrium delay variable Z is given by:

$$G_Z(t) = Pr[Z \leq t] = \frac{1}{\langle t \rangle} \int_0^t S_X(\tau) d\tau, \quad \tau \geq 0 \tag{A.22}$$

Note, that the *pdf* of the excess lifetime for the Poisson process is the same as the *pdf* for interarrival times. This is the only renewal process with this property [Medhi 1994].

A.4.4 Interarrival time CDF in an unsynchronized ensemble

Here, we make use of Eqs. A.11 and A.13 for the first order statistics to express the CDF in the pooled process (superscript S instead of $(1, N)$) in terms of a single source probability distribution. As mentioned earlier, in order to tackle events generated by an ensemble of *unsynchronized* processes we cannot use the interarrival time CDF of the single source directly in these equations. Instead, we substitute the excess lifetime CDF of a single source given by Eq. A.22:

$$\begin{aligned} G_Z^{(S)}(t) &= 1 - (1 - G_Z(t))^N & \text{(A.23)} \\ g_Z^{(S)}(t) &= \frac{d}{dt} G_Z^{(S)}(t) \\ &= N\, g(t)\, (1 - G_Z(t))^{N-1} & \text{(A.24)} \end{aligned}$$

Further, we substitute the excess lifetime *pdf* and CDF in the above relation with Eqs. A.21 and A.22:

$$\frac{S_X^{(1,N)}(t)}{\langle t \rangle^{(S)}} = \frac{N}{\langle t \rangle} S_X(t) \left(1 - \frac{1}{\langle t \rangle} \int_0^t S_X(\tau) d\tau\right)^{N-1}, \tag{A.25}$$

where $\langle t \rangle^{(S)}$ is the mean interarrival time in the pooled process. For instance, for the superposition of *iid* IPP processes this mean is related to the single source mean (Eq. 2.12a) in the following manner:

$$\langle t \rangle^{(S)} = \frac{\langle t \rangle}{N} \tag{A.26}$$

Finally, the SP in a pooled IPP process can be written in terms of a single-source SP $S_X(t)$:

$$S_X^{(S)}(t) = S_X(t) \left(1 - \frac{1}{\langle t \rangle} \int_0^t S_X(\tau) d\tau\right)^{N-1} \tag{A.27}$$

Appendix B

Basics of diffusion-limited reactions

B.1 Theory: single molecule

First-passage time theory is a useful theoretical framework to analyze diffusion of macromolecules in geometries like the problem analyzed in Chapter 3: a 3D sphere with (partially) absorbing boundaries [Redner 2001, Weiss 1967].

Diffusive movement of a single macromolecule in a potential free domain can be described by the following Fokker-Planck equation, also referred to as the Einstein diffusion equation [Schulten & Kosztin 2000, Weiss 1994, Weiss 1967, Ebeling *et al.* 2003]:

$$\frac{\partial c\left(\vec{r}, t|\vec{r_0}, t_0\right)}{\partial t} = D\nabla^2 c\left(\vec{r}, t|\vec{r_0}, t_0\right). \tag{B.1}$$

The continuous variable c is the probability density for finding a particle at position $\vec{r}$ at time t given it was at $\vec{r}_0$ at the earlier time t_0. Designation of the function by a letter c is intentional because it is analogous to the macroscopic continuous concentration of diffusing molecules.

In this case, the diffusion coefficient D is assumed constant in the whole domain. Equation B.1 implies that the molecule diffuses freely, that is, the mean square displacement is linearly proportional to time:

$$\left\langle \left(\vec{r}(t) - \vec{r}(t_0)\right)^2 \right\rangle = 2^d Dt, \tag{B.2}$$

where d denotes the dimensionality.

B.1.1 Incorporating reactions through boundary conditions

In order to include chemical reactions into the model we need to prescribe appropriate boundary conditions for the volume Ω. The list below summarizes four different types of such conditions corresponding to various interactions.

Reflecting: the particle does not cross the boundary:

$$\hat{\mathbf{n}}(\mathbf{r}) \cdot \hat{\mathbf{J}}(\mathbf{r})\, c(\mathbf{r}, t|\mathbf{r}_0, t_0) = 0, \qquad \mathbf{r} \in \partial\Omega, \tag{B.3}$$

where $\hat{\mathbf{n}}(\mathbf{r})$ is a unit vector normal to $\partial\Omega$, the boundary of the volume Ω. The term $\hat{\mathbf{J}}(\mathbf{r}) = D\nabla$ is the flux operator acting on the solution of the Einstein diffusion equation. It yields the local probability flux.

Reaction (absorbing): the particle arriving at the boundary is annihilated with 100% efficiency:

$$c(\mathbf{r}, t|\mathbf{r}_0, t_0) = 0, \qquad \mathbf{r} \in \partial\Omega. \tag{B.4}$$

Radiation: intermediate reactivity at the boundary:

$$\hat{\mathbf{n}}(\mathbf{r}) \cdot \hat{\mathbf{J}}(\mathbf{r})\, c(\mathbf{r}, t|\mathbf{r}_0, t_0) = \kappa_r\, c(\mathbf{r}, t|\mathbf{r}_0, t_0), \qquad \mathbf{r} \in \partial\Omega, \tag{B.5}$$

where κ_r is intrinsic reaction rate at contact [Agmon & Szabo 1990].

Reversible (back-reaction): after reaction the particle can undergo dissociation with intrinsic rate coefficient κ_d [Agmon & Szabo 1990]:

$$\hat{\mathbf{n}}(\mathbf{r}) \cdot \hat{\mathbf{J}}(\mathbf{r})\, c(\mathbf{r}, t|\mathbf{r}_0, t_0) = \kappa_r\, c(\mathbf{r}, t|\mathbf{r}_0, t_0) - \kappa_d \left[1 - Sep(t|\mathbf{r}_0, t_0)\right], \qquad \mathbf{r} \in \partial\Omega, \tag{B.6}$$

where $Sep(t|\mathbf{r}_0, t_0)$ is the *separation* probability, i.e. the probability that the molecule is unbound [Agmon & Szabo 1990].

B.1.2 Solution for the spherically symmetric system

Since we are considering a strongly diffusion-limited binding, we assume that the time to overcome the free energy barrier is negligible compared to the time required for the two molecules to meet. Hence, we will use the **reaction** boundary condition B.4 to describe the binding.

Once we assume radial symmetry of the problem, $c(\vec{r}, t) \to c(r, t)$, where $r = |\vec{r}|$. Then, Eq. B.1 can be rewritten in the following way:

$$\frac{\partial c(r,t)}{\partial t} = D\left[\frac{\partial^2 c(r,t)}{\partial r^2} + \frac{d-1}{r}\frac{\partial c(r,t)}{\partial r}\right], \tag{B.7}$$

where d is the dimensionality of the system.

We solve this equation with the absorbing boundary condition at the inner concentric sphere $r = R_-$, and reflecting at the outer boundary $r = R_+$:

$$c(R_-, t) = 0, \tag{B.8}$$

$$\left.\frac{\partial c(r,t)}{\partial r}\right|_{r=R_+} = 0. \tag{B.9}$$

We assume the initial condition to be a spherical shell of probability density (concentration) at $r = r_0$:

$$c(r, t=0) = \delta(r - r_0)\frac{1}{\Omega_d\, r_0^{d-2}}, \tag{B.10}$$

where $\delta(r)$ is the Dirac delta function, Ω_d is the surface area of a d-dimensional unit sphere for normalization.

The solution is obtained by means of the Laplace transform in the time domain [Redner 2001]:

$$c(x, s) = \frac{1}{s}\left(\frac{s}{D}\right)^{d/s} \frac{(xx_0)^\nu}{\Omega_d} \frac{\mathcal{C}_\nu(x_<, x_-)\, \mathcal{D}_{\nu,-}(x_>, x_+)}{\mathcal{D}_{\nu,-}(x_-, x_+)} \tag{B.11}$$

where:

- d - dimensionality, and $\nu = 1 - d/2$ - dimensionality factor,
- $x = r\sqrt{s/D}$, radius reexpressed in terms of the dimensionless radial coordinate; similarly, $x_\pm = R_\pm\sqrt{s/D}$,
- $x_> = max(x, x_0)$, $x_< = min(x, x_0)$,
- $\mathcal{C}_\nu(a,b) \equiv I_\nu(a)K_\nu(b) - I_\nu(b)K_\nu(a)$, where I_ν and K_ν are modified Bessel functions of the first and second kind, respectively,
- $\mathcal{D}_{\nu,\pm}(a,b) \equiv I_\nu(a)K_{\nu\pm1}(b) - K_\nu(a)I_{\nu\pm1}(b)$.

A related problem, where the boundary conditions are switched, that is reflective at R_- and absorbing at R_+, involves only interchanging variables x_- and x_+.

B.1.3 First-passage time

We denote the single-particle first-passage time (FPT) *pdf* as $f(\vec{r}_f, \vec{r}_0, t)$. It is the probability density for a single particle, initially placed at position r_0 at time $t = 0$, to reach the target at r_f at time t for the first time. The FPT is related to the occupation probability $c(\vec{r}_f, \vec{r}_0, t)$ which is the probability to be at r_f at time t, irrespective of when it arrived:

$$\begin{aligned} c(\vec{r}_f, \vec{r}_0, t) &= \left(f(\vec{r}_f, \vec{r}_0) * c(\vec{r}_f, \vec{r}_f)\right)(t) \\ &= \int_0^t f(\vec{r}_f, \vec{r}_0, \tau)\, c(\vec{r}_f, \vec{r}_f, t-\tau) d\tau. \end{aligned} \tag{B.12}$$

This convolution can be explained in the following way. The occupation probability at the final position r_f at time t has two contributions. The first is the FPT *pdf* to visit the final position r_f at the earlier time τ for the first time, $f(\vec{r}_f, \vec{r}_0, \tau)$. The latter is the occupation *pdf* for starting at r_f and returning to the same position within time τ', $c(\vec{r}_f, \vec{r}_f, \tau')$. Finally, we need to sum all the terms $f(\tau)c(\tau')$, where $\tau + \tau' = t$, because we are looking for the occupation at the final time t. Hence the convolution $f(\tau) * c(\tau')$ in the above equation.[1]

The relation B.12 is very useful if rewritten in the Laplace form. It allows to find the FPT *pdf* from the occupation probability in the following way:

$$f(\vec{r}_f, \vec{r}_i, \tau) = \mathcal{L}^{-1}\left\{\frac{\mathcal{L}\left\{c(\vec{r}_f, \vec{r}_i, t)\right\}}{\mathcal{L}\left\{c(\vec{r}_f, \vec{r}_f, t-\tau)\right\}}\right\}. \tag{B.13}$$

B.1.4 First-passage related to flux

The first-passage probability density of a diffusing particle to a certain area, e.g. an absorbing boundary, is the **total current** (flux) of the density $c(\vec{r}, t)$ to that boundary (provided the initial condition is normalized). Equation B.11 gives the Laplace transform of the time-dependent occupation probability density in a radially symmetric domain with

[1] In probability theory the sum of two independent random variables is the convolution of their individual probability densities.

absorbing boundary at R_- and reflective at R_+. We are interested in the first-passage time to the absorbing boundary, hence we compute the flux at R_-:

$$\begin{aligned} J_-(x_0,s) &\equiv \tilde{f}(x_0,s) \\ &= \int_{S_-} D\,\frac{\partial c}{\partial r}\bigg|_{r=R_-} \\ &= \left(\frac{x_0}{x_-}\right)^{\nu} \frac{\mathcal{D}_{\nu,-}(x_0,x_+)}{\mathcal{D}_{\nu,-}(x_-,x_+)} \end{aligned} \tag{B.14}$$

Note that $J_-(x_0,s)$ in the above equation still depends on s through $x_\pm$ and not on the real time t. We are still in the Laplace frequency domain. Dependency on x_0 denotes the initial position expressed in terms of the dimensionless radial coordinate.

The relation of the FPT to the flux is more general and applies also to the FPT of chemical events occurring in reaction networks. For example, if one wishes to compute the FPT to a certain discrete state in a reacting system described by the chemical master equation, the FPT *pdf* consists of all possible transitions that end up in that particular system.

B.1.5 Evaluation for 1D

We shall consider a 1D domain of length $R_+ \equiv L$. In order to obtain the first moment, or the mean, of the first-passage probability in 1D we evaluate expression B.14 for $d = 1$, expand to the lowest order around small s (equivalent to taking $t \to \infty$), and compare coefficients to Eq. A.6. Setting the initial position to $x_0 = L$ (reflective end), and then assuming $R_- = 0$ we obtain the expression for the mean first-passage time to an absorbing end at $x = 0$ of a particle diffusing on a 1D line of length L:

$$\langle t \rangle_L = \frac{L^2}{2D}. \tag{B.15}$$

One can recognize in the above expression the average time needed for a diffusing molecule to traverse a distance L.

The Laplace transform of the FPT *pdf* takes the form:

$$\tilde{f}(x_0 = L, s) = sech\left(L\sqrt{s/D}\right). \tag{B.16}$$

In this case it is possible to invert this relation and to obtain the explicit expression for the time-dependent FPT *pdf* [Nagar & Pradhan 2003]:

$$f(t) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{(2n-1)\pi D}{L^2} e^{-(n-\frac{1}{2})^2 \pi^2 D t / L^2}. \tag{B.17}$$

Similarly, when a particle initiates its motion randomly in the domain, the FPT *pdf* in the Laplace form equals:

$$\tilde{f}(s) = \frac{1}{L}\int_0^L \tilde{f}(x_0,s)dx_0 = \frac{1}{\sqrt{s/D}} tanh\left(L\sqrt{s/D}\right). \tag{B.18}$$

The MFPT for random initial conditions is:

$$\langle t \rangle_{rnd} = \frac{L^2}{3D} = \frac{2}{3}\langle t \rangle_L. \tag{B.19}$$

Appendix C

Mesoscopic models and computational methods

C.1 Mesoscopic models

Here we present various theoretical approaches to the modeling of biochemical reactions on the mesoscopic level. All models are stochastic with discrete molecule detail. Only Brownian dynamics and the reaction-diffusion master equation account for space.

Brownian dynamics (BD) is a computational model which proved to be very useful in modeling biochemical systems. It has been applied, to name only a few studies, to the enzyme-ligand binding or to determine kinetic parameters [Northrup *et al.* 1984, Tan *et al.* 1993]. In this approach the solvent is treated as a continuous medium while solute molecules are modeled explicitly in space. Their trajectory is described by a random walk due to collisions with the much smaller solvent atoms. Additionally the interaction potential, for example due to electrostatic forces, can be included. This dramatically reduces the order of complexity of the model since the majority of degrees of freedom is characterized only by the fluctuating force. As a result the computational cost is much smaller than that of methods based on molecular dynamics (MD) where the positions and velocities of all atoms are traced. However, even BD is very expensive when applied to a large biochemical system because of the relatively small simulation time step needed to resolve a collision event leading to a chemical reaction.

Brownian dynamics with chemical reactions. A spherical molecule undergoing a random walk close to a target to which it can bind reversibly can be modeled by the Smoluchowski diffusion equation (Eq. B.1). This is a PDE for the probability distribution of finding a particle at a certain point in space with boundary conditions dependent on the intrinsic association and dissociation rates (Eqs. B.3-B.6) [Rice 1985, Agmon & Szabo 1990, Weiss 1994]. This mathematical abstraction corresponds to a pair of molecules diffusing and reacting in a closed or infinite volume. Far from the target the solution is very well represented by a Gaussian distribution (free diffusion) since the influence of the reactive trap is negligible. For short distances the inter-particle distribution is modified significantly due to the possibility of binding and geminate recombination. For that reason it is necessary to use smaller time steps in computer simulations, since for a given distance the influence of the trap on the diffusing particle is decreased on a shorter time scale. In order to maximize the simulation time step an analytical solution of the diffusion equation for a pair of particles can be employed [Lamm & Schulten 1981]. One of the quantities obtained in such an operation is

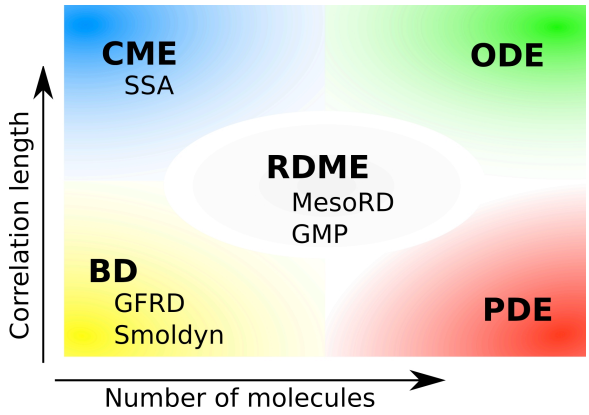


Figure C.1: Mesoscopic models and computational methods for solving biological problems in different regimes. Abbreviations: BD - Brownian dynamics, CME - chemical master equation, ODE - ordinary differential equation, PDE - partial differential equation, RDME - reaction-diffusion master equation. For method abbreviation see text.

the probability of binding after a certain time step given the initial position at the earlier time. An example analytical solution for a 1-D domain and an absorbing boundary condition is given in Section B.1.5. In this way diffusion and reaction steps are combined together which, in principle, allows to propagate the system with arbitrarily large time steps. The analytical solution (a Green's function or propagator) has been employed in an efficient BD simulation method in 1-D [Edelstein & Agmon 1993], further extended to 3-D [Kim & Shin 1999], and used extensively as a reference result for testing theories of diffusion-limited reactions [Popov & Agmon 2001a, Popov & Agmon 2001b]. The above ideas are also implemented in two of the methods for modeling biochemical networks, GFRD and Smoldyn, outlined in Section C.2. They have been respectively applied to study fluctuations in gene expression [van Zon & ten Wolde 2005a, van Zon *et al.* 2006], and to model signal transduction in *E. coli* chemotaxis [Lipkow *et al.* 2005].

Chemical master equation. In the commonly used ODE approach to modeling biochemical networks concentrations of chemical species are treated as continuous variables. Differential equations describe deterministic propagation of these variables in time. In this approach, the discrete nature of molecules is discarded, which is valid if concentrations are *high*. For large numbers of molecules the continuum assumption is an adequate model, however for low concentrations the inherent randomness of the reactants' collisions requires a stochastic description. The chemical master equation (CME) is such a formulation. It describes the evolution of the probability distribution of finding the system in a certain discrete state given by a vector with integer-valued populations of molecular species reacting in the system (see [van Kampen 1997, Gillespie 1992] for a derivation). We shall exemplify the structure of the CME using an example of the synthesis-degradation reaction scheme depicted in Fig. 1.2 in the Introduction.

In our example, molecules of species X are synthesized at the rate k_p, and degraded at the rate $k_d \cdot x$, where x denotes the number of molecules. The stochastic description

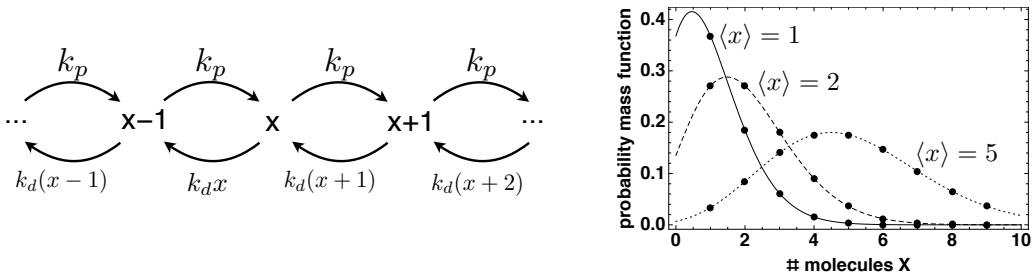


Figure C.2: (Left) Discrete states in the production-degradation model, $\varnothing \to X \to \varnothing$. Arrows depict chemical reactions that transfer the system to a new state. The chemical master equation (Eq. C.1) describes temporal changes of the probability distribution for a single state x. (Right) The steady-state solution, the Poisson distribution, shown for three different steady-state averages, $\langle x(t) \rangle_t = k_p/k_d = 1$, 2, and 5.

accounts for the discreteness of molecules and fluctuations in the synthesis and degradation rates; $1/k_p$ has the interpretation of the mean time between two production events and $1/k_d$ is the mean lifetime of a single molecule X. Both intervals fluctuate and the function describing their distribution is an exponential. The extensive discussion of the effects of other waiting time probability distributions on the steady-state distribution of X can be found in the Introduction (Chapter 1).

The quantity evolved in the CME is the probability distribution $P(x,t)$, which is the probability of finding the system in a state with x molecules at time t. Gains and losses of a single state due to chemical reactions are depicted in Fig. C.2; mathematically they are formalized in the following way:

$$\begin{aligned} \frac{d}{dt}P(x,t) &= k_p\,P(x-1,t) + k_d\,(x+1)\,P(x+1,t) \\ &- P(x,t)\,(k_p + k_d\,x)\;. \end{aligned} \tag{C.1}$$

The first line of the right hand side includes "gains", that is all chemical reactions that result in state x. These are the synthesis of a *single* X molecule in state $x-1$ and the degradation of a *single* X molecule in state $x+1$. Similarly, the second line describes, respectively, the synthesis and the degradation event that lead to abandonment of the state x – the "losses". Each state with x molecules is described by such an equation. The steady-state solution, i.e. $dP(x,t)/dt = 0$, leads to a familiar Poisson distribution (Fig. C.2). The CME approach to model chemical reactions remains relevant as long as the system is in thermal equilibrium and is well-mixed, i.e. there are many more non-reactive than reactive collisions and the probability of binding does not depend on the molecules' position.

The familiar macroscopic gain-and-loss equation for the deterministic evolution of the continuous concentration $c(t) = x(t)/V$, where V is the system's volume and $x(t)$ is the number of molecules at time t, reads:

$$\frac{d}{dt}c(t) = \kappa_p - \kappa_d\,c(t)\,. \tag{C.2}$$

The relation between Equation C.1 and Equation C.2 becomes clear after applying

the ensemble average to both sides of the former. This average is taken over all discrete states that the system might be in, weighted with the probability of every state, i.e.

$$\langle x(t)\rangle \equiv \sum_{x=0}^{\infty} x\, P(x,t)\,. \tag{C.3}$$

Note that the dependency of x on time on the left hand side stems from the time dependency of P. For the linear chemical scheme, the equivalence $\langle x(t)\rangle/V = c(t)$ is exact. The same applies to reaction-rate constants, k_p and k_d, which are numerically equal to macroscopic κ_p and κ_d, respectively. The mean of a stochastic trajectory obtained from the CME agrees exactly with that obtained from the corresponding macroscopic ODE only when reaction rates depend linearly on the number of molecules. In general, the macroscopic equation is only an approximation to the CME, which omits higher moments representing fluctuations around the average. Reaction-rate constants will also differ (albeit by constant factors only) between the stochastic and the deterministic approaches in case of non-linear dependencies [Gillespie 1977]. We shall illustrate these issues with the following simple non-linear scheme [Gardiner 1983]:

$$X \underset{k_d}{\overset{k_p}{\rightleftharpoons}} 2X\,. \tag{C.4}$$

The backward reaction proceeds at the rate $1/2\, k_d\, x\,(x-1)$ (again, x denotes the number of molecules X), which is quadratic with respect to x. The factor $1/2$ stems from the fact that there are only $x(x-1)/2!$ distinct pairs of molecules X available for the backward reaction. The CME for the above scheme reads,

$$\begin{aligned}\frac{d}{dt}P_x &= k_p\,(x-1)\,P_{x-1} + \frac{k_d}{2}\,x(x+1)\,P_{x+1} \\ &- P_x\left(k_p\,x + \frac{k_d}{2}\,x(x-1)\right)\,,\end{aligned} \tag{C.5}$$

where $P_x \equiv P(x,t)$. Applying the average over x to both sides of Equation C.5 yields:

$$\sum_{x=0}^{\infty} x\,\frac{d}{dt}P_x = k_p \sum_{x=0}^{\infty} xP_x - \frac{k_d}{2}\sum_{x=0}^{\infty} x(x-1)P_x \tag{C.6}$$

As a result of Eq. C.3, every term can be ascribed to a respective ensemble average, i.e. $d\langle x(t)\rangle/dt$, $\langle x(t)\rangle$, and $\langle x(t)(x(t)-1)\rangle$. The equivalence between this equation and the macroscopic solution, $dc(t)/dt = \kappa_p c(t) - \kappa_d c(t)^2$, can be established only by assuming that $\langle x(t)(x(t)-1)\rangle \approx \langle x(t)\rangle\langle x(t)\rangle$. This assumption, known as the mean-field approximation, is valid for large number of molecules. Additionally, the correspondence between backward reaction-rate constants follows, $\kappa_d = V\, k_d/2$.

Analytical solutions of the CME are known only for a handful of reaction schemes [McQuarrie 1967]. More complicated systems require numerical methods such as the Stochastic Simulation Algorithm (SSA) by Gillespie [Gillespie 1977, Gillespie 2007]. It has been applied in numerous studies that investigate the influence of noise on biological systems [Arkin *et al.* 1998, Kierzek *et al.* 2001, Krishna *et al.* 2005].

Reaction-diffusion master equation The reaction-diffusion master equation is an extension of the CME model for systems with correlation length smaller than the size of the system. Like the CME it is an equation for the probability of the realization of a certain state of the system. Space is incorporated by dividing the volume into smaller sub-volumes, which allows to tackle inhomogeneities due to diffusion [Gardiner 1983]. Tracking a single molecule is not possible in this model; unlike in BD, apart from the occupancy of the sub-volumes, no exact positions of molecules are stored. Diffusion in the RDME is added as a state change with rate proportional to the macroscopic diffusion coefficient (and inversely proportional to the size of the sub-volume), which is equivalent to adding a unimolecular reaction. The final form of the RDME is analogous to a semi-discrete form of the reaction-diffusion PDE with the diffusion term discretized using a second-order centered scheme [Bernstein 2005].

The change of state in one sub-volume depends on the rates of the chemical reactions that occur locally within the sub-volume (like in the CME) and the diffusive flux, i.e. the rate of diffusion between the sub-volume and neighboring sub-volumes. It is assumed that within one sub-volume the probability of a reaction does not depend on the position of the molecules, i.e. locally the system is well-stirred. That obviously sets a requirement for the maximum size of the sub-volume if one wants to model a system where diffusion limits the rates of chemical reactions. In principle the size of one sub-volume should be of the order of the correlation length [Baras & Mansour 1996]. Although correct, this estimate is rather difficult to asses for a complex reaction network. The correlation length, also referred as the Kuramoto length, is usually estimated as the square root of the mean time between reactions times the relative diffusion coefficient of the species involved in that reaction [Togashi & Kaneko 2005, van Kampen 1997].

The availability of published software implementations of RDME like MesoRD and Gillespie Multi-Particle (GMP), has recently allowed to study the spatio-temporal dynamics of processes in the cell [Fange & Elf 2006, Elf & Ehrenberg 2004, Rodríguez *et al.* 2006].

C.2 Computational methods

In this section we describe the main features of the stochastic methods used in the Chapter 4. The methods are based on the mesoscopic models presented in the previous section.

Green's Function Reaction Dynamics (GFRD) [van Zon & ten Wolde 2005b, van Zon & ten Wolde 2005a] implements the idea of using analytical solutions to the Smoluchowski diffusion equation (an outline of the algorithm can be found in Tab. C.1). It is used to simulate biochemical networks, and should be especially efficient for those processes consisting of small numbers of molecules. For such sparse systems longer simulation times can be taken because of larger average distances to the nearest-neighboring reacting molecule. Numerical experiments showed that an increase in time step by around four orders of magnitude can be achieved compared to conventional brute-force BD methods (see Fig. 9 of [van Zon & ten Wolde 2005a] for the distribution of propagation times). On the other hand the usage of analytical solutions requires a pretabulation of probability distributions in computer memory with sufficiently fine discretization. Those tables have to be searched every time step in order to obtain the new radial and angular position

Table C.1: An outline of the GFRD algorithm for one binding site and many molecules diffusing around the reactive target (e.g. the simple gene expression described in the main article). Probability distributions for finding the time of the next reaction and the new position around the absorbing target are computed from the solution of the Smoluchowski diffusion equation. For efficiency these distributions are stored in look-up tables which are searched every time step.

Step	Action
1	Loop over all molecules and find the shortest distance to the binding site.
2	Set the maximum possible time step Δt_{max} such that only the closest molecule can reach the target, within a certain threshold.
3A	If the binding site is occupied, move molecules to new positions with Δt_{max}, go to 1.
3B	Else, use the solution of the Smoluchowski equation for a pair of molecules to find the time of the next reaction t_R from the appropriate probability distribution.
4A	If $t_R < \Delta t_{max}$ do the reaction, find new positions for the rest of the molecules, go to 1.
4B	Else do not react, find new positions for all molecules, go to 1.

around the target. This can be a costly procedure which reduces the gain of the increase of the simulation time step.

In general, GFRD is an appealing approach because it is an exact method to solve the diffusion problem with chemical reactions. It is efficient for sparse systems (nanomolar concentrations) and this is exactly the situation where spatial fluctuations can be significant. GFRD could also be easily used in a hybrid method dealing with multi-scale problems since coupling to continuum methods requires only conservation of mass. The main drawback is the complexity of the method. Also a usable, general tool based on this approach is not available.

Smoldyn The idea behind Smoldyn (Smoluchowski Dynamics) [Andrews & Bray 2004] is to set microscopic parameters like the binding and the unbinding radius for the reversible reaction such that the simulation reproduces the known macroscopic behavior. A bimolecular, reversible reaction is modeled such that particles react immediately if their centers are within a certain reaction distance (every collision leads to a reaction). Therefore, instead of sampling from the exact single-particle probability distributions, binding and unbinding radii are adjusted to obtain correct macroscopic steady-state values. In reality, not every collision leads to a reaction (potential barrier), thus Smoldyn uses the *effective binding radius* σ_b. The backward reaction creates a problem in Smoldyn because the particle cannot be placed at contact (with a trap) since it would lead to immediate rebinding with probability 1. Instead, the *unbinding radius* σ_u is introduced. Its magnitude is obtained using the argument that at equilibrium there is no net flux of particles towards a trap: in a given time interval the same number of molecules is lost in a forward reaction as in a backward reaction. Thus the source of molecules at σ_u must match the sink at σ_b. Smoldyn is available as a convenient tool (with source code included).

Table C.2: An outline of the Smoldyn method.

Step	Action
0	Set binding and unbinding radii for all reaction channels.
	Loop over the molecules
1	Execute unimolecular reactions (including dissociation of complexes).
2	Execute bimolecular reactions if two molecules are within the binding radius for a given reaction.
3	Assign new positions to the molecules.
4	Go to 1.

Table C.3: An outline of the SSA algorithm, Gillespie's *direct* method.

Step	Action
1	Compute propensity functions a_ν for all M reaction channels.
2	Calculate the global propensity function $a_0 = \sum_{\nu=1}^{M} a_\nu$.
3	Generate two random numbers r_1 and r_2 uniformly distributed in the interval [0,1).
4	Compute the time of the next reaction: $\tau_R = \frac{1}{a_0} \ln \frac{1}{r_1}$.
5	Compute which reaction ν will occur, i.e. choose ν as the smallest integer for which $\sum_{\nu'=1}^{\nu} a_{\nu'} > r_2\, a_0$.
6	Advance the simulation time by τ_R.
7	Update the number of species affected by the change after reaction ν.
8	Go to 1.

Stochastic Simulation Algorithm (SSA) is the *direct method* developed by Gillespie in 1976 [Gillespie 1976, Gillespie 1977]. Together with his other approach, the *first-reaction method*, they belong to the broader class of Kinetic/Dynamic Monte Carlo methods (see [Fichthorn & Weinberg 1991] and [Voter 2005] for reviews) here applied to a chemically reacting, well-mixed (space not included) system. SSA computes the trajectory of a stochastic discrete system described by the chemical master equation (algorithm in Tab. C.3).

The method scales linearly with the number of reactions which becomes a problem for large chemical networks. Gibson and Bruck [Gibson & Bruck 2000] proposed the *next reaction method*, an exact scheme which is proportional to the logarithm of the number of reactions. This is achieved by minimizing the number of recalculations of the propensity function a_μ. A single reaction changes only a certain fraction of a_μ's in a system with many reaction channels. By using a dependency graph one can compute beforehand which reactions change which propensity functions. In this method only one random number per event is required.

MesoRD is a tool for simulating trajectories of discrete, stochastic systems with space as described by the reaction-diffusion master equation. It uses the *next subvolume method* [Elf & Ehrenberg 2004] which is an analogue of the *next reaction method* [Gibson & Bruck 2000] for systems without space (algorithm in Tab. C.4). The efficiency gain of this approach is especially advantageous when dealing with reaction-diffusion. As it was mentioned in Section C.1 adding diffusion to the CME is in fact equivalent to

adding reactions representing the flow of molecules between the sub-volumes. The event queue (a binary tree) implemented in the *next sub-volume method*, which is equivalent to the dependency graph of Gibson and Bruck [Gibson & Bruck 2000], identifies the sub-volume where the next reaction will trigger. Therefore it is not necessary to scan all the sub-volumes searching for the scheduled event. The method has been shown to produce the same results as if the *direct method* has been employed instead of the *next sub-volume* (see supplementary material of [Elf & Ehrenberg 2004]). MesoRD is available as a convenient and easy-to-use tool [Hattne *et al.* 2005], it accepts input files written in SBML format.

Table C.4: An outline of the next sub-volume algorithm used in MesoRD (adopted from the supplementary material of [Elf & Ehrenberg 2004]). Diffusion between sub-volumes is treated as a unimolecular reaction with the propensity function weighed by the diffusion coefficient and the size of the sub-volume.

Step	Action
	Initialization
0A	For each sub-volume:
	1. compute propensity functions for reactions and diffusion,
	2. calculate time of the next event like in the SSA
	(uses one random number r_0 uniformly distributed in the interval [0,1)).
0B	Sort sub-volumes in an event queue according to the time of the next event.
	Iterations
1	Let α be the sub-volume from the event queue where the next event occurs.
2	Generate a random number $r_1 \in [0, 1)$.
3	Using r_1 compute whether a reaction or a diffusion event occurs.
	Reaction
4A	Reuse r_1 to determine which reaction occurs in the direct SSA method.
4B	Recalculate the propensity function for the sub-volume α.
4C	Generate a new random number r_2, and calculate the time of the next-event t_α.
4D	Insert the next event in the event queue, sort the queue.
4E	Go to 1.
	Diffusion
5A	Reuse r_1 to choose the type of the molecule which diffuses.
5B	Reuse r_1 again to choose the direction of diffusion.
5C	Update propensity functions for the sub-volume α and the neighbor β which got an additional molecule.
5D	Generate a new random number r_2, and calculate the time of the next event in both sub-volumes, α and β.
5E	Insert next events in the event queue, sort the queue.
5F	Go to 1.

Gillespie Multi-Particle Reaction-diffusion processes may also be computed with the Gillespie Multi-Particle algorithm (source code available)[Rodríguez *et al.* 2006]. The basis of this method is the Lattice Gas Automata algorithm [Chopard *et al.* 1994], which is

used for modeling diffusion processes. Single discrete molecules perform a random walk on the lattice; it was demonstrated by Chopard that the macroscopic diffusion equation is reproduced in the limit of $L_{sv} \to 0$, where L_{sv} is the linear size of the lattice site (a sub-volume in RDME terminology). For every species with a diffusion coefficient D, and for a given lattice size L_{sv}, the diffusion time step[6] τ_D is fixed. For a small number of molecules diffusion jumps are performed individually, but above a certain threshold (approximately 90 particles in one site) it is allowed and computationally more efficient to move particles in a bulk according to a Gaussian distribution.

Reaction events are executed between diffusion steps using the SSA algorithm in every sub-volume, and the simulation time is advanced by time intervals between reactions. Once the simulation time reaches one of the fixed diffusion time steps for a diffusing species, the molecules of the respective species are moved to neighboring sub-volumes. As in RDME, it is assumed that chemical reactions are local and that inside the sub-volume the probability of reactions is independent of the molecules' position. Additionally, concentrations should not change significantly between two diffusion steps because reaction and diffusion processes are split. This requirement has an important implication: if the chemical reactions in the system under study are much more frequent than diffusion events, then smaller diffusion time intervals should be used in order to diminish the splitting error. Since τ_D is prescribed by the size of the sub-volumes, this is equivalent to using a finer lattice, which of course hampers the performance. A schematic algorithm is listed in Tab. C.5.

Table C.5: An outline of the GMP method [Rodríguez *et al.* 2006].

Step	Action
	Initialization
0A	Find the diffusion time step τ_{D_S} for every species S.
0B	Find $t_S = \min\{\tau_{D_S}\}$.
0C	Set $n_S = 1$ for all species, $t_{sim} = 0$.
	Iterations
	Reaction
1	While $t_{sim} < t_S$, do:
1A	Reactions in every lattice site with time step τ_R according to the SSA algorithm.
1B	Advance the simulation time $t_{sim} = t_{sim} + \tau_R$.
	Diffusion
2A	Diffuse species S for which $t_S = t_{sim}$.
2B	Increment iteration n_S for the diffused species.
2C	Choose the next diffusion time step $t_S = \min\{\tau_{D_S} \cdot n_S\}$.
3	Go to 1.

[6]The diffusion time step equals $\tau_D = \frac{1}{2d}\frac{L_{sv}^2}{D}$, where d is the dimensionality of the system, L_{sv} – the linear size of the sub-volume, and D – the diffusion coefficient of a given species.

Bibliography

[Acar *et al.* 2005] Murat Acar, Attila Becskei and Alexander van Oudenaarden. *Enhancement of cellular memory by reducing stochastic transitions.* Nature, vol. 435, no. 7039, pages 228–32, 2005. 96

[Acar *et al.* 2008] Murat Acar, Jerome T Mettetal and Alexander van Oudenaarden. *Stochastic switching as a survival strategy in fluctuating environments.* Nat Genet, vol. 40, no. 4, pages 471–5, 2008. 15, 27, 101

[Ackermann *et al.* 2008] Martin Ackermann, Bärbel Stecher, Nikki E Freed, Pascal Songhet, Wolf-Dietrich Hardt and Michael Doebeli. *Self-destructive cooperation mediated by phenotypic noise.* Nature, vol. 454, no. 7207, pages 987–90, 2008. 16

[Agmon & Szabo 1990] Noam Agmon and Attila Szabo. *Theory of reversible diffusion-influenced reactions.* J. Chem. Phys., vol. 92, page 5270, 1990. 92, 118, 121

[Allen & Tildesley 2002] MP Allen and DJ Tildesley. Computer simulation of liquids. 2002. 85

[Alon 2006] Uri Alon. An introduction to systems biology: Design principles of biological circuits. Chapman and Hall/CRC, 2006. 17

[Andrews & Bray 2004] Steven S Andrews and Dennis Bray. *Stochastic simulation of chemical reactions with spatial resolution and single molecule detail.* Physical Biology, vol. 1, no. 3-4, pages 137–51, 2004. 89, 126

[Andrews *et al.* 2010] Steven S Andrews, Nathan J Addy, Roger Brent and Adam P Arkin. *Detailed simulations of cell biology with Smoldyn 2.1.* PLoS Comp. Biol., vol. 6, no. 3, page e1000705, 2010. 83

[Arkin *et al.* 1998] A Arkin, J Ross and H H McAdams. *Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells.* Genetics, vol. 149, no. 4, pages 1633–48, 1998. 83, 124

[Artsimovitch & Landick 2000] I Artsimovitch and R Landick. *Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals.* PNAS, vol. 97, no. 13, pages 7090–5, 2000. 33, 37, 101

[Ashall *et al.* 2009] Louise Ashall, Caroline A Horton, David E Nelson, Pawel Paszek, Claire V Harper, Kate Sillitoe, Sheila Ryan, David G Spiller, John F Unitt, David S Broomhead, Douglas B Kell, David A Rand, Violaine Sée and Michael R H White. *Pulsatile stimulation determines timing and specificity of NF-kappaB-dependent transcription.* Science, vol. 324, no. 5924, pages 242–6, 2009. 102

[Balaban *et al.* 2004] Nathalie Q Balaban, Jack Merrin, Remy Chait, Lukasz Kowalik and Stanislas Leibler. *Bacterial persistence as a phenotypic switch.* Science, vol. 305, no. 5690, pages 1622–5, 2004. 3, 16

[Bar-Even *et al.* 2006] Arren Bar-Even, Johan Paulsson, Narendra Maheshri, Miri Carmi, Erin K O'Shea, Yitzhak Pilpel and Naama Barkai. *Noise in protein expression scales with natural protein abundance.* Nat Genet, vol. 38, no. 6, pages 636–43, 2006. 27, 66

[Bar-Nahum *et al.* 2005] Gil Bar-Nahum, Vitaly Epshtein, Andrei E Ruckenstein, Ruslan Rafikov, Arkady Mustaev and Evgeny Nudler. *A ratchet mechanism of transcription elongation and its control.* Cell, vol. 120, no. 2, pages 183–93, 2005. 33, 37

[Baras & Mansour 1996] F Baras and M. Malek Mansour. *Reaction-diffusion master equation: A comparison with microscopic simulations.* Physical Review E (Statistical Physics, vol. 54, page 6139, 1996. 91, 125

[Batchelor & Goulian 2003] Eric Batchelor and Mark Goulian. *Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system.* PNAS, vol. 100, no. 2, pages 691–6, 2003. 55, 104

[Beaumont *et al.* 2009] Hubertus J E Beaumont, Jenna Gallie, Christian Kost, Gayle C Ferguson and Paul B Rainey. *Experimental evolution of bet hedging.* Nature, vol. 462, no. 7269, pages 90–3, 2009. 16

[Becskei *et al.* 2005] Attila Becskei, Benjamin B Kaufmann and Alexander van Oudenaarden. *Contributions of low molecule number and chromosomal positioning to stochastic gene expression.* Nature Genetics, vol. 37, no. 9, pages 937–44, 2005. 84

[Berg & Purcell 1977] H C Berg and E M Purcell. *Physics of chemoreception.* Biophys J, vol. 20, no. 2, pages 193–219, 1977. 56, 59, 104

[Berg & von Hippel 1985] O G Berg and P H von Hippel. *Diffusion-controlled macromolecular interactions.* Annual review of biophysics and biophysical chemistry, vol. 14, pages 131–60, 1985. 55

[Berg *et al.* 1981] O G Berg, R B Winter and P H von Hippel. *Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory.* Biochemistry, vol. 20, no. 24, pages 6929–48, 1981. 65

[Bernstein 2005] David Bernstein. *Simulating mesoscopic reaction-diffusion systems using the Gillespie algorithm.* Phys. Rev. E, vol. 71, page 41103, 2005. 125

[Besemer *et al.* 2001] J Besemer, A Lomsadze and M Borodovsky. *GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions.* Nucleic Acids Res, vol. 29, no. 12, pages 2607–18, 2001. 64, 79

[Bhalla 2004] Upinder S Bhalla. *Signaling in small subcellular volumes. I. Stochastic and diffusion effects on individual pathways.* Biophys J, vol. 87, no. 2, pages 733–44, 2004. 82

[Binder 2009] P-M Binder. *Computation: The edge of reductionism.* Nature, vol. 459, no. 7245, pages 332–4, 2009. 2

[Bishop *et al.* 2007] Amy L Bishop, Faiza A Rab, Edward R Sumner and Simon V Avery. *Phenotypic heterogeneity can enhance rare-cell survival in 'stress-sensitive' yeast populations.* Molecular Microbiology, vol. 63, no. 2, pages 507–20, 2007. 16

[Blake *et al.* 2003] William J Blake, Mads Kaern, Charles R Cantor and James J Collins. *Noise in eukaryotic gene expression.* Nature, vol. 422, no. 6932, pages 633–7, 2003. 105

[Blake *et al.* 2006] William J Blake, Gábor Balázsi, Michael A Kohanski, Farren J Isaacs, Kevin F Murphy, Yina Kuang, Charles R Cantor, David R Walt and James J Collins. *Phenotypic consequences of promoter-mediated transcriptional noise.* Molecular Cell, vol. 24, no. 6, pages 853–65, 2006. 3, 16, 101, 102

[Bremer *et al.* 2003] H Bremer, P Dennis and M Ehrenberg. *Free RNA polymerase and modeling global transcription in Escherichia coli.* Biochimie, vol. 85, no. 6, pages 597–609, 2003. 38

[Bridgham *et al.* 2009] Jamie T Bridgham, Eric A Ortlund and Joseph W Thornton. *An epistatic ratchet constrains the direction of glucocorticoid receptor evolution.* Nature, vol. 461, no. 7263, pages 515–9, 2009. 17

[Bruggeman *et al.* 2009] Frank J Bruggeman, Nils Blüthgen and Hans V Westerhoff. *Noise management by molecular networks.* PLoS Comp. Biol., vol. 5, no. 9, page e1000506, 2009. 3, 17

[Buchan & Stansfield 2007] J Ross Buchan and Ian Stansfield. *Halting a cellular production line: responses to ribosomal pausing during translation.* Biol Cell, vol. 99, no. 9, pages 475–87, 2007. 33, 101

[Cai & Inouye 2002] Sheng Jian Cai and Masayori Inouye. *EnvZ-OmpR interaction and osmoregulation in Escherichia coli.* J Biol Chem, vol. 277, no. 27, pages 24155–61, 2002. 65

[Cai *et al.* 2006] Long Cai, Nir Friedman and X Sunney Xie. *Stochastic protein expression in individual cells at the single molecule level.* Nature, vol. 440, no. 7082, pages 358–62, 2006. 27, 31, 36, 55, 66

[Cai *et al.* 2008] Long Cai, Chiraj K Dalal and Michael B Elowitz. *Frequency-modulated nuclear localization bursts coordinate gene regulation.* Nature, vol. 455, no. 7212, pages 485–90, 2008. 102

[Carlberg & Dunlop 2006] Carsten Carlberg and Thomas W Dunlop. *An integrated biological approach to nuclear receptor signaling in physiological control and disease.* Crit Rev Eukaryot Gene Expr, vol. 16, no. 1, pages 1–22, 2006. 66

[Chang *et al.* 2008] H Chang, M Hemberg, M Barahona, D Ingber and S Huang. *Transcriptome-wide noise controls lineage choice in mammalian progenitor cells.* Nature, vol. 453, no. 7194, pages 544–547, 2008. 3, 15, 16

[Chopard *et al.* 1994] Bastien Chopard, Laurent Frachebourg and Michel Droz. *Multiparticle Lattice Gas Automata for Reaction Diffusion Systems.* International Journal of Modern Physics C, vol. 5, page 47, 1994. 89, 128

[Chubb *et al.* 2006] Jonathan R Chubb, Tatjana Trcek, Shailesh M Shenoy and Robert H Singer. *Transcriptional pulsing of a developmental gene.* Curr Biol, vol. 16, no. 10, pages 1018–25, 2006. 27, 31, 36, 38

[Clarke & Liu 2008] David C Clarke and Xuedong Liu. *Decoding the quantitative nature of TGF-beta/Smad signaling.* Trends in Cell Biology, vol. 18, no. 9, pages 430–42, 2008. 66

[Cluzel *et al.* 2000] P Cluzel, M Surette and S Leibler. *An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells.* Science, vol. 287, no. 5458, pages 1652–5, 2000. 17

[Cohen *et al.* 2008] A A Cohen, N Geva-Zatorsky, E Eden, M Frenkel-Morgenstern, I Issaeva, A Sigal, R Milo, C Cohen-Saidon, Y Liron, Z Kam, L Cohen, T Danon, N Perzov and U Alon. *Dynamic proteomics of individual cancer cells in response to a drug.* Science, vol. 322, no. 5907, pages 1511–6, 2008. 16, 101

[Collier *et al.* 2007] Justine Collier, Harley H McAdams and Lucy Shapiro. *A DNA methylation ratchet governs progression through a bacterial cell cycle.* PNAS, vol. 104, no. 43, pages 17111–6, Oct 2007. 17

[Colquhoun & Hawkes 1982] D Colquhoun and A G Hawkes. *On the stochastic properties of bursts of single ion channel openings and of clusters of bursts.* Philos Trans R Soc Lond, B, Biol Sci, vol. 300, no. 1098, pages 1–59, 1982. 26

[Core & Lis 2008] Leighton J Core and John T Lis. *Transcription regulation through promoter-proximal pausing of RNA polymerase II.* Science, vol. 319, no. 5871, pages 1791–2, 2008. 101

[Cox & Smith 2006] D Cox and W Smith. *On The Superposition Of Renewal Processes.* Selected Statistical Papers of Sir David Cox, 2006. 115

[Davidson & Surette 2008] Carla J Davidson and Michael G Surette. *Individuality in bacteria.* Annu Rev Genet, vol. 42, pages 253–68, 2008. 15

[Degenhardt *et al.* 2009] Tatjana Degenhardt, Katja N Rybakova, Aleksandra Tomaszewska, Martijn J Moné, Hans V Westerhoff, Frank J Bruggeman and Carsten Carlberg. *Population-level transcription cycles derive from stochastic timing of single-cell transcription.* Cell, vol. 138, no. 3, pages 489–501, 2009. 12

[Dekel & Alon 2005] Erez Dekel and Uri Alon. *Optimality and evolutionary tuning of the expression level of a protein.* Nature, vol. 436, no. 7050, pages 588–92, 2005. 60

[Dobrzyński & Bruggeman 2009] Maciej Dobrzyński and Frank J Bruggeman. *Elongation dynamics shape bursty transcription and translation.* PNAS, vol. 106, no. 8, pages 2583–5, 2009. 22

[Dobrzyński *et al.* 2007] Maciej Dobrzyński, Jordi Vidal Rodríguez, Jaap A Kaandorp and Joke G Blom. *Computational methods for diffusion-influenced biochemical reactions.* Bioinformatics, vol. 23, no. 15, pages 1969–77, 2007. 23

[Dobrzyński 2008] Maciej Dobrzyński. *When do diffusion-limited trajectories become memoryless?* In Proceedings of the 5^{th} Workshop on Computation of Biochemical Pathways and Genetic Network, Heidelberg, Germany, 2008. BioQuant. 21, 23

[Doubrovinski & Howard 2005] Konstantin Doubrovinski and Martin Howard. *Stochastic model for Soj relocation dynamics in Bacillus subtilis.* PNAS, vol. 102, no. 28, pages 9808–13, 2005. 82

[Dronamraju 1999] K R Dronamraju. *Erwin Schrödinger and the origins of molecular biology.* Genetics, vol. 153, no. 3, pages 1071–6, 1999. 3

[Ebeling *et al.* 2003] Werner Ebeling, Lutz Schimansky-Geier and Yuri M. Romanovsky. *Stochastic Dynamics of Reacting Biomolecules.* page 318, 2003. 117

[Echeveria *et al.* 2007] Carlos Echeveria, Kay Tucci and Raymond Kapral. *Diffusion and reaction in crowded environments.* J Phys-Condens Mat, vol. 19, no. 6, page 065146, 2007. 97

[Edelstein & Agmon 1993] Arieh L Edelstein and Noam Agmon. *Brownian dynamics simulations of reversible reactions in one dimension.* J. Chem. Phys., vol. 99, page 5396, 1993. 122

[Eldar & Elowitz 2010] Avigdor Eldar and Michael B Elowitz. *Functional roles for noise in genetic circuits.* Nature, vol. 467, no. 7312, pages 167–73, 2010. 82, 105

[Elf & Ehrenberg 2003] Johan Elf and Måns Ehrenberg. *Fast evaluation of fluctuations in biochemical networks with the linear noise approximation.* Genome Res, vol. 13, no. 11, pages 2475–84, 2003. 26, 62

[Elf & Ehrenberg 2004] J Elf and M Ehrenberg. *Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases.* Syst Biol (Stevenage), vol. 1, no. 2, pages 230–6, 2004. 83, 89, 125, 127, 128

[Elf *et al.* 2007] Johan Elf, Gene-Wei Li and X Sunney Xie. *Probing transcription factor dynamics at the single-molecule level in a living cell.* Science, vol. 316, no. 5828, pages 1191–4, 2007. 55, 65

[Ellis 2001] R J Ellis. *Macromolecular crowding: obvious but underappreciated.* Trends Biochem Sci, vol. 26, no. 10, pages 597–604, 2001. 60

[Elowitz *et al.* 2002] Michael B Elowitz, Arnold J Levine, Eric D Siggia and Peter S Swain. *Stochastic gene expression in a single cell.* Science, vol. 297, no. 5584, pages 1183–6, 2002. 3, 5, 26, 54, 55

[English *et al.* 2006] Brian P English, Wei Min, Antoine M van Oijen, Kang Taek Lee, Guobin Luo, Hongye Sun, Binny J Cherayil, S C Kou and X Sunney Xie. *Ever-fluctuating single enzyme molecules: Michaelis-Menten equation revisited.* Nat Chem Biol, vol. 2, no. 2, pages 87–94, 2006. 105

[Fange & Elf 2006] David Fange and Johan Elf. *Noise-induced Min phenotypes in E. coli.* PLoS Comp. Biol., vol. 2, no. 6, page e80, 2006. 82, 83, 97, 125

[Feinerman *et al.* 2008] Ofer Feinerman, Joël Veiga, Jeffrey R Dorfman, Ronald N Germain and Grégoire Altan-Bonnet. *Variability and robustness in T cell activation from regulated heterogeneity in protein levels.* Science, vol. 321, no. 5892, pages 1081–4, 2008. 3

[Fichthorn & Weinberg 1991] Kristen A Fichthorn and W. H Weinberg. *Theoretical foundations of dynamical Monte Carlo simulations.* J. Chem. Phys., vol. 95, page 1090, 1991. 127

[Fischer & Sauer 2005] Eliane Fischer and Uwe Sauer. *Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism.* Nature Genetics, vol. 37, no. 6, pages 636–40, 2005. 54

[Francke *et al.* 2003] Christof Francke, Pieter W Postma, Hans V Westerhoff, Joke G Blom and Mark A Peletier. *Why the phosphotransferase system of Escherichia coli escapes diffusion limitation.* Biophys J, vol. 85, no. 1, pages 612–22, 2003. 84

[Friedman *et al.* 2006] Nir Friedman, Long Cai and X. Sunney Xie. *Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression.* Phys Rev Lett, vol. 97, page 168302, 2006. 27, 105

[Galburt *et al.* 2007] Eric A Galburt, Stephan W Grill, Anna Wiedmann, Lucyna Lubkowska, Jason Choy, Eva Nogales, Mikhail Kashlev and Carlos Bustamante. *Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner.* Nature, vol. 446, no. 7137, pages 820–3, 2007. 33

[Gardiner 1983] Crispin W Gardiner. Handbook of stochastic methods: for physics, chemistry and the natural sciences. Springer, 1983. 85, 124, 125

[Gibson & Bruck 2000] MA Gibson and J Bruck. *Efficient exact stochastic simulation of chemical systems with many species and many channels.* J Phys Chem A, vol. 104, no. 9, pages 1876–1889, 2000. 10, 98, 127, 128

[Gillespie 1976] Daniel T Gillespie. *A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions.* Journal of Computational Physics, vol. 22, page 403, 1976. 90, 127

[Gillespie 1977] Daniel T Gillespie. *Exact Stochastic Simulation of Coupled Chemical Reactions.* J. Phys. Chem., vol. 81, no. 25, 1977. 90, 124, 127

[Gillespie 1992] Daniel T Gillespie. *A rigorous derivation of the chemical master equation.* Physica A, vol. 188, page 404, 1992. 90, 122

[Gillespie 2007] Daniel T Gillespie. *Stochastic simulation of chemical kinetics.* Annual review of physical chemistry, vol. 58, pages 35–55, 2007. 10, 124

[Golding *et al.* 2005] Ido Golding, Johan Paulsson, Scott M Zawilski and Edward C Cox. *Real-time kinetics of gene activity in individual bacteria.* Cell, vol. 123, no. 6, pages 1025–36, 2005. 26, 27, 31, 36, 37, 66, 96

[Gore *et al.* 2009] Jeff Gore, Hyun Youk and Alexander van Oudenaarden. *Snowdrift game dynamics and facultative cheating in yeast.* Nature, vol. 459, no. 7244, pages 253–6, 2009. 16

[Greive & von Hippel 2005] Sandra J Greive and Peter H von Hippel. *Thinking quantitatively about transcriptional regulation.* Nat Rev Mol Cell Biol, vol. 6, no. 3, pages 221–32, 2005. 33, 37, 101

[Grigoriev *et al.* 2002] Igor V Grigoriev, Yurii A Makhnovskii, Alexander M Berezhkovskii and Vladimir Yu Zitserman. *Kinetics of escape through a small hole.* J Chem Phys, vol. 116, page 9574, 2002. 57

[Halford & Marko 2004] Stephen E Halford and John F Marko. *How do site-specific DNA-binding proteins find their targets?* Nucleic Acids Res, vol. 32, no. 10, pages 3040–52, 2004. 55, 65, 84

[Hatoum & Roberts 2008] Asma Hatoum and Jeffrey Roberts. *Prevalence of RNA polymerase stalling at Escherichia coli promoters after open complex formation.* Mol Microbiol, vol. 68, no. 1, pages 17–28, 2008. 33, 37

[Hattne *et al.* 2005] Johan Hattne, David Fange and Johan Elf. *Stochastic reaction-diffusion simulation with MesoRD.* Bioinformatics, vol. 21, no. 12, pages 2923–4, 2005. 89, 128

[Hayes & Sauer 2003] Christopher S Hayes and Robert T Sauer. *Cleavage of the A site mRNA codon during ribosome pausing provides a mechanism for translational quality control.* Molecular Cell, vol. 12, no. 4, pages 903–11, 2003. 33, 101

[Hellingwerf 2005] Klaas J Hellingwerf. *Bacterial observations: a rudimentary form of intelligence?* Trends in Microbiology, vol. 13, no. 4, pages 152–8, 2005. 13

[Hoch & Silhavy 1995] James A. Hoch and Thomas J. Silhavy. *Two-component signal transduction.* Book, page 488, 1995. 54, 55

[Hooshangi *et al.* 2005] Sara Hooshangi, Stephan Thiberge and Ron Weiss. *Ultrasensitivity and noise propagation in a synthetic transcriptional cascade.* PNAS, vol. 102, no. 10, pages 3581–6, 2005. 3, 17

[Hopfield 1974] J J Hopfield. *Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity.* PNAS, vol. 71, no. 10, pages 4135–9, 1974. 17

[Iber 2006] Dagmar Iber. *A quantitative study of the benefits of co-regulation using the spoIIA operon as an example.* Mol Syst Biol, vol. 2, page 43, 2006. 62

[Imlay 2003] James A Imlay. *Pathways of oxidative damage.* Annu Rev Microbiol, vol. 57, pages 395–418, 2003. 54

[Isaacs *et al.* 2003] Farren J Isaacs, Jeff Hasty, Charles R Cantor and James J Collins. *Prediction and measurement of an autoregulatory genetic module.* PNAS, vol. 100, no. 13, pages 7714–9, 2003. 14

[Jülicher & Bruinsma 1998] F Jülicher and R Bruinsma. *Motion of RNA polymerase along DNA: a stochastic model.* Biophys J, vol. 74, no. 3, pages 1169–85, 1998. 64

[Kaern *et al.* 2005] Mads Kaern, Timothy C Elston, William J Blake and James J Collins. *Stochasticity in gene expression: from theories to phenotypes.* Nat Rev Genet, vol. 6, no. 6, pages 451–64, 2005. 26, 82

[Kamoun & All 1994] F Kamoun and M All. *Statistical Analysis of the Traffic Generated by the Superposition of N Independent Interrupted ....* Information Theory and Applications: Third Canadian Workshop ..., 1994. 48

[Kaufmann & van Oudenaarden 2007] Benjamin B Kaufmann and Alexander van Oudenaarden. *Stochastic gene expression: from single molecules to the proteome.* Curr Opin Genet Dev, vol. 17, no. 2, pages 107–12, 2007. 26, 31

[Kepler & Elston 2001] Thomas B Kepler and Timothy C Elston. *Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations.* Biophys J, vol. 81, no. 6, pages 3116–36, 2001. 16

[Keseler *et al.* 2009] Ingrid M Keseler, César Bonavides-Martínez, Julio Collado-Vides, Socorro Gama-Castro, Robert P Gunsalus, D Aaron Johnson, Markus Krummenacker, Laura M Nolan, Suzanne Paley, Ian T Paulsen, Martin Peralta-Gil, Alberto Santos-Zavaleta, Alexander Glennon Shearer and Peter D Karp. *EcoCyc: a comprehensive view of Escherichia coli biology.* Nucleic Acids Res, vol. 37, no. Database issue, pages D464–70, 2009. 54, 65

[Khare *et al.* 2009] Anupama Khare, Lorenzo A Santorelli, Joan E Strassmann, David C Queller, Adam Kuspa and Gad Shaulsky. *Cheater-resistance is not futile.* Nature, vol. 461, no. 7266, pages 980–2, 2009. 16

[Kholodenko *et al.* 1999] Boris N Kholodenko, Oleg V Demin, Gisela Moehren and Jan B Hoek. *Quantification of Short Term Signaling by the Epidermal Growth Factor Receptor.* J. Biol. Chem., vol. 274, no. 42, pages 30169–30181, 1999. 66

[Kierzek *et al.* 2001] A M Kierzek, J Zaim and P Zielenkiewicz. *The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression.* J Biol Chem, vol. 276, no. 11, pages 8165–72, 2001. 83, 124

[Kim & Shin 1999] Hyojoon Kim and Kook Joe Shin. *Exact Solution of the Reversible Diffusion-Influenced Reaction for an Isolated Pair in Three Dimensions.* Phys Rev Lett, vol. 82, page 1578, 1999. 122

[Krishna *et al.* 2005] Sandeep Krishna, Bidisha Banerjee, T V Ramakrishnan and G V Shivashankar. *Stochastic simulations of the origins and implications of long-tailed distributions in gene expression.* PNAS, vol. 102, no. 13, pages 4771–6, 2005. 83, 124

[Kuczura 1973] Anatol Kuczura. *The Interrupted Poisson Process As an Overflow Process.* The Bell System Technical Journal, vol. 52, no. 3, pages 437–448, 1973. 26, 27, 38

[Kussell & Leibler 2005] Edo Kussell and Stanislas Leibler. *Phenotypic diversity, population growth, and information in fluctuating environments.* Science, vol. 309, no. 5743, pages 2075–8, 2005. 14

[Kussell *et al.* 2005] Edo Kussell, Roy Kishony, Nathalie Q Balaban and Stanislas Leibler. *Bacterial persistence: a model of survival in changing environments.* Genetics, vol. 169, no. 4, pages 1807–14, 2005. 16

[Lamm & Schulten 1981] Gene Lamm and Klaus Schulten. *Extended Browian dynamics approach to diffusion-controlled processes.* J. Chem. Phys., vol. 75, page 365, 1981. 121

[Legewie *et al.* 2008] Stefan Legewie, Hanspeter Herzel, Hans V Westerhoff and Nils Blüthgen. *Recurrent design patterns in the feedback regulation of the mammalian signalling network.* Mol Syst Biol, vol. 4, page 190, 2008. 17

[Lengeler *et al.* 1999] Joseph W. Lengeler, Gerhart Drews and Hans Günter Schlegel. *Biology of the prokaryotes.* page 955, 1999. 63

[Li & Qian 2002] Guangpu Li and Hong Qian. *Kinetic timing: a novel mechanism that improves the accuracy of GTPase timers in endosome fusion and other biological processes.* Traffic, vol. 3, no. 4, pages 249–55, 2002. 17

[Lipkow *et al.* 2005] Karen Lipkow, Steven S Andrews and Dennis Bray. *Simulated diffusion of phosphorylated CheY through the cytoplasm of Escherichia coli.* J Bacteriol, vol. 187, no. 1, pages 45–53, 2005. 83, 87, 88, 93, 97, 122

[Lipkow 2006] Karen Lipkow. *Changing cellular location of CheZ predicted by molecular simulations.* PLoS Comp. Biol., vol. 2, no. 4, page e39, 2006. 83, 97

[Løvdok *et al.* 2007] Linda Løvdok, Markus Kollmann and Victor Sourjik. *Co-expression of signaling proteins improves robustness of the bacterial chemotaxis pathway.* J Biotechnol, vol. 129, no. 2, pages 173–80, 2007. 62

[Løvdok *et al.* 2009] Linda Løvdok, Kajetan Bentele, Nikita Vladimirov, Anette Müller, Ferencz S Pop, Dirk Lebiedz, Markus Kollmann and Victor Sourjik. *Role of translational coupling in robustness of bacterial chemotaxis pathway.* PLoS Biol, vol. 7, no. 8, page e1000171, 2009. 106

[Marion *et al.* 2002] Glenn Marion, Xuerong Mao, Eric Renshaw and Junli Liu. *Spatial heterogeneity and the stability of reaction states in autocatalysis.* Phys Rev E Stat Nonlin Soft Matter Phys, vol. 66, no. 5 Pt 1, page 051915, 2002. 82

[McAdams & Arkin 1999] H H McAdams and Adam P Arkin. *It's a noisy business! Genetic regulation at the nanomolar scale.* Trends Genet, vol. 15, no. 2, pages 65–9, 1999. 3, 17, 82

[McQuarrie 1967] Donald McQuarrie. *Stochastic Approach to Chemical Kinetics.* Journal of Applied Probability, vol. 4, no. 3, pages 413–478, 1967. 124

[Medhi 1994] J Medhi. *Stochastic Processes.* 1994. 116

[Medhi 2003] J Medhi. *Stochastic Models in Queueing Theory.* page 450, 2003. 114, 115

[Metzler 2001] Ralf Metzler. *The Future is Noisy: The Role of Spatial Fluctuations in Genetic Switching.* Phys Rev Lett, vol. 87, page 68103, 2001. 85

[Milne 1982] C Milne. *Transient Behaviour of the Interrupted Poisson Process.* Journal of the Royal Statistical Society. Series B, 1982. 38

[Mitarai *et al.* 2008] Namiko Mitarai, Ian B Dodd, Michael T Crooks and Kim Sneppen. *The generation of promoter-mediated transcriptional noise in bacteria.* PLoS Comp. Biol., vol. 4, no. 7, page e1000109, 2008. 26, 27

[Mitchell *et al.* 2009] Amir Mitchell, Gal H Romano, Bella Groisman, Avihu Yona, Erez Dekel, Martin Kupiec, Orna Dahan and Yitzhak Pilpel. *Adaptive prediction of environmental changes by microorganisms.* Nature, vol. 460, no. 7252, pages 220–4, 2009. 13, 17

[Morelli & ten Wolde 2008] Marco J Morelli and Pieter Rein ten Wolde. *Reaction Brownian dynamics and the effect of spatial fluctuations on the gain of a push-pull network.* J Chem Phys, vol. 129, no. 5, page 054112, 2008. 82

[Nagar & Pradhan 2003] Apoorva Nagar and Punyabrata Pradhan. *First passage time distribution in random walks with absorbing boundaries.* Physica A, vol. 320, page 141, 2003. 120

[Newman *et al.* 2006] John R S Newman, Sina Ghaemmaghami, Jan Ihmels, David K Breslow, Matthew Noble, Joseph L DeRisi and Jonathan S Weissman. *Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise.* Nature, vol. 441, no. 7095, pages 840–6, 2006. 27, 54, 105

[Northrup *et al.* 1984] Scott H Northrup, Stuart A Allison and J. Andrew McCammon. *Brownian dynamics simulation of diffusion-influenced bimolecular reactions.* J. Chem. Phys., vol. 80, page 1517, 1984. 121

[Oliveira *et al.* 2010] Rodrigo F Oliveira, Anna Terrin, Giulietta Di Benedetto, Robert C Cannon, Wonryull Koh, MyungSook Kim, Manuela Zaccolo and Kim T Blackwell. *The role of type 4 phosphodiesterases in generating microdomains of cAMP: large scale stochastic simulations.* PLoS ONE, vol. 5, no. 7, page e11725, 2010. 83

[Ozbudak *et al.* 2002] Ertugrul M Ozbudak, Mukund Thattai, Iren Kurtser, Alan D Grossman and Alexander van Oudenaarden. *Regulation of noise in the expression of a single gene.* Nat Genet, vol. 31, no. 1, pages 69–73, 2002. 26, 55

[Paulsson 2004] Johan Paulsson. *Summing up the noise in gene networks.* Nature, vol. 427, no. 6973, pages 415–8, 2004. 26, 62

[Pedraza & Paulsson 2008] Juan M Pedraza and Johan Paulsson. *Effects of molecular memory and bursting on fluctuations in gene expression.* Science, vol. 319, no. 5861, pages 339–43, 2008. 11, 17, 26

[Pedraza & van Oudenaarden 2005] Juan M Pedraza and Alexander van Oudenaarden. *Noise propagation in gene networks.* Science, vol. 307, no. 5717, pages 1965–9, 2005. 96

[Perkins & Swain 2009] T Perkins and P Swain. *Strategies for cellular decision-making.* Mol Syst Biol, vol. 5, page 326, 2009. 82, 105

[Popov & Agmon 2001a] Alexander V Popov and Noam Agmon. *Three-dimensional simulation verifies theoretical asymptotics for reversible binding.* Chem Phys Lett, vol. 340, pages 151–156, 2001. 122

[Popov & Agmon 2001b] Alexander V Popov and Noam Agmon. *Three-dimensional simulations of reversible bimolecular reactions: The simple target problem.* J Chem Phys, vol. 115, page 8921, 2001. 122

[Price 2008] David H Price. *Poised polymerases: on your mark...get set...go!* Molecular Cell, vol. 30, no. 1, pages 7–10, 2008. 38

[Qian 2006] Hong Qian. *Reducing intrinsic biochemical noise in cells and its thermodynamic limit.* Journal of Molecular Biology, vol. 362, no. 3, pages 387–92, 2006. 17

[Qian 2008] Hong Qian. *Cooperativity and specificity in enzyme kinetics: a single-molecule time-based perspective.* Biophys J, vol. 95, no. 1, pages 10–7, 2008. 55, 105, 106

[Raj *et al.* 2006] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas and Sanjay Tyagi. *Stochastic mRNA synthesis in mammalian cells.* Plos Biol, vol. 4, no. 10, page e309, 2006. 27, 31, 36, 103

[Raser & O'Shea 2004] Jonathan M Raser and Erin K O'Shea. *Control of stochasticity in eukaryotic gene expression.* Science, vol. 304, no. 5678, pages 1811–4, 2004. 103

[Redner 2001] Sidney Redner. *A Guide to First-Passage Processes.* 2001. 38, 58, 69, 70, 85, 92, 112, 117, 118

[Rice 1985] Steven A Rice. Diffusion limited reactions. 1985. 121

[Rodríguez *et al.* 2006] Jordi Vidal Rodríguez, Jaap A Kaandorp, Maciej Dobrzyński and Joke G Blom. *Spatial stochastic modelling of the phosphoenolpyruvate-dependent phosphotransferase (PTS) pathway in Escherichia coli.* Bioinformatics, vol. 22, no. 15, pages 1895–901, 2006. 83, 89, 91, 125, 128, 129

[Rosenfeld *et al.* 2005] Nitzan Rosenfeld, Jonathan W Young, Uri Alon, Peter S Swain and Michael B Elowitz. *Gene regulation at the single-cell level.* Science, vol. 307, no. 5717, pages 1962–5, 2005. 26, 54, 66, 96

[Roussel & Zhu 2006] Marc R Roussel and Rui Zhu. *Stochastic kinetics description of a simple transcription model.* Bull Math Biol, vol. 68, no. 7, pages 1681–713, 2006. 49

[Salazar & Höfer 2009] Carlos Salazar and Thomas Höfer. *Multisite protein phosphorylation–from molecular mechanisms to kinetic models.* FEBS J, vol. 276, no. 12, pages 3177–98, 2009. 17

[Santorelli *et al.* 2008] Lorenzo A Santorelli, Christopher R L Thompson, Elizabeth Villegas, Jessica Svetz, Christopher Dinh, Anup Parikh, Richard Sucgang, Adam Kuspa, Joan E Strassmann, David C Queller and Gad Shaulsky. *Facultative cheater mutants reveal the genetic complexity of cooperation in social amoebae.* Nature, vol. 451, no. 7182, pages 1107–10, 2008. 16

[Savageau 1998] M A Savageau. *Demand theory of gene regulation. I. Quantitative development of the theory.* Genetics, vol. 149, no. 4, pages 1665–76, 1998. 37

[Schäferjohann *et al.* 1996] J Schäferjohann, R Bednarski and B Bowien. *Regulation of CO2 assimilation in Ralstonia eutropha: premature transcription termination within the cbb operon.* Journal of Bacteriology, vol. 178, no. 23, pages 6714–9, 1996. 63

[Schulten & Kosztin 2000] K Schulten and I Kosztin. *Lectures in Theoretical Biophysics.* University of Illinois, 2000. 117

[Schuss *et al.* 2007] Z Schuss, A Singer and David Holcman. *The narrow escape problem for diffusion in cellular microdomains.* PNAS, vol. 104, no. 41, pages 16098–103, 2007. 56

[Selinger *et al.* 2003] Douglas W Selinger, Rini Mukherjee Saxena, Kevin J Cheung, George M Church and Carsten Rosenow. *Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation.* Genome Res, vol. 13, no. 2, pages 216–23, 2003. 64

[Shahrezaei & Swain 2008] Vahid Shahrezaei and Peter S Swain. *Analytical distributions for stochastic gene expression.* PNAS, vol. 105, no. 45, pages 17256–61, 2008. 105

[Shalem *et al.* 2008] Ophir Shalem, Orna Dahan, Michal Levo, Maria Rodriguez Martinez, Itay Furman, Eran Segal and Yitzhak Pilpel. *Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation.* Mol Syst Biol, vol. 4, page 223, 2008. 54

[Sharma *et al.* 2010] Sreenath V Sharma, Diana Y Lee, Bihua Li, Margaret P Quinlan, Fumiyuki Takahashi, Shyamala Maheswaran, Ultan McDermott, Nancy Azizian, Lee Zou, Michael A Fischbach, Kwok-Kin Wong, Kathleyn Brandstetter, Ben Wittner, Sridhar Ramaswamy, Marie Classon and Jeff Settleman. *A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations.* Cell, vol. 141, no. 1, pages 69–80, 2010. 101

[Shen *et al.* 2008] Xiling Shen, Justine Collier, David Dill, Lucy Shapiro, Mark Horowitz and Harley H McAdams. *Architecture and inherent robustness of a bacterial cell-cycle control system.* PNAS, vol. 105, no. 32, pages 11340–5, 2008. 17, 65

[Shinar *et al.* 2007] Guy Shinar, Ron Milo, María Rodríguez Martínez and Uri Alon. *Input output robustness in simple bacterial signaling systems.* PNAS, vol. 104, no. 50, pages 19931–5, 2007. 55, 104

[Shnerb *et al.* 2000] N M Shnerb, Y Louzoun, E Bettelheim and S Solomon. *The importance of being discrete: life always wins on the surface.* PNAS, vol. 97, no. 19, pages 10322–4, 2000. 82

[Shoup & Szabo 1982] D Shoup and Attila Szabo. *Role of diffusion in ligand binding to macromolecules and cell-bound receptors.* Biophys J, vol. 40, no. 1, pages 33–9, 1982. 59, 104

[Sigal *et al.* 2006] Alex Sigal, Ron Milo, Ariel Cohen, Naama Geva-Zatorsky, Yael Klein, Yuvalal Liron, Nitzan Rosenfeld, Tamar Danon, Natalie Perzov and Uri Alon. *Variability and memory of protein levels in human cells.* Nature, vol. 444, no. 7119, pages 643–6, 2006. 16, 26, 61, 66, 101, 106

[Simpson *et al.* 2003] Michael L Simpson, Chris D Cox and Gary S Sayler. *Frequency domain analysis of noise in autoregulated gene circuits.* PNAS, vol. 100, no. 8, pages 4551–6, 2003. 26

[Singer *et al.* 2006a] A Singer, Z Schuss, D Holcman and R. S Eisenberg. *Narrow Escape, Part I.* Journal of Statistical Physics, vol. 122, page 437, 2006. 70

[Singer *et al.* 2006b] A Singer, Z Schuss and David Holcman. *Narrow escape. II. The circular disk.* Journal of Statistical Physics, vol. 122, no. 3, pages 465–489, 2006. 70

[Skupsky *et al.* 2010] R Skupsky, J Burnett, J Foley, D Schaffer and A Arkin. *HIV Promoter Integration Site Primarily Modulates Transcriptional Burst Size Rather Than Frequency.* PLoS Comp. Biol., vol. 6, page e1000952, 2010. 103

[Smits *et al.* 2006] Wiep Klaas Smits, Oscar P Kuipers and Jan-Willem Veening. *Phenotypic variation in bacteria: the role of feedback regulation.* Nat Rev Micro, vol. 4, no. 4, pages 259–71, 2006. 15

[Spencer *et al.* 2009] Sabrina L Spencer, Suzanne Gaudet, John G Albeck, John M Burke and Peter K Sorger. *Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis.* Nature, vol. 459, no. 7245, pages 428–32, 2009. 3, 16, 101

[Stehfest 1970] Harald Stehfest. *Algorithm 368: Numerical inversion of Laplace transforms [D5].* Communications of the ACM, vol. 13, no. 1, 1970. 21

[Stock *et al.* 2000] Ann M Stock, V L Robinson and P N Goudreau. *Two-component signal transduction.* Annual Review of Biochemistry, vol. 69, pages 183–215, 2000. 56, 65

[Stundzia & Lumsden 1996] Audrius B Stundzia and Charles J Lumsden. *Stochastic Simulation of Coupled Reaction-Diffusion Processes.* J. Comp. Phys., vol. 127, pages 196–207, 1996. 82

[Sunohara *et al.* 2004] Takafumi Sunohara, Kaoru Jojima, Hideaki Tagami, Toshifumi Inada and Hiroji Aiba. *Ribosome stalling during translation elongation induces cleavage of mRNA being translated in Escherichia coli.* J Biol Chem, vol. 279, no. 15, pages 15368–75, 2004. 33, 101

[Suzek *et al.* 2001] B E Suzek, M D Ermolaeva, M Schreiber and S L Salzberg. *A probabilistic method for identifying start codons in bacterial genomes.* Bioinformatics, vol. 17, no. 12, pages 1123–30, 2001. 64, 79

[Swain 2004] Peter S Swain. *Efficient attenuation of stochasticity in gene expression through post-transcriptional control.* Journal of Molecular Biology, vol. 344, no. 4, pages 965–76, 2004. 62

[Swameye *et al.* 2003] I Swameye, T G Muller, J Timmer, O Sandra and U Klingmuller. *Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling.* PNAS, vol. 100, no. 3, pages 1028–33, 2003. 66

[Takahashi *et al.* 2005] Kouichi Takahashi, Satya Nanda Vel Arjunan and Masaru Tomita. *Space in systems biology of signaling pathways–towards intracellular molecular crowding in silico.* FEBS Letters, vol. 579, no. 8, pages 1783–8, 2005. 83

[Tan *et al.* 1993] Raymond C Tan, Thanh N Truong and J Andrew McCammon. *Acetylcholinesterase: electrostatic steering increases the rate of ligand binding.* Biochemistry, vol. 32, pages 401–403, 1993. 121

[Tănase-Nicola & ten Wolde 2008] Sorin Tănase-Nicola and Pieter Rein ten Wolde. *Regulatory control and the costs and benefits of biochemical noise.* PLoS Comp. Biol., vol. 4, no. 8, page e1000125, 2008. 5

[Tănase-Nicola *et al.* 2006] Sorin Tănase-Nicola, Patrick B Warren and Pieter Rein ten Wolde. *Signal detection, modularity, and the correlation between extrinsic and intrinsic noise in biochemical networks.* Phys Rev Lett, vol. 97, no. 6, page 068102, 2006. 17

[Thattai & van Oudenaarden 2002] Mukund Thattai and Alexander van Oudenaarden. *Attenuation of noise in ultrasensitive signaling cascades.* Biophys J, vol. 82, no. 6, pages 2943–50, 2002. 3

[To & Maheshri 2010] Tsz-Leung To and Narendra Maheshri. *Noise can induce bimodality in positive transcriptional feedback loops without bistability.* Science, vol. 327, no. 5969, pages 1142–5, 2010. 16

[Togashi & Kaneko 2001] Y Togashi and K Kaneko. *Transitions induced by the discreteness of molecules in a small autocatalytic system.* Phys Rev Lett, vol. 86, no. 11, pages 2459–62, 2001. 82

[Togashi & Kaneko 2004] Yuichi Togashi and Kunihiko Kaneko. *Molecular discreteness in reaction-diffusion systems yields steady states not seen in the continuum limit.* Phys Rev E Stat Nonlin Soft Matter Phys, vol. 70, no. 2 Pt 1, page 020901, 2004. 82

[Togashi & Kaneko 2005] Yuichi Togashi and Kunihiko Kaneko. *Discreteness-induced stochastic steady state in reaction diffusion systems: self-consistent analysis and stochastic simulations.* Physica D: Nonlinear Phenomena, vol. 205, page 87, 2005. 91, 125

[Ulrich & Zhulin 2007] Luke E Ulrich and Igor B Zhulin. *MiST: a microbial signal transduction database.* Nucleic Acids Res, vol. 35, no. Database issue, pages D386–90, 2007. 62, 64, 65, 78, 105

[Ulrich *et al.* 2005] Luke E Ulrich, Eugene V Koonin and Igor B Zhulin. *One-component systems dominate signal transduction in prokaryotes.* Trends in Microbiology, vol. 13, no. 2, pages 52–6, 2005. 55

[van Kampen 1997] Nico G van Kampen. Stochastic processes in physics and chemistry. North Holland, 1997. 55, 84, 122, 125

[van Zon & ten Wolde 2005a] Jeroen S van Zon and Pieter Rein ten Wolde. *Green's-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space.* J. Chem. Phys., vol. 123, no. 23, page 234910, 2005. 55, 82, 88, 92, 122, 125

[van Zon & ten Wolde 2005b] Jeroen S van Zon and Pieter Rein ten Wolde. *Simulating biochemical networks at the particle level and in time and space: Green's function reaction dynamics.* Phys Rev Lett, vol. 94, no. 12, page 128103, 2005. 85, 125

[van Zon *et al.* 2006] Jeroen S van Zon, Marco J Morelli, Sorin Tănase-Nicola and Pieter Rein ten Wolde. *Diffusion of transcription factors can drastically enhance the noise in gene expression.* Biophys J, vol. 91, no. 12, pages 4350–67, 2006. 26, 82, 96, 122

[Voliotis *et al.* 2008] Margaritis Voliotis, Netta Cohen, Carmen Molina-París and Tanniemola B Liverpool. *Fluctuations, pauses, and backtracking in DNA transcription.* Biophys J, vol. 94, no. 2, pages 334–48, 2008. 37

[Voter 2005] Arthur F Voter. *Introduction to the Kinetic Monte Carlo Method.* 2005. 127

[Wagner 2007] Andreas Wagner. Robustness and evolvability in living systems. Princeton University Press, 2007. 17

[Walczak *et al.* 2005] Aleksandra M Walczak, Masaki Sasai and Peter G Wolynes. *Self-consistent proteomic field theory of stochastic gene switches.* Biophys J, vol. 88, no. 2, pages 828–50, 2005. 26, 27

[Weiss *et al.* 1983] George H Weiss, Kurt E Shuler and Katja Lindenberg. *Order statistics for first passage times in diffusion processes.* Journal of Statistical Physics, vol. 31, no. 2, pages 255–278, 1983. 113

[Weiss *et al.* 1992] V Weiss, F Claverie-Martin and B Magasanik. *Phosphorylation of nitrogen regulator I of Escherichia coli induces strong cooperative binding to DNA essential for activation of transcription.* PNAS, vol. 89, no. 11, pages 5088–92, 1992. 65

[Weiss 1967] George H Weiss. *First passage time problems in chemical physics.* Advances in Chemical Physics, vol. 13, pages 1–18, 1967. 69, 117

[Weiss 1994] George H. Weiss. *Aspects and Applications of the Random Walk.* Book, 1994. 117, 121

[Wen *et al.* 2008] Jin-Der Wen, Laura Lancaster, Courtney Hodges, Ana-Carolina Zeri, Shige H Yoshimura, Harry F Noller, Carlos Bustamante and Ignacio Tinoco. *Following translation by single ribosomes one codon at a time.* Nature, vol. 452, no. 7187, pages 598–603, 2008. 33

[West & Stock 2001] A H West and Ann M Stock. *Histidine kinases and response regulator proteins in two-component signaling systems.* Trends Biochem Sci, vol. 26, no. 6, pages 369–76, 2001. 74

[Wolf *et al.* 1995] D A Wolf, L J Strobl, A Pullner and D Eick. *Variable pause positions of RNA polymerase II lie proximal to the c-myc promoter irrespective of transcriptional activity.* Nucleic Acids Res, vol. 23, no. 17, pages 3373–9, 1995. 38

[Wunderlich & Mirny 2008] Zeba Wunderlich and Leonid A Mirny. *Spatial effects on the speed and reliability of protein-DNA search.* Nucleic Acids Res, vol. 36, no. 11, pages 3570–8, 2008. 55, 65

[Yin *et al.* 1999] H Yin, I Artsimovitch, R Landick and J Gelles. *Nonequilibrium mechanism of transcription termination from observations of single RNA polymerase molecules.* PNAS, vol. 96, no. 23, pages 13124–9, 1999. 33

[Yu *et al.* 2006] Ji Yu, Jie Xiao, Xiaojia Ren, Kaiqin Lao and X Sunney Xie. *Probing gene expression in live cells, one protein molecule at a time.* Science, vol. 311, no. 5767, pages 1600–3, 2006. 5, 16, 26, 27, 31, 36, 37, 55, 101

[Yuste *et al.* 2001] Santos B Yuste, L Acedo and K Lindenberg. *Order statistics for d-dimensional diffusion processes.* Physical Review E, vol. 64, no. 5, page 052102, 2001. 60

[Zhdanov 2002] Vladimir P Zhdanov. *Cellular oscillator with a small number of particles.* Eur. Phys. J. B, vol. 29, pages 485–489, 2002. 82

[Zwanzig 1990] R Zwanzig. *Diffusion-controlled ligand binding to spheres partially covered by receptors: an effective medium treatment.* PNAS, vol. 87, no. 15, pages 5856–7, 1990. 59, 104

Summary

The focus of this thesis revolves around stochastic effects in the physiology of a single cell. Fluctuations in the concentration of chemical interactants in biochemical processes pose a fundamental physical limit on the precision and robustness of the cellular function. My main objectives are to (i) elucidate the influence of these fluctuations on the behavior of a biological cell, (ii) analyze some of the mechanisms leading to reduction or magnification of molecular noise to the benefit of the organism, (iii) identify biological regimes where costly computer simulations of spatial stochastic biochemical processes can be significantly reduced, (iv) evaluate the performance and accuracy of computational methods dealing with spatial stochastic problems in cellular biology. Below I shall briefly explain the key terminology used in the text such as *stochasticity*, *gene expression*, *population heterogeneity*, *signaling networks*, *memorylessness*, and summarize the main findings.

Stochasticity, or randomness, is inherent to all biochemical processes and stems largely from thermal fluctuations which affect the motion of biomolecules and their relative position thus biasing rates of chemical reactions. Changes in reaction rates, in turn, desynchronize synthesis and degradation of chemical species such that their concentrations vary over time. This variability becomes significant when the number of molecules engaged in a particular biochemical reaction is small, a dozen for instance. This is the case for a number of cellular processes. *Gene expression*, for example, is potentially susceptible to such fluctuations. It is the process of synthesizing the sequence of amino-acids that constitute a protein based on the information contained in a gene. Gene expression often takes place from a single copy of a gene. The extent of the protein synthesis period changes randomly as the molecules responsible for maintaining the gene in the active state (i.e. allowing for expression and subsequent protein synthesis) repeatedly interrupt the chemical binding with the regulatory part of the gene. Consequently, the intervals (or *waiting times*) at which a newly synthesized protein appears in the cell are highly stochastic. As a result, alternating periods of activity and rest in the protein production appear, so called stochastic *bursts*.

A number of additional processes in gene expression may affect the amount of noise in the synthesis intervals. RNA polymerase is a molecule that uses a DNA gene as template to synthesize a complementary strain with genetic information used at a further stage to build a protein. The polymerase moves step by step along the gene pausing occasionally. Depending on the circumstances, randomness in the polymerase's motion may enhance or diminish the noise providing cells with an extra layer of noise control. We investigate this issue in **Chapter 2**. The comparison of various mechanisms generating bursts requires unambiguous measures such as burst size, duration and significance. We introduce and derive them in this chapter. The measures are also applicable to other areas of science dealing with stochastic events such as telecommunication or queueing theory.

Stochastic protein production, bursts in particular, heavily influence the spread, or variance, of the protein number distribution in a cell at a particular instance of time. Since proteins are involved in virtually all cellular processes, large uncertainty in their

concentrations may have far-reaching consequences on the cellular physiology. However, contrary to popular belief, molecular noise is not always a disruptive force. The generation of *population heterogeneity* is one of the processes that benefits from stochastic gene expression. A group of cells with the same set of genes, an isogenic bacterial population, for instance, can differ significantly in the amount of protein species they consist of due to gene expression noise. The state, active or silent, of hundreds of genes can differ from cell to cell resulting in variations in their molecular make-up. Differences in protein expression levels induced solely by stochasticity prescribe contrasting capacities to interact with the environment across the population. By doing so, the population as a whole hedges its existence against fatal and unexpected changes in the surroundings. When a stressful condition arises, at least some cells are prepared to fight it off. This survival strategy known as stochastic *bet-hedging* turns out to be a universal mechanism employed across various evolutionary levels, from microbes, to birds and human societies, to minimize the risk. We illustrated the effects of molecular noise on protein distributions and discussed the physiological consequences of various distributions in **Chapter 1**.

The majority of cellular processes have evolved, however, to circumvent the influence of molecular noise. Single cells constantly monitor the environment and their own state to adjust the gene expression to meet new nutritional conditions or physical parameters such as temperature or osmotic pressure. Noise in the concentration of proteins, the constituents of *signaling networks* responsible for sensing that information, might disrupt processing of the signal and affect the accuracy of decisions taken by the cell. Therefore, the arrangement of biochemical reaction networks into specific structures, such as multi-level cascades or feedback loops, or large concentrations of chemical species became typical adaptations allowing for fast and robust operation largely independent of concentration fluctuations. But how much is a *large* concentration? How many molecules are required to assure a reliable signal sensing and transfer from the outside of the cell to effect an appropriate change in the gene expression? Spatial aspects come into play here; proteins occupying the cell have to find their cognate targets in space to complete the signal relay. One of the simplest schemes, yet ubiquitous in microbes, a so called two-component signaling network was the topic of our analysis in **Chapter 3**. Using theory and computations we characterized the optimal number of proteins comprising the signaling network that effect a quick response to a stimulus without a large energetic investment in the production of network constituents. Our theoretical results provide insights into the maximum swiftness of this type of networks arguably explaining the absence of the two-component-like arrangement in the signaling of higher organisms.

The mathematical description of stochastic biochemical processes typically requires computer simulations as the equations are too involved to evaluate in a pen-and-paper manner. Problems accounting for spatial organization of the cell and stochastic motion of chemical species are particularly challenging. Due to this complexity computations might be lengthy and hungry for CPU power. Various theoretical approximations have been proposed to optimize some aspects of the computations. We discuss them extensively in **Chapter 4** and perform comparisons using two biological test problems. Our results indicate that significant differences in the magnitude of molecular noise may arise across the methods; an important observation given the prime objective of these methods: the simulation of spatial stochastic biochemical networks.

The computational cost of biological problems involving stochasticity and space can

be greatly reduced, however. We devote parts of the Introduction, **Chapter 3** and the Discussion to that topic. The issue is best illustrated by looking at a signaling network just like the one described earlier. In order to communicate the change in extracellular conditions to the machinery able to respond to that change inside the cell the following takes place: a protein with an altered state has to travel from the cellular membrane (which is exposed to the outside world) to an appropriate gene or a set of genes to initiate expression and production of a new set of proteins. In the simplest case, proteins diffuse freely in the cell and search their targets by means of a random walk without any "guidance". Quantification of the speed of this search typically calls for computer simulations that track the random motion of every molecule – a costly procedure from a computational point of view. An approximation can be made, however. The search like the one taking place in cell signaling involves very small targets compared to the size of the cell. Therefore, the search trajectory of a single protein molecule becomes so long that it "forgets" its initial position. The trajectory becomes *memoryless*. In this regime, mathematics is very simple; the time to reach the target is still stochastic but the equation describing the distribution of these times is a well-known exponential function. It is much easier to handle analytically and the spatial aspect of the problem is reduced. Effectively the system becomes independent of the spatial geometry – arguably a mechanism allowing organisms to decouple search times from cell's shape and internal organization.

The thesis contributes to our understanding of the sources of stochasticity in cellular processes and the way the demand for enhancing or diminishing molecular noise shapes cellular structures. The focus on waiting times in stochastic processes and the use of first-passage time theory yields a powerful framework useful in the analysis of single-cell, single-molecule measurements being actively pursued at the moment. The problems studied in this work open a number of avenues for further theoretical research and also for future experimental verification. For instance, the bursts analyzed in **Chapter 2** refer only to temporal patterns in the intervals of protein synthesis. The full description should also account for protein degradation. Bursts as "packets" of proteins existing for a certain amount of time are the topic of current research at CWI, Amsterdam. Similarly, the effects of various statistics in the waiting times of protein synthesis and degradation on the steady-state protein distributions are being actively studied there in order to obtain analytical solutions. Here, only numerical results have been shown in the Introduction. The role of RNA polymerase collisions and pausing during gene transcription in amplifying or attenuating the noise postulated in **Chapter 2** has not been verified experimentally yet. Although very plausible, it is still not clear whether the composition of two-component signaling networks is indeed optimized for speed in signal relay. **Chapter 3** provides some intuitive theoretical results on that issue, however, only an experimental setup could give a definitive answer to an intriguing question: to what extent gene expression noise, gene order or receptor location on the membrane affect performance of the two-component signaling.

Samenvatting[1]

De focus van dit proefschrift is gericht op stochastische effecten in de fysiologie van een enkele cel. Fluctuaties in de concentratie van chemische stoffen die in biochemische processen in interactie zijn stellen een fundamentele fysische limiet aan de precisie en robuustheid van het functioneren van de cel. Mijn belangrijkste doelen zijn (i) de invloed van deze fluctuaties op het gedrag van een biologische cel bloot te leggen, (ii) een aantal mechanismen te analyseren die leiden tot vermindering of vergroting van de moleculaire fluctuaties die in het voordeel zijn van het organisme, (iii) biologische regimes te identificeren waarvoor tijdrovende computersimulaties van ruimtelijke stochastische biochemische processen aanzienlijk kunnen worden gereduceerd, (iv) de efficiëntie en de nauwkeurigheid te evalueren van de computationele methoden om ruimtelijke stochastische problemen in cellulaire biologie door te rekenen. Hieronder zal ik kort de belangrijkste terminologie die in de tekst wordt gebruikt toelichten, zoals *stochasticiteit*, *genexpressie*, *populatie heterogeniteit*, *signaleringsnetwerken* en *geheugenloosheid*. Ook zal ik de belangrijkste bevindingen in dit proefschrift samenvatten.

Stochasticiteit, of willekeur, is onlosmakelijk verbonden met alle biochemische processen. Het vloeit grotendeels voort uit thermische fluctuaties die de beweging van biomoleculen en hun relatieve positie, en dus de snelheid van chemische reacties, beïnvloeden. Op hun beurt verstoren veranderingen in reactiesnelheden de synchronisatie tussen aanmaak en afbraak van chemische stoffen zodanig dat de concentraties variëren in de tijd. Deze variabiliteit wordt significant wanneer slechts een klein aantal moleculen betrokken is bij een bepaalde biochemische reactie, bijvoorbeeld een tiental. Dit is het geval bij een aantal processen in de cel zoals bijvoorbeeld bij *genexpressie*. Dit is het proces van de synthese van de aminozuurketen van een eiwit op basis van de informatie in een gen. Genexpressie vindt vaak plaats vanuit slechts één enkele kopie van een gen en speciale moleculen moeten dat gen in de actieve toestand houden (i.e. het mogelijk maken van expressie en de daaropvolgende eiwitsynthese). De lengte van de eiwitsyntheseperiode verandert willekeurig als die moleculen de chemische binding met het regulatie deel van het gen herhaaldelijk onderbreken. Het gevolg is dat de intervallen (of *wachttijden*) tussen het ontstaan van nieuw gesynthetiseerde eiwitten in de cel zeer stochastisch zijn. Daardoor zullen zich afwisselend periodes van activiteit en rust in de eiwitproductie voordoen, de zogenaamde stochastische *uitbarsting*, of burst.

Een aantal bijkomende processen in de genexpressie kunnen invloed hebben op de hoeveelheid ruis in de synthese-intervallen. RNA-polymerase is een molecuul dat een DNA gen als mal gebruikt om een complementaire keten met genetische informatie te synthetiseren. Deze keten wordt in een later stadium gebruikt om een eiwit te bouwen. De polymerase beweegt stap-voor-stap langs het gen daarbij af en toe pauzerend. Afhankelijk van de omstandigheden, kan willekeur in de beweging van de polymerase de ruis verhogen of verminderen. Hierdoor krijgen de cellen een extra mogelijkheid tot ruiscontrole. Dit

[1] Vertaald door Peter van Heijster

wordt bestudeerd in **Hoofdstuk 2**. Om de verschillende mechanismen die bursts genereren te kunnen vergelijken zijn eenduidige maten nodig zoals omvang, duur en significantie van de burst. Deze worden geïntroduceerd en afgeleid in dit hoofdstuk. Deze maten zijn ook toepasbaar binnen andere gebieden van de wetenschap die te maken hebben met stochastische gebeurtenissen, zoals telecommunicatie of wachtrijtheorie.

Stochastische eiwitproductie in het algemeen en bursts in het bijzonder, hebben grote invloed op de spreiding (of variantie) van de distributie van het aantal moleculen van een eiwit in een cel op een bepaald tijdstip. Aangezien eiwitten betrokken zijn bij vrijwel alle cellulaire processen, kan grote willekeur in de eiwitconcentraties verregaande gevolgen hebben voor de cellulaire fysiologie. Echter, in tegenstelling tot de populaire opvatting, heeft moleculaire ruis niet altijd een ontwrichtende werking. De vorming van *populatie heterogeniteit* is een van de processen die profiteert van stochastische genexpressie. In een groep cellen met dezelfde set genen, een isogene bacteriële populatie bijvoorbeeld, kan als gevolg van ruis in de genexpressie de hoeveelheid eiwitten per cel aanzienlijk verschillen. De toestand, aan of uit, van honderden genen kan verschillen van cel tot cel resulterend in verschillen in hun moleculaire samenstelling. Verschillen in niveaus van eiwitexpressie die uitsluitend veroorzaakt zijn door stochasticiteit zorgen dat er binnen de populatie een contrasterende capaciteit ontstaat om te reageren op de omgeving. Hierdoor beschermt de gehele populatie zichzelf tegen fatale en onverwachte veranderingen in de omgeving. Wanneer een stressvolle toestand ontstaat zullen door de variatie ten minste sommige cellen in staat zijn om zich te verweren. Deze overlevingsstrategie, bekend als stochastische *bet-hedging*, blijkt een universeel mechanisme te zijn om risico's te minimaliseren dat werkzaam is op verschillende evolutionaire niveaus, van microben tot vogels en menselijke samenlevingen. In **Hoofdstuk 1** onderzoeken wij de effecten van moleculaire ruis op eiwitdistributies en bespreken we de fysiologische gevolgen van de verschillende distributies.

De meeste cellulaire processen zijn zodanig geëvolueerd dat de invloed van de moleculaire ruis wordt omzeild. Individuele cellen monitoren voortdurend de omgeving en hun eigen toestand om de genexpressie aan te passen aan nieuwe voedselcondities of fysische parameters zoals temperatuur of osmotische druk. Ruis in de concentratie van eiwitten, de bouwstenen van *signaleringsnetwerken* welke verantwoordelijk zijn voor het waarnemen van die informatie, kan de verwerking van het signaal verstoren en de juistheid van de beslissingen die genomen worden door de cel beïnvloeden. Daarom werden de organisatie van biochemische reactienetwerken in specifieke structuren, zoals "multi-level cascades" of "feedback-loops", of grote concentraties van chemische componenten typerende aanpassingen die een snelle en robuuste werking mogelijk maken die grotendeels onafhankelijk is van fluctuaties in concentraties. Maar wat is een *grote* concentratie? Hoeveel moleculen zijn er nodig voor een betrouwbare waarneming en verwerking van een signaal buiten de cel zodanig dat een juiste verandering in de genexpressie optreedt? Ook ruimtelijke aspecten spelen hierbij een rol: eiwitten in de cel moeten verwante doelen zien te vinden in de ruimte om het signaal te kunnen doorgeven. Een van de eenvoudigste systemen, dat desondanks alomtegenwoordig is in microben, het zogenaamde twee-componenten signaleringsnetwerk, is onderwerp van onze analyse in **Hoofdstuk 3**. Zowel theoretisch als met behulp van computerberekeningen hebben we het optimale aantal eiwitten in het signaleringsnetwerk gekarakteriseerd dat een snelle reactie geeft op een stimulus zonder dat er een grote energetische investering vereist is bij de productie van de bestanddelen. Onze theoretische resultaten geven inzicht in de maximale snelheid van dit soort netwer-

ken en verklaren de afwezigheid van de twee-componenten-achtige signalering in hogere organismen.

De wiskundige beschrijving van stochastische biochemische processen vereist doorgaans computersimulaties omdat de vergelijkingen te ingewikkeld zijn om met pen en papier op te lossen. Vooral beschrijvingen die rekening houden met de ruimtelijke organisatie van de cel en de stochastische bewegingen van de chemische stoffen zijn uitdagend. Vanwege deze complexiteit zijn berekeningen tijdrovend en vergen ze veel CPU-capaciteit. Er zijn dan ook al verschillende theoretische benaderingen voorgesteld om bepaalde aspecten van de berekeningen te optimaliseren. We bespreken deze benaderingen uitgebreid in **Hoofdstuk 4** en vergelijken ze met behulp van twee biologische testproblemen. Uit onze resultaten blijkt dat de verschillende methoden significant verschillende resultaten in de omvang van de moleculaire ruis geven. Dit is een belangrijke constatering, aangezien de voornaamste doelstelling van deze methoden de simulatie van ruimtelijke stochastische biochemische netwerken is.

Echter, de rekentijden voor biologische problemen met stochasticiteit en ruimte kunnen sterk worden verminderd. We besteden delen van de Inleiding, **Hoofdstuk 3** en de Discussie aan dit onderwerp. Deze kwestie wordt het best geïllustreerd door te kijken naar een signaleringsnetwerk zoals eerder beschreven. Het volgende moet gebeuren om de verandering in de extracellulaire condities te communiceren naar de machinerie in de cel die op deze verandering kan reageren: een eiwit met een veranderde toestand moet van het celmembraan (dat is blootgesteld aan de buitenwereld) naar een passend gen of een set van genen reizen om de expressie en de productie van een nieuwe verzameling eiwitten mogelijk te maken. In het eenvoudigste geval kunnen eiwitten zich vrij verspreiden in de cel en zoeken ze hun doelwit door middel van een random walk zonder enige "sturing". Kwantificering van de snelheid van deze zoektocht vraagt normaal gesproken om computersimulaties die de willekeurige beweging van alle moleculen volgen – een kostbare procedure vanuit een computationeel oogpunt. Echter, er kan een benadering worden gemaakt: een zoektocht zoals bij celsignalering is er één naar zeer kleine doelwitten in vergelijking met de grootte van de cel. Hierdoor is de baan van een enkel eiwitmolecuul zo lang dat het zijn aanvankelijke startpunt "vergeet". De baan is zogezegd *geheugenloos*. In dit regime is de wiskunde heel eenvoudig: de tijd om het doelwit te bereiken is nog steeds stochastisch, maar de vergelijking die de verdeling van deze tijden beschrijft is een bekende exponentiële functie. Dit probleem is analytisch veel makkelijker en het ruimtelijke aspect van het probleem wordt gereduceerd. Effectief gezien wordt het systeem onafhankelijk van de ruimtelijke geometrie – dit is aantoonbaar een mechanisme waardoor organismen zoektijden kunnen ontkoppelen van celvorm en interne organisatie.

Dit proefschrift draagt bij aan ons begrip van de bronnen van onzekerheid in cellulaire processen en hoe de eis tot vergroten of verminderen van de moleculaire ruis de celstructuren vormt. De focus op wachttijden in stochastische processen en het gebruik van de *first-passage time* theorie levert een solide kader op dat zeer nuttig zal zijn bij de analyse van metingen aan individuele cellen en aan individuele moleculen. Dit type metingen zijn momenteel onderwerp van lopend onderzoek. De problemen bestudeerd in dit proefschrift geven een aantal suggesties voor verder theoretisch onderzoek en ook voor toekomstige experimentele verificatie. Bijvoorbeeld, de bursts geanalyseerd in **Hoofdstuk 2** hebben alleen betrekking op de temporele patronen in de intervallen van de aanmaak van eiwitten. Een volledige beschrijving dient ook rekening te houden met de afbraak van eiwitten.

Bursts als "pakketjes" eiwitten die voor een bepaalde tijd bestaan zijn onderwerp van lopend onderzoek bij het CWI, Amsterdam. Ook worden hier de effecten van verschillende statistieken in de wachttijden van de aanmaak en afbraak van eiwitten op de eiwitdistributies in evenwicht onderzocht, om zodoende analytische oplossingen te verkrijgen. Merk op dat in de Inleiding alleen numerieke resultaten zijn weergegeven. De invloed van RNA-polymerase botsingen en pauzes tijdens gentranscriptie op het versterken of verzwakken van de ruis die gepostuleerd is in **Hoofdstuk 2** is nog niet experimenteel geverifieerd. Hoewel het zeer aannemelijk is, is het nog niet bewezen dat de samenstelling van twee-componenten signaleringsnetwerken inderdaad optimaal is voor een snelle signaaloverdracht. **Hoofdstuk 3** bevat een aantal intuïtieve theoretische resultaten op dit punt. Echter, slechts een experimentele opstelling kan een definitief antwoord geven op de volgende intrigerende vraag: in welke mate hebben genexpressieruis, genvolgorde en de locatie van de receptor op het membraan invloed op de prestaties van het twee-componenten signaleringsnetwerk.

Dla laików

Zrozumienie procesów zachodzących w organizmach żywych jest wielkim wyzwaniem dla nauki na najbliższe dziesięciolecia. Zadanie nie jest łatwe. Nawet pojedyncza komórka biologiczna o długości jednej tysięcznej milimetra, to miliony cząsteczek chemicznych, m.in. białek, tłuszczy, cukrów, mikroelementów, które oddziaływują ze sobą w dziesiątkach tysięcy reakcji chemicznych. Co więcej, komórka, to zwykle część większego układu, na przykład organizmu. Jej składowe podlegają ciągłym przekształceniom w procesie ewolucji pod wpływem oddziaływania ze środowiskiem.

Aby przetrwać, każdy organizm musi być zdolny do przetwarzania informacji o otoczeniu oraz do podejmowania odpowiednich działań na podstawie jego zmian. Ilość oraz rodzaj pożywienia (np. glukoza lub laktoza) w otoczeniu jednokomórkowej bakterii jest ważną informacją, bowiem każdy rodzaj pożywienia wymaga nieco innego „przetwórstwa". Z kolei jego brak może wymusić na bakterii przekształcenie w formę uśpioną lub migrację w bardziej zasobne rejony. Wprowadzenie w życie każdej z takich decyzji wymaga użycia nieco innego zestawu cząsteczek chemicznych. Niektóre z nich są dostępne w komórce „od ręki", inne należy dopiero wyprodukować na podstawie genów zawartch w DNA. Produkcja nowych komponentów odpowiednich dla nowych warunków otoczenia oznacza wydatek energetyczny dla organizmu, a zła decyzja może być bardzo kosztowna.

Komórka biologiczna to niezwykle skomplikowana, lecz czasem niezbyt rzetelna fabryka. Pomyłki w reakcjach na zmiany się zdarzają, choć jak się okazuje część z nich może być całkiem zamierzona. Brak precyzji bierze się głównie z losowości reakcji biochemicznych. Aby nastąpiła reakcja pomiędzy dwoma rodzajami cząsteczek, obie muszą się spotkać. Nie jest to łatwe biorąc pod uwagę ogromną ilość najróżniejszych substancji w malutkiej komórce i ogólny chemiczny tłok, jaki w niej panuje. Fizyka działa tu również na niekorzyść. To samo zjawisko, które skłania składniki gotującej się zupy do ruchu powoduje, że ruch cząsteczek chemicznych wewnątrz komórki jest dość chaotyczny (tzw. fluktuacje termiczne). Aby mieć pełen obraz, należy również wiedzieć, że część procesów biochemicznych odbywa się z udziałem niewielu cząsteczek, kilkunastu bądź kilkudziesięciu. A przecież przy tak małej ilości komponentów nadmiar lub brak nawet jednej cząsteczki (np. enzymu) może być znaczący. Nawet nieznaczne różnice w ilości cząsteczek przekładają się na szybkość procesów chemicznych, na szybkość reakcji na zmiany otoczenia lub na wybór decyzji co do dalszego postępowania w danych warunkach środowiskowych. Mimo to komórki nie rozpadają się; tworzą komunikujące się ze sobą struktury jak na przykład kolonie bakterii, ewoluują w wielokomórkowe organizmy. Są częścią składową nas samych, myślących organizmów zdolnych do tworzenia nauki, sztuki, poezji. W jaki sposób reakcje chemiczne, nawet na poziomie pojedynczej komórki tworzą spójny i harmonijny taniec, pomimo niesprzyjającej losowości zdarzeń chemicznych?

Na to pytanie staramy się odpowiedzieć w rozdziale 3. na przykładzie prostego, lecz powszechnego u bakterii układu przesyłającego stan otoczenia do wnętrza komórki. Używając wyłącznie opisu matematycznego reakcji chemicznych oraz symulacji komputerowych, jesteśmy w stanie określić jak szybko ten układ jest w stanie zareagować na zmiany

oraz ile cząsteczek chemicznych wystarczy, aby przesłać sygnał szybko i bez przekłamań. Narzędzia matematyczne, których używamy i które rozwijamy w rozdziałach 2. i 3. rzadko stosowane są do tego rodzaju problemów. Ich przybliżenie w kontekście biologii komórki stanowi zatem bazę do analizy dalszych, bardziej złożonych układów.

Ze względu na stopień komplikacji wiele problemów analizowanych w tej pracy wymaga obliczeń przy pomocy komputerów. Porównanie metod służących takim obliczeniom stanowi treść rozdziału 4. Okazuje się, że ze względu na przyjęte założenia niektóre z tych metod mogą dawać całkiem różne wyniki. Z drugiej strony kosztowne obliczenia komputerowe mogą być czasem znacząco ograniczone, dzięki stosownym przybliżeniom. Jedno z takich przybliżeń rozwijamy w rozdziale 1. i stosujemy do analizy zjawisk w dalszych częściach tej pracy. Dzięki temu jesteśmy w stanie zredukować czas pracy komputera kilkunastokrotnie.

Efekty losowe (tzw. stochastyczne) w życiu komórki są powszechne, lecz nie zawsze stanowią przeszkodę dla funkcjonowania organizmu. Niedawne odkrycia pokazują fascynujące mechanizmy, w których losowość jest kluczowa dla przeżycia lub rozwoju organizmu żywego. I tak na przykład pojedyncze osobniki w populacji bakterii są w stanie zróżnicować swoje cechy w sposób całkowicie losowy. Dzięki temu każdy z nich jest przygotowany na innego rodzaju warunki otoczenia. Przygotowanie się na wiele różnych zmian jest zbyt kosztowne dla pojedynczej bakterii. Rozdzielają więc zadania losowo i działają w grupie minimalizując w ten sposób ryzyko na poziomie populacji. Strategia ta znana jest inwestorom giełdowym. Korzystają z niej również bakterie, aby przeżyć, gdy zostaną potraktowane antybiotykami, a także komórki rakowe podczas chemioterapii. Zrozumienie i opis teoretyczny tego rodzaju zjawisk w biologii wymaga jeszcze wielu dalszych badań, a w miarę odkrywania kolejnych prawidłowości rodzą się nowe pytania. Analiza efektów losowych w procesach komórkowych przedstawiona w niniejszej pracy stanowi pomoc w zrozumieniu źródeł owej losowości oraz pokazuje sposób jej matematycznego opisu.