# Stochastic Modelling of Gene Expression: From Single Molecules to Populations

**by**

*Margaritis Voliotis*

**Submitted in accordance with the requirements for the degree of Doctor of Philosophy.**


**The University of Leeds**
**School of Computing**
**School of Mathematics**

**August 2009**

# Declarations

Some parts of the work presented in this thesis (chapters 4, 5, and 6) have been published in the following articles:

**M. Voliotis, N. Cohen, C. Molina-París, and T. B. Liverpool**, "Fluctuations, pauses, and backtracking in DNA transcription", *Biophysical journal*, 94 (2008) 334–348.

**M. Voliotis, N. Cohen, C. Molina-París, and T. B. Liverpool**, "Backtracking and Proofreading in DNA Transcription", *Physical Review Letters*, 102 (2009) 258101.

The research was designed by MV, NC, CMP, and TBL. The research was conducted by MV under the supervision and guidance of NC, CMP, and TBL. The manuscripts were prepared by MV, NC, CMP, and TBL.

# Acknowledgements

I would like to express my deepest gratitude to all my supervisors Netta Cohen, Carmen Molina-París, and Tanniemola B. Liverpool for their guidance and support along the way. Throughout these years they helped me develop my research skills and nourished my love for science. Along I would like to thank my master's supervisor Samuel L. Braustein who brought me in touch with NC.

I am especially grateful to my parents Vlasis Voliotis and Konstantina Volioti and my sister Areti Volioti. They always stood by me and encouraged me to seek my dreams.

A big thanks to all the people of the Biosystems reading group at the university of Leeds for the fruitful discussions we had during our weekly meetings and over lunch time. Special thanks to James Watson who kindly provided me with pieces of his code and Stefano Beri who was there to answer my stupid biology questions and who dared to take me in a biology lab. I would also like to acknowledge the helpful discussions I had with the people of the Amorph group

Finally, I would like to express my gratitude to my housemates and friends Ioannis Klapaftis, Dimitrios Kolovos and Jenny Orfanou. Thank you guys for putting up with me!

The financial support received from the University of Leeds is greatly acknowledged.

# Abstract

Gene expression constitutes a vital life process through which pieces of genetic information stored in the nucleotide sequence of DNA are transformed into functional molecules, namely proteins and RNA chains. These molecules and the intricate network of interactions among them are the driving force behind most cellular processes, including gene expression itself. Also, of particular importance is the regulation of gene expression. By modulating the levels of proteins they produce, cells manage to synchronise their internal workings and adapt to various environmental conditions. Moreover, in this manner cells manage to coordinate their genetically prescribed behaviour when present in populations, such as a developing embryo or a bacterial colony. This thesis presents a theoretical study of gene expression within the context of different organisational levels from the molecular to the cell population level.

On the single molecule level special emphasis is given on the dynamics of the RNA polymerase, the enzyme that carries out the transcription of DNA into RNA. Recent single molecule experiments have shed light on the dynamical behaviour of this molecule as it transcribes DNA. Of particular importance is the direct observation of transient pauses in the process of transcription, induced by the he backward translocation of the enzyme along the DNA template, a phenomenon dubbed *backtracking*. Motivated by this finding and the implications transcriptional pausing has for the regulation of DNA transcription, our work aims at providing a quantitative characterisation of backtracking and the effect of such pauses on the temporal dynamics of the process. Our results indicate that the lifetime of such pauses should obey a wide distribution and can have dramatic effects on the temporal statistics of the transcription process.

A particularly interesting function of backtracking is transcriptional error correction. Indeed, RNAP does not copy the genetic information accurately; thermal fluctuations introduce errors to the process that must be corrected on the fly. A proposed mechanism of transcriptional error correction involves backtracking of the RNA polymerase and the subsequent cleavage of the the erroneous RNA segment. Based on the picture of DNA transcription provided by single molecule experiments we propose a putative model of this editing process. Our work offers a quantitative picture of transcriptional error correction, predicting the error rate in terms of microscopic rates parameters and allowing one to assess the role of backtracking in transcriptional fidelity. Furthermore, our model puts the specific mechanism of error correction into context by linking it to *kinetic proofreading*, a general principle of biological accuracy.

On a different level, the microscopic dynamics of the DNA transcription ought to have direct implications regarding fluctuations in the numbers of RNA species observed within

the cell. These fluctuations have on their turn far-reaching implications regarding cell fate, behaviour and function. To study the effect transcriptional pauses have on the statistics of RNA production we propose an integrated model of DNA transcription. A key element of our model is that several RNAP molecules can transcribe DNA at the same time, moving in tandem on the template. Our results indicate that transcriptional pauses and exclusive interactions between the RNAP molecules, lead to bursts of RNA production and therefore make the process appear more random. Interestingly such pattern of mRNA production has been observed experimentally and hence our model provides a possible explanations of the phenomenon. It also demonstrates how interactions between molecules can affect behaviour at cellular level by introducing fluctuations in the process of gene expression.

At an even higher level, one should appreciate the fact that cells rarely exist in isolation. At this level of description we are interested in how intra-cellular fluctuations of molecular species affect the behaviour of populations of cells. In particular, motivated by the complex social behaviour observed in certain bacterial species, we propose an *in-silico* paradigm of bacterial communication. In a nutshell, the circuit enables cells to communicate and choose between two antagonistic social behaviours. We find that owing to intra-cellular fluctuations the population can exist in two states: for low values of intra-cellular coupling the population appears mixed (disordered), with approximately one half of the cells adopting each behaviour. As the coupling is increased the population a consensus state starts to appear. We study the transition between the two regimes of behaviour and find that intra-cellular fluctuations as well as the size of the population affect the steepness of this transition.

**In memory of my grandmother Katerina**

# Contents

# List of Figures

# List of Tables

# List of Symbols and Acronyms

| | |
|---|---|
| $CV_X$ | coeeficient of variation for $X$ |
| $J$ | probability flux |
| $k_B$ | Boltzmann constant |
| $P, \Pi$ | probability density function |
| $\mathcal{T}$ | first passage time |
| $\beta = \frac{1}{k_B T}$ | $k_B$ - Boltzmann constant, $T$ - absolute temperature |
| $\Gamma(x)$ | Gamma function |
| $\delta(x)$ | Dirac delta function |
| $\delta_{i,j}$ | Kronecker delta |
| $\sigma_X^2$ | variance of $X$ |
| $\langle X \rangle, \mu_X$ | mean value of $X$ |
| $\tilde{f}(s) = \int_0^\infty e^{-st} f(t) dt$ | Laplace transform of $f(t)$ |
| bp | base pair |
| CV | Coefficient of variation |
| DNA | deoxyribonucleic acid |
| dsDNA | double stranded DNA |
| kbp (kb) | kilo base pairs |
| KP | kinetic proofreading |
| MGF | moment generating function |
| NP | nucleolytic proofreading |
| nt | nucleotide |
| NTP | nucleotide triphosphate |
| PDF | probability density function |
| RNA | ribonucleic acid |
| RNAP | RNA polymerase |
| sec | second |
| TEC | transcription elongation complex |

# Chapter 1

# Introduction

This thesis presents a theoretical study of *gene expression*, the vital cellular process through which genetic information is transformed into cell function and structure. The stochastic nature of the process poses as a unifying theme in our work. Indeed, it has long been appreciated that within the cellular environment stochasticity and noise ought to play an important role [122]. In particular, thermal noise constitutes a major player at the molecular level; driving the motion of bio-molecules and the interactions between them. At a higher organisational level, these interactions give rise to cellular processes, such as the one of gene expression. However, due to the stochastic and discrete nature of molecular interactions, cellular processes are endowed with a certain degree of variability. For example, genetically identical cells, under the same environmental conditions can display wide variations in growth rates and physiology [86, 102]; and in general all cellular function and behaviour is subject to probability laws rather than being deterministic. The scope of our work is two-fold: (i) to quantitatively understand certain microscopic aspects of gene expression and characterise phenomena observed at the single-molecule level and (ii) to understand from a bottom-up perspective how dynamics at single molecule level give rise to fluctuations at the cellular level and in turn how these fluctuations affect cellular behaviour.

Cells constitute the building blocks of life [3]. Their essence lies in DNA, the molecule that stores the *genetic information*. During the life-time of a cell, pieces of DNA are constantly transformed into functional molecules, namely proteins and RNA chains, through a process known as *gene expression*. These molecular species participate in the various

structural entities of the cell, drive the various catalytic reactions – including those that are necessary for gene expression – and in general their interactions allow for structure and function to emerge at higher organisational levels. Not surprisingly, in the last century most scientific efforts of understanding life had been in terms of cataloging and characterising (functionally and structurally) these molecules – an approach termed *reductionism*. More recently, advancements in experimental techniques have allowed for a more comprehensive molecular picture to emerge. In particular, the advent of single molecule manipulation techniques [70] has enabled the study of bio-molecules with unprecedented spatial and temporal resolution and has provided a dynamical characterisation of the processes underpinning life at the molecular level.

Part of our work considers the single molecule dynamics of the RNA polymerase (RNAP) – a key player in the process of gene expression. RNAP is the molecule that carries out DNA transcription, copying genetic information from DNA into RNA molecules. RNA transcripts are subsequently used as templates for protein synthesis and in many cases participate actively in other cellular processes. Owing to its essential role, RNAP has been the subject of extensive study and scientific endeavours leading to the discovery and characterisation of RNAP have rewarded researchers with prestigious Nobel prises. More recently, RNAP has also been put under the the scrutiny of single-molecule manipulation techniques [63]. These studies revealed, for example, how RNAP molecules harness thermal fluctuations to drive their motion along the DNA [1]. They also reported frequent pauses during the process of transcription [47, 64, 124]. Such transcriptional pauses had been a well known phenomenon for quite some time and their implications regarding the regulation of the process well appreciated [58,119]. However, single molecule studies provided for the first time a close look at how some of these pauses are induced. In particular, they reported that during some pauses the RNAP translocates backward along the DNA template, a phenomenon dubbed *backtracking*. Motivated by these findings and the biological implications transcriptional pausing could have for DNA transcription, our work aims at providing a quantitative characterisation of backtracking. Our results indicate that the lifetime of backtracking pauses should obey a wide distribution and can have dramatic effects on the temporal statistics of the transcription process.

Backtracking has also been implicated with transcriptional error correction [3]. Indeed, RNAP does not copy the genetic information accurately. Thermal fluctuations driving the motion of the RNAP along the DNA also introduce errors to the process. These errors must be corrected on the fly to allow for functional RNAs and proteins to be produced [3]. One proposed mechanisms of transcriptional error correction involves a transient pause during which the RNAP steps back along the DNA to allow cleavage of the

erroneous RNA segment [3,58]. However key questions still remain open [30]. How does the RNAP know where to cleave? What fidelity levels are accomplished through such a mechanism? Based on the picture of DNA transcription provided by single-molecule experiments we propose a putative model of this editing process. Our model offers a quantitative picture of transcriptional error correction that allows one to assess the role of backtracking in providing the necessary levels of transcriptional fidelity. Furthermore, our model puts the specific mechanism of error correction into context by linking it to *kinetic proofreading* [68, 101], a general principle regarding accuracy in biological processes.

Transcriptional pauses, however, can also have implications that are perhaps better appreciated at a higher level of organisation. Inside cells, bio-molecules are constantly interacting with each other. It is this inherently complex network of interactions that gives rise to interesting behaviour not seen in inanimate physical systems. Here, one customarily thinks in terms of modules instead of individual molecules [62]. These modules, similar to engineering disciplines, correspond to small groups of interacting components that give rise to quasi-independent functions such as gene expression, signal transduction and cell division, to name a few. The study of life this level of organisation provides a complementary picture to that of reductionism and has lately has lately come to be known as *molecular systems biology* [75]. At this level, one is particularly interested in the role of gene expression noise and how fluctuations in the levels of molecular species affect the functions and behaviour of the cell [83].

Transcriptional pauses affect the temporal dynamics of transcription and hence ought to have a direct effect on the fluctuations in the levels of RNAs and proteins within cells. These fluctuations have on their turn far-reaching implications regarding cell fate, behaviour and functioning [25]. To study the effect transcriptional pauses have on the statistics of RNA populations we propose and study an integrated model of DNA transcription. A key element of our model is that several RNAP molecules can transcribe DNA at the same time, moving in tandem on the template. Our results indicate that due to transcriptional pauses and exclusive interactions between the RNAP molecules, RNA production appears more random, occurring in bursts. Interestingly, this pattern of RNA production has been experimentally observed [27, 55, 114]. Our model, therefore, provides a possible explanations of the phenomenon. It also demonstrates how interactions between molecules can affect behaviour at cellular level by introducing fluctuations in the process of gene expression.

At an even higher level, one should appreciate the fact that cells rarely exist in isolation. Higher organisms (*eukaryotes*) usually consist of a number of cells. These cells constantly communicate and interact to achieve common goals. During development, for

example, cells are constantly coordinated through chemical signals and differentiate to achieve the genetically prescribed anatomy of the organisms. Unicellular organisms are also capable of communication when present in populations or colonies. Communication enables bacterial cells to coordinate their behaviour with respect to environmental stimuli and renders them with astonishingly complex social behaviours [150]. Moreover, it enables certain species to break the barriers of unicellularity and behave remarkably similarly to multi-cellular organisms, cooperating for the survival of the whole rather than the individual [131].

At this level of description we are interested in how sub-cellular fluctuations in the levels of molecular species affect the behaviour of populations of cells. In particular, motivated by the complex social behaviour observed in certain bacterial species, we propose an *in-silico* paradigm of bacterial communication. In a nutshell, the circuit enables cells to communicate and choose between two antagonistic social behaviours. We find that owing to sub-cellular fluctuations the population can exist in two states: for low values of intra-cellular coupling the population appears mixed (disordered), with approximately one half of the cells adopting each behaviour. As the coupling is increased the population a consensus state starts to appear. We study the transition between the two regimes and find that sub-cellular fluctuations hinder the ability of cell to synchronise their behaviour.

The thesis is organised as follows. Chapters 2 and 3 present background material that is regarded essential for the reading of the thesis. In particular, Chapter 2 introduces the reader to some key concepts of molecular biology, focusing on the processes of gene expression and DNA transcription. Chapter 3 provides a brief introduction to the mathematical and computational tools used throughout the thesis. More specifically, the theory of stochastic processes is reviewed, and the reader is introduced to the Master equation and existing analytical and computational methods used for solving it. Chapter 4-7 present the main results of the thesis. In Chapter 4 a stochastic model of the transcription elongation dynamics is presented and used to study transcriptional pausing. In Chapter 5 we build upon the model of the elongation dynamics focusing on a quantitative characterisation of transcriptional error correction. Next (Chapter 6), an integrated model of DNA transcription is presented and used to study the effect of transcriptional pauses on the statistics of RNA production. Finally, Chapter 7 focuses on the cell population level: the *in-silico* model of bacterial communication is presented and the effects of sub-cellular fluctuations on the population wide dynamics are considered. The final chapter of the thesis (Chapter 8) includes a summary of the different results presented along with some concluding remarks.

# Chapter 2

# Molecular Biology of Gene Expression

Gene expression is a vital life process through which genetic information is transformed into functional and structural molecules. The aim of this Chapter is to give a brief overview of the process: introducing the reader to the key steps and the major players involved, and highlighting its vital role for cell behaviour and fate. Special attention is given to DNA transcription – the first step of gene expression – and in particular to new knowledge regarding this process gained from single molecule experiments. The new, dynamical picture of DNA transcription revealed by such experiments facilitates, for the first time, the development of quantitative and predictive models of the process. Such models will be the subject of the following chapters.

## 2.1 Gene Expression

DNA (deoxyribonucleic acid) is the molecule of life; it contains the *genetic information* that defines every living organism. From the rod-like shape of *Escherichia coli* cells to complex human bodies and from bacterial chemotaxis to the sexual preferences of peahens, characteristics or even behaviours have a basis on pieces of information stored in the DNA, called *genes*. Species perpetuate and evolve as this genetic information is replicated and passed down to next generations. Moreover, during the lifetime of an individual this information is constantly accessed and cell constituents are produced from it, namely RNA (ribonucleic acid) and proteins. Complex interactions between these

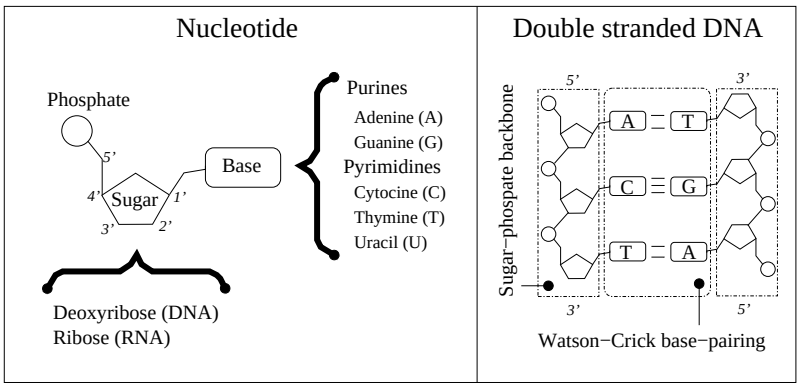



Figure 2.1: Simplified illustration of the nucleotide and DNA structure.

molecules and DNA give rise to the complex cellular behaviour we perceive.

In the remainder of this section we review some well established biological facts regarding how genetic information is stored, transformed and managed. This review is meant to provide the biological context for the work presented in the following chapters.

### 2.1.1 DNA Structure

DNA is a polymer made up of simple units called *nucleotides*. Each of these small monomers ($\sim 3.4$Å), consists of three parts (see Fig. 2.1):

1. a core sugar made up of five carbons (pentose)
2. a *base* attached to the $2$-carbon of the sugar
3. a phosphate group attached to the $5$-carbon of the sugar

As the term deoxyribonucleic acid suggests, nucleotides that make up DNA carry the sugar deoxyribose. Additionally, they can be loaded with four different bases: *adenine* (A), *cytocine* (C) *guanine* (G) and *thymine* (T). Physically, genetic information is stored in the sequence of these four types of nucleotides along the DNA. Two nucleotides are linked together via bonds that are created between the phosphate group of the first one and the $3$-carbon site of the second. Successive sugars on the DNA are, therefore, linked via phosphodiester bonds between their $5$ and $3$-carbon sites. Owning to the asymmetric structure of the nucleotides and the resulting asymmetry in their bonding, DNA is endowed with directionality. Customarily, the notation $3'$ and $5'$ is used to denote the ends of a DNA chain with regard to which carbon site is free at the terminal nucleotide.

Within cells, DNA usually occurs in a stable, double-stranded form (dsDNA), which when relaxed attains the familiar double helical structure [151]. The two strands run on

different directions and are linked to each other through what is known as *complementary* or *Watson-Crick base-pairing*. Bases come in two types (i) purines, consisting of A and G and (ii) pyrimidines, consisting of C and T. Hydrogen bonds can be formed between purines and pyrimidines: A binds to T via 2 hydrogen bonds and G binds to C via 3 hydrogen bonds (see Fig. 2.1). In other words, A is complementary to T as G is to C, and DNA strands with complementary sequences can base-pair with one another forming dsDNA.

### 2.1.2 From DNA to RNA

The stability of DNA makes it ideal as a long-term storage device for genetic information. However, for cells to function, pieces of the genetic information, customarily called *genes*, [1] must be *expressed* – transformed into protein molecules that carry out specialised functions. DNA transcription refers to the initial step of gene expression, where genetic information is read from DNA and copied onto RNA.

RNA molecules, similar to DNA, are a polymeric chains made up of four nucleotides. Nucleotides that comprise RNA are, however, slightly different form those used in DNA. The first difference lies in the sugar component, where ribose is is used instead of deoxyribose (hence the name ribonucleic acid). Furthermore, RNA nucleotides use a slightly different set of bases, namely A, C, G, U. Here U stands for the base *uracil* which is the RNA analog of T (thymine).

An RNA chain that has been transcribed from a gene on the DNA is referred to as messenger RNA (mRNA). Messenger RNAs are subsequently used as templates for proteins synthesis. At this step, dubbed *translation*, the genetic code conveyed by the mRNA is finally decoded a protein – a sequence of amino acids.[2] Proteins constitute the functional and structural elements of cells, participating for example in various reactions as catalysts (including DNA and RNA synthesis) or as building blocks in various cellular structures (*e.g.*, cytoskeleton).

Unlike mRNAs, certain classes of RNA molecules transcribed from the DNA are not used as templates for protein synthesis (non-coding RNAs). Being single stranded, RNA is a rather flexible molecule, which can fold in a sequence dependent manner forming various distinctive structures (*e.g.*, RNA hairpins) [3]. These structures can in some cases recognise other molecules and participate in various catalytic reactions, hence enabling RNA molecules to play various functional roles within the cell [152]. For example, tRNAs and rRNA are two classes of functional RNA molecules that participate in translation

[1] see Ref. [110] for a detailed discussion on the definition of the gene.

[2] Every three nucleotides in the sequence of the mRNA map to an amino acid in the protein sequence.

and are constantly expressed from the DNA. Finally, it has been appreciated that various small RNA molecules can have regulatory functions, dictating which of the genes get to be expressed [78–81].

The crucial role RNA molecules play in cell function places DNA transcription among the most vital life processes. The microscopic dynamics of the process will be the subject of Chapter 4 and in Chapter 6 will shall focus on how these dynamics affect RNA production.

### 2.1.3 RNA Polymerase

Across all domains of life, transcription is carried out by specialised enzymes known as *RNA polymerases* (RNAPs). These remarkable enzymes slide along the DNA producing RNA. To do so, they possess an impressive repertoire of functions. Initially RNAP binds to DNA and unwinds (melts) the double helix. Subsequently, the RNAP moves along the DNA in a stepwise fashion, using the one strand of the DNA as a template for the production of the RNA chain. At each step the RNAP selects the RNA nucleotide that base-pairs with the corresponding DNA nucleotide, and catalyses the creation of the phosphodiester bond linking the nucleotide to the rest of the RNA chain. An additional important feature of the RNAP is its ability to catalyse the cleavage of the RNA chain (*nucleolytic* activity). As we will see in more detail in Chapter 5 such a function is crucial for the correction of errors (misincorporated nucleotides) that occur due to thermal fluctuations.

### 2.1.4 Orchestrating the Code

To keep pace with environmental changes and synchronise its internal workings the cell must be able to control the timing and levels of gene expression. This vital ability, referred to as *gene regulation*, constitutes the very essence of cellular behaviour and fate.

In the 1960s seminal work by Jacob and Monod [93] showed that the process of DNA transcription of specific genes can be turned on and off in response to environmental stimuli. More recently it has been appreciated that this mechanism, known as *transcriptional regulation*, is just one of the many that cells use to the modulate the expression of their genes. In fact within cells, proteins, RNA molecules and genes form complex networks of interactions. As we will see in more detail in Chapter 7, such networks produce non-trivial genetic behaviour at the cellular level. In this manner, for example, different cell types of multicellular organisms can demonstrate different physiology and functionality despite the fact that they all share the same genetic information. Similarly, bacteria can switch

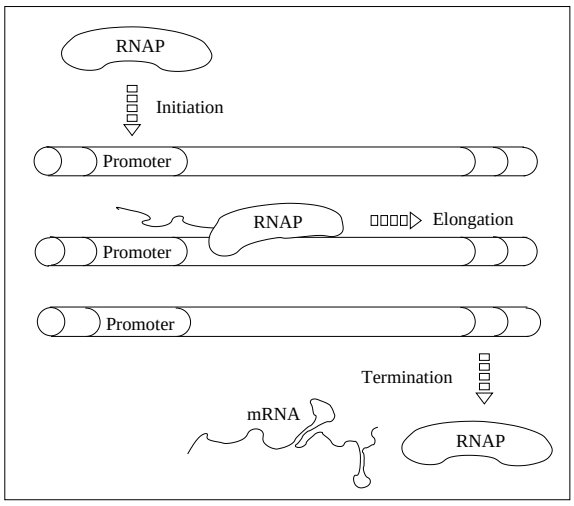



Figure 2.2: Simplified illustration of the transcription cycle.

between alternate genetic programs enabling them to survive under a wide spectrum of environmental conditions.

## 2.2 Dissecting DNA Transcription

DNA transcription is a rather intricate process. Several key events are involved some of which we are only beginning to understand in fine detail [119]. To this end single molecule techniques as well as crystallography proved to be powerfull tools, enabling the study of transcription at an unprecedented scale [31, 33, 63]. In this section we briefly review knowledge of the process that we have gained from such experimental studies. Such knowledge facilitates and motivates the development of predictive models of transcription, which will be the subject of subsequent chapters.

Like every life process, DNA transcription is subject to the laws of evolution. With this in mind, it should be noted that differences exist in the actual process between the different domains of life [3]. However, the vital role of DNA transcription for life is exemplified by the conservation of the core process across all organisms. In this respect, the overview presented below is meant to be as general as possible, focusing on our knowledge from bacterial transcription and pointing out similarities and differences with eukaryotic transcription.

On a crude level, the process of DNA transcription can be broken up into three main phases: (see Fig. 2.2)

1. initiation,

2. elongation,

3. termination.

In the initiation phase the RNAP recognises and binds to specific DNA sequences, which mark the beginning of genes. During the subsequent phase of transcription elongation the enzyme translocates along DNA using the $3' \rightarrow 5'$ strand as a template for the polymerisation of the RNA chain. Finally, sites of transcriptional termination cause disassociation of the RNAP from the DNA and the release of the transcript. In the remainder of this section for the sake of completeness we consider all three stages. Special emphasis is given, however, to the elongation phase that is the major subject of study in our work.

### 2.2.1 Initiation

The initiation phase involves loading of the RNAP onto the DNA template and the subsequent transcription of the first few nucleotides [58]. To accomplish the former, the RNAP is capable of binding to specific DNA sites, dubbed *promoters*.[3] Physically, within cells, DNA occupies some $3$D volume most often in a highly condensed form. Hence, finding the right place to bind is a non-trivial problem. It has been proposed, that the dimensionality of the promoter search problem is reduced by a combination of $1$D and $3$D diffusion [145]; the RNAP scans for promoters by binding weakly and sliding along non-specific DNA and occasionally jumps between distant DNA segments. Such a mechanism explains the rapid promoter binding, which can be as fast as a few seconds [15].

The initial loading of the RNAP on the DNA is a major step of transcriptional regulation across all domains of life. Specific proteins, known as *transcription factors* (TFs), can assist or hinder the binding of RNAP on the DNA, either through direct interactions with the enzyme or indirectly by exposing or hiding DNA promoter sequences [112]. In this manner the expression of genes is tuned in response to various cues through the action of one or more TFs. In addition "master" TFs, having under their control a large number of genes, add higher layers of genetic regulation.

Once bound and properly positioned on the DNA, the RNAP unwinds the double helix, uncovering the template strand to be transcribed. Then, the RNAP attempts to initiate the processive elongation of the transcript through a process known as *abortive initiation* [58]. During this stage the initial fragment of the DNA template is repeatedly transcribed and cleaved, owing to the inability of the RNAP to efficiently disassociate from the promoter and proceed further downstream [73]. The eventual clearance of the

[3] In eukaryotes promoter binding is mediated by accessory proteins [3]

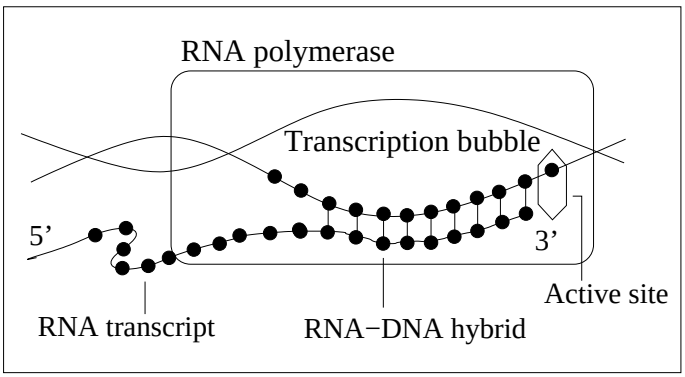



Figure 2.3: Schematic illustration of transcription elongation complex (TEC), consisting of the RNA polymerase, a region of melted DNA (transcription bubble) and the RNA-DNA hybrid. Polymerisation of the nascent RNA is catalysed by the active site of the polymerase.

promoter results in the formation of a stable complex, know as transcription elongation complex, and which signals the entrance into the elongation phase [58].

### 2.2.2 Elongation

During the elongation phase the RNAP slides along the DNA template polymerising the transcript at a rate of $30 - 100$ nucleotides/sec. However, processive RNA synthesis is often disrupted by specific DNA sequences; lesions or roadblocks present in the DNA; nucleotide misincorporation events; and proteins that regulate RNAP function. Recently *in-vivo* and *in-vitro* experimental studies have demonstrated the prevalence of these pauses during DNA transcription and highlighted their possible biological significance [32, 47, 53, 65, 95, 99, 124, 136, 149].

Of particular relevance to our work is the dynamical picture of the elongation phase uncovered by single molecule manipulation experiments. These studies provided a more thorough understanding of how the RNAP motors along the DNA producing the RNA transcript [1, 65]. They also observed frequent pausing by the RNAP and shed light on some of the mechanisms inducing these pauses [47, 99, 124]. Below we briefly review some key experimental findings.

#### Transcription Elongation Complex

As the elongation phase is entered the RNAP forms a stable complex along with the DNA and the RNA transcript. This complex is known as the *transcription elongation complex* (TEC). The TEC covers a region of approximately 25 DNA base pairs (bp), the central

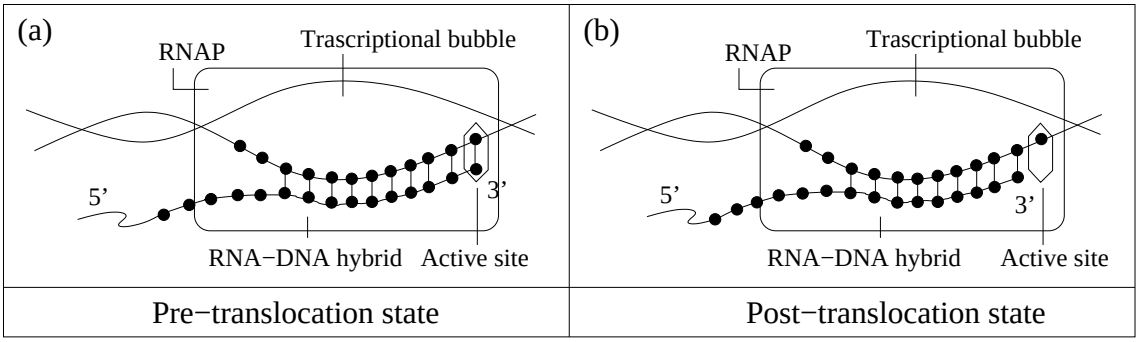



Figure 2.4: Schematic illustration of the pre- and post-translocation state of the transcription elongation complex (TEC). The pre-translocation state corresponds to the state immediately after the polymerisation of a nucleotide. The post-translocation state corresponds to the state after the forward translocation of the TEC and before polymerisation of the next nucleotide takes place.

part ($12$ bp) of which is "melted", [4] forming the *transcription bubble* [76]. Within the bubble a double stranded helix (approximately $8 - 9$ bp long) is formed between the nascent RNA and the DNA template. This structure is know as the *RNA-DNA hybrid*. The RNA-DNA hybrid as well as nonspecific interactions between the RNAP, the DNA and the RNA are the major contributors to the stability of the complex [104]. Upstream of the RNA-DNA hybrid, the RNA chain exits the complex through the *RNA exit channel* of the polymerase. Free nucleotides (NTP) enter the complex through the *secondary channel* of the RNAP and are polymerised at the $3'$ end of the transcript by the *active site* of the RNAP. A schematic illustration of the TEC is given in Fig. 2.3

### Single Nucleotide Addition Cycle

The elongation of the RNA transcript is accomplished through the *polymerase* and *helicase* capabilities of the RNAP. The former corresponds to the ability of the RNAP to catalyse the addition of nucleotides at the $3'$ end of the RNA chain, while the latter to the ability of translocating along the DNA template while unwinding the double helix.

The two activities operate in tandem, so that each polymerisation event is closely followed by the forward *translocation* of the TEC by one nucleotide. Experimental evidence suggests that the two steps are not energetically coupled, *i.e.*, no energy exerted during the polymerisation step is utilised for translocation [1]. Rather, the TEC is behaving like a thermal ratchet with forward translocation driven solely by diffusion. Schematically, the

[4] The two DNA stands are separated.

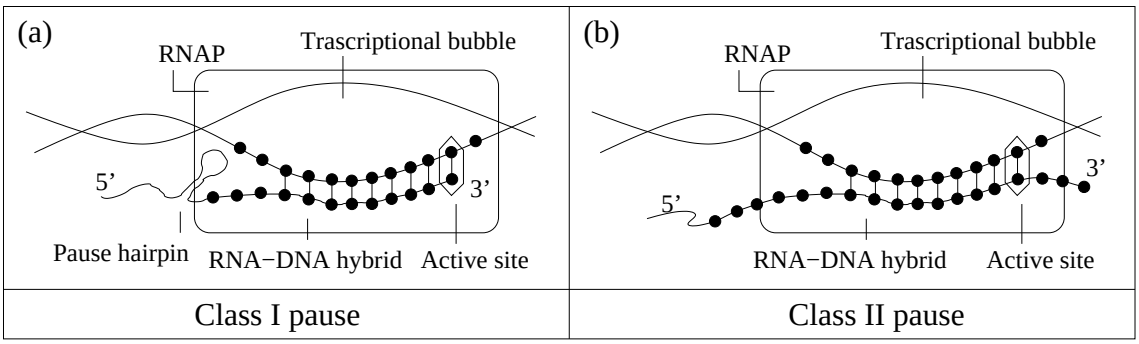



Figure 2.5: Schematic illustration of class I and class II transcriptional pauses. Class I pauses are induced by RNA hairpins that interact with the RNAP, while class II pauses involve the backward translocation of TEC along the DNA, a phenomenon dubbed as backtracking.

single nucleotide addition cycle is captured by

$$\mathrm{TEC}_{n,}\mathrm{pre} \;\rightleftarrows\; \mathrm{TEC}_{n,}\mathrm{post} \;\xrightarrow{+NTP}\; \mathrm{TEC}_{n+1,}\mathrm{pre} + \mathrm{PP}_{i}. \tag{2.1}$$

where $\mathrm{TEC}_{n,}\mathrm{pre}$ and $\mathrm{TEC}_{n,}\mathrm{post}$ correspond to the *pre-* and *post-translocated* state of the complex; before and after the translocation step has been achieved and prior to the polymerisation of the next nucleotide ($n \rightarrow n+1$) (see Fig. 2.4).

### Transcriptional Pausing

Often, the processive polymerisation of the RNA transcript is disrupted by pauses of the RNAP, a phenomenon dubbed *transcriptional pausing*. Still a hot subject of biological research, transcriptional pausing has been endowed with a wide variety of roles. For example, it has has been proposed that transcriptional pausing can assist the recruitment of regulatory proteins to the TEC [100]; function as a precursor to transcriptional termination [94]; and play a role in transcriptional error correction [124, 147].

Early biochemical assays with bacterial RNAP focused on the identification of pauses induced by specific DNA sequences. In particular, these studies revealed two distinct classes of DNA signals that give rise to transcriptional pausing [7] (see Fig. 2.5). The major difference between the two classes of signals lies in the mechanistic details through which pausing is induced: Class I signals encode for RNA hairpins that interact with the RNAP, blocking its movement, whereas class II signals result in repositioning of the active site and the backward translocation of the RNAP on the DNA template (*backtracking*). Additionally, it was shown that specific proteins (*e.g.*, NusA, NusG and GreA) help the

RNAP recover such pauses, pointing to a novel mode of transcriptional regulation [7].

A more thorough investigation of transcriptional pausing came with advancements in single molecule manipulation techniques. Such studies with bacterial RNAP reported a wide distribution of pauses, ranging from a few seconds up to several minutes [46,99,124]. In particular, Shaevitz *et al.* [124] directly observed that particularly long pauses ($> 25$ s) were induced by backtracking of the RNAP with a frequency of $1$ pause/kbp. More recently, backtracking and the wide temporal distribution of pauses were also observed for the case of eukaryotic transcription [47].

Shorter pauses ($1 - 6$ s) were found to be insensitive to hindering or assisting loads acting on the RNAP [99]. This observation suggested that the pauses did not involve any translocation of TEC whatsoever. Rather, it was proposed that they form a separate class of elemental pauses, termed *ubiquitous pauses* [99]. Such pauses seem to occur due to small conformational changes of the RNAP molecule, which are induced by DNA sequences with specific characteristics [64].

**RNA Polymerase Backtracking**

Backtracking is a major player in transcriptional pausing [47, 124]. At each template position, backtracking constitutes an alternative reaction pathway that is in kinetic competition with polymerisation [58]. Entrance into this pathway is particularly favoured in the presence of a weak RNA-DNA hybrid, such as in the case of a misincorporated nucleotide [124] or when strong forces are exerted on the polymerase while it transcribes the DNA [47, 124].

During backtracking the RNAP freely diffuses back and forth along the DNA template [58]. In particular, the backward translocation of the RNAP causes the $3'$ end of the transcript to break loose from the RNA-DNA hybrid and move out of the complex (through the secondary channel) while the two DNA strands are rejoined. Similarly, at the $5'$ end of the transcription bubble dsDNA is re-opened and part of the RNA transcript is moved inside the complex (through the RNA-exit channel) where it becomes rehybridised with DNA. Once backtracked, the TEC can presumably slide back and forth until it retains its polymerisation-competent state, with the $3'$ end of the transcript positioned in the active site.

In general, during backtracking the TEC can move as far as $8-9$ nucleotides from the transcriptional starting point. Moving past this point is thermodynamically unfavourable since it would result in shortening of the RNA-DNA hybrid and destabilisation of the complex. Such extensive backtracks, however, are thought to be precluded mainly due to structural elements (*e.g.*, hairpins) of the transcript that interact with the TEC [58].

### 2.2.3 Termination

Termination corresponds to the disassociation of the RNAP from the DNA and the release of the RNA transcript. In bacterial transcription, termination is usually marked by specific sequences, termed *intrinsic terminators*; they code for an RNA hairpin structure followed a U-rich sequence. Such sequences destabilise the TEC and causes the transcript to be released [146]. Regulated termination, mediated by specific proteins, is also widespread. For example, the *Rho* factor binds to specific sites on the nascent RNA and slides along it towards the TEC causing it to terminate transcription. On the contrary, the *Mfd* factor does not recognise any particular sequence but directly interacts with paused TECs causing them to collapse. This last case of regulated terminations exemplifies the role of transcriptional pausing in regulating the process of transcription.

## 2.3 Summary

In this Chapter we presented a brief review of the biology that has motivated the work presented in the subsequent chapters. Some elements of molecular biology regarding how genetic information is stored and expressed were presented along with a more detailed description of DNA transcription, the process through which the genetic information stored in the DNA is copied in RNA.

Recent advancements in experimental techniques have enabled the study of transcription at the single molecule level [47, 124, 147]. This unprecedented level of detail has highlighted interesting phenomena, such as transcriptional pausing, with important biological implications for the regulation of the process and therefore the functionality of the cell. Moreover, it facilitates the development quantitative models that can explain existing data and make quantitative predictions regarding the process. Such models will be the main subject of the chapters to follow.

# Chapter 3

# Theoretical Background

In this Chapter we give an overview of the mathematical and computational tools that will be used throughout this thesis. In particular, it includes a brief introduction to probability theory and stochastic processes. Our main aim is to enable the non-expert reader to understand key concepts that will be used in subsequent Chapters without being referred to the vast literature. For the sake of brevity, rigorous derivations and technical details are skipped and in this respect the material presented should be considered as a catalogue of key concepts, facts and notations that will be used later on.

## 3.1 Elements of Probability Theory

In everyday life, we all have a rather intuitive understanding of what probability is: it merely quantifies our expectations of how likely it is for a certain event to occur. Imagine, for example, a not so serious gambler stepping into a casino in Monte Carlo and placing all his money at a roulette table on $18$ red. Before the croupier spins the wheel, we would expect that the odds of our friend winning are $1/37$. Our intuition is based on the assumption (and trust in the casino owners) that all $37$ possible events are equally likely. This simple example allows us to sketch how probability is formulated on solid mathematical grounds. For more rigorous definition however the reader is referred to any advanced textbook on probability theory (*e.g.*, see Refs. [44, 106]).

### 3.1.1 Basic Concepts and Notation

A central concept in probability theory is that of a *stochastic* or *random variable* $X$. Stochastic variables are used to describe real-world observations or the outcome of certain actions such as spinning the roulette wheel or throwing a die. They can also be multidimensional objects, in which case they are conveniently thought of vectors $\mathbf{X}$, such as the position vector of a small particle suspended in water (*Brownian particle*). Depending on the system at hand, $X$ can attain certain values (or states) $x$ that constitute a set, customarily denoted by $\Omega$ and called *sample space* or *set of states*. For example in the case of a roulette consists of all possible outcomes, i.e., $0, 1R, 2B, \ldots 36R$. A function, called the *distribution function*, is then defined over $\Omega$ mapping to every subset $A$ of $\Omega$ a real-valued number, representing the probability that $X$ attains a value within $A$. To satisfy our intuition that probability is non-negative and must always sum to $1$ one would have to impose certain restrictions on the choice of the distribution function.

When $\Omega$ consists of discrete values (states) the distribution function is

$$P_X(x) \quad = \quad \text{Prob}(X = x), \tag{3.1}$$

subject to the conditions

$$\begin{array}{lrcl} \text{(i)} & P_X(x) & \geq & 0, \\ \text{(ii)} & \displaystyle\sum_x P_X(x) & = & 1. \end{array}$$

Establishing $\Omega$ and $P_X$ is the key step for any practical application of probability theory. In every case, these are constructed based on prior knowledge and intuition as well as on physical consideration of the specific problem at hand. In this sense, the term *a priori probabilities* is often used for $P_X$ to stress the fact that in most case $P_X$ is just assumed and therefore subject to experimental validation [141].

When the values $x$ form a continuous range, $P_X(x)$ is used to denote the *probability density function* (PDF). Then the probability $X$ attains a value between $x$ and $x + dt$ is

$$\text{Prob}(x \leq X \leq +dt) \quad = \quad P_X(x)dx. \tag{3.2}$$

One immediately sees that this probability goes to zero as $dt \rightarrow 0$. Therefore, the probability that $X$ has exactly the value $x$ is zero. A way around this is to make use of *Dirac delta function* defined as

$$\delta(x) \quad = \quad \begin{cases} \infty, & x = 0 \\ 0, & x \neq 0 \end{cases} \tag{3.3}$$

subject to the additional constrain

$$\int_{-\infty}^{+\infty} \delta(x)dx \;=\; 1. \tag{3.4}$$

For example, $P_X(x) = \delta(x - x_a)$ defined over $(-\infty, +\infty)$ states that the probability of observing any value other that $x_a$ is zero. In fact, one can use delta functions to rewrite any discrete probability distribution $\Pi_N(n)$ in terms of a probability density function $P_X(x)$. In particular, one has

$$P_X(x) \;=\; \sum_n \Pi_N(n)\delta(x - x_n), \tag{3.5}$$

where the discrete stochastic variable $N$ is mapped into a a continuous one, $X$ and the discrete values $n$ are mapped to set of points $x_n$ embedded in a continuous interval. Having said that, in what follows we focus on continuous stochastic variables.

### 3.1.2 Moments

The *mean value* of a stochastic variable $X$ is given by

$$\langle X \rangle \;=\; \int x P_X(x)dx. \tag{3.6}$$

More generally, one can define the average of any function $f(x)$ as

$$\langle f(X) \rangle \;=\; \int f(x) P_X(x)dx. \tag{3.7}$$

One is particularly interested in the quantities

$$\mu'_m \;\equiv\; \langle X^m \rangle, \tag{3.8a}$$

$$\mu'_m \;\equiv\; \langle (X - \mu_1)^m \rangle \tag{3.8b}$$

which are called the *raw* and *central moments* of the distribution respectively.

### 3.1.3 Other Important Functions

In addition to the PDF $P_X(x)$, a few other functions are also of key importance in probability theory. In particular, the *cumulative distribution function* (CDF), gives the total

probability of $X$ attaining a value less than or equal to $x$, *i.e.*,

$$\text{Prob}(X \leq x) \;=\; F_X(x) \;=\; \int_{-\infty}^{x} P_X(x')dx' \tag{3.9}$$

In fact, for real-valued stochastic variables, first the CDF is defined over the state space and subsequently the PDF is obtained as the derivative of the CDF,

$$P_X(x) \;=\; F'_X(x). \tag{3.10}$$

That is why the name probability function is often used (especially in the mathematical literature) instead of CDF. However, since "probability function" has also been used for the discrete version of $P_X(x)$, we shall stick to the term CDF term to avoid confusion.

A closely related function is the *survival function*, $S_X(x)$, describing the probability of the stochastic variable $X$ to attain a value greater than $x$. One readily obtains

$$\begin{aligned} S_X(x) + F_X(x) &= 1 \Rightarrow \\ F'_X(x) &= -S'_X(x) \Rightarrow \\ P_X(x) &= -S'_X(x), \end{aligned} \tag{3.11}$$

which relates the PDF of a stochastic variable to its survival function.

The *characteristic function* (CF) is yet another alternative description to the PDF (see [106]). The CF, $G_X(k)$, of real-valued stochastic variable $X$ is defined as the Fourier transform of the PDF:

$$G_X(k) \;=\; \int_{-\infty}^{+\infty} e^{ikx} P_X(x)dx \;=\; \langle e^{ikX} \rangle. \tag{3.12}$$

$G_X(k)$ also allows us to illustrate the notion of a *moment generating function* (MGF). In particular, $G_X(k)$ encodes all raw moments in the coefficients of its Taylor expansion in $k$:

$$\begin{aligned} G_X(k) &= \int_{-\infty}^{+\infty} e^{ikx} P_X(x)dx \\ &= \int_{-\infty}^{+\infty} \left[1 + ikx - \frac{(kx)^2}{2} + \ldots\right] P_X(x)dx \\ &= 1 + ik\mu'_1 - \frac{k^2}{2}\mu'_2 + \ldots \\ &= \sum_{m=0}^{\infty} \frac{(ik)^m}{m!}\mu'_m. \end{aligned} \tag{3.13}$$

Finally, alternative MGFs can be constructed, using for example $\langle e^{sX} \rangle$,$\langle e^{-sX} \rangle$ and $\langle z^X \rangle$.

These different formulations of the MGF offer certain advantages depending on the range over which $X$ is defined [141].

### 3.1.4 Multivariate Distributions

As noted above, the notion of a random variable can also be generalised to $n$-dimensions by regarding a vector $\mathbf{X}$ consisting of $n$ components $X_1, X_2 \ldots, X_n$. Here, we catalogue some special density functions that are relevant to this case. For the sake of brevity we restrict ourselves to the two-dimensional case, noting that results can readily be generalised to more dimensions.

Let $\mathbf{X} = (X_1, X_2)$ be a two component stochastic variable. The probability that $X_1$ has a value between $x_1$ and $x_1 + dx_1$ and that $X_2$ a value between $x_2$ and $x_2 + dx_2$ is given by:

$$P_{\mathbf{X}}(x_1, x_2)dx_1dx_2. \tag{3.14}$$

$P_{\mathbf{X}}(x_1, x_2)$ denotes the PDF of the composite variable $\mathbf{X}$ or the *joint probability density function* of the two variables $X_1$ and $X_2$ and is subject to the normalisation condition:

$$\int P_{\mathbf{X}}(x_1, y_2)dx_1dx_2 \;=\; 1. \tag{3.15}$$

The *marginal probability density functions* are concerned with each stochastic variable regardless the value of the other one. For example, the marginal PDF of $X_1$ can be obtained from the joint PDF as

$$P_{X_1}(x_1) \;=\; \int P_{\mathbf{X}}(x_1, x_2)dx_2 \tag{3.16}$$

One can now consider the distribution of one variable given that the other variable has some fixed value . For example, the *conditional probability density function* of $X_1$ conditional on $X_2$ having the value $x_2$ is denoted by

$$P_{X_1|X_2}(x_1|x_2). \tag{3.17}$$

According to *Bayes' rule* the conditional PDF can be written as

$$P_{X_1|X_2}(x_1|x_2) \;=\; \frac{P_{X_1,X_2}(x_1, x_2)}{P_{X_2}(x_2)} \tag{3.18}$$

A final point is that of statistical independence. Two stochastic variables are said to be statistically independent if their joint PDF can be factorised into the product of the

marginal ones, viz.

$$P_{X_1,X_2}(x_1,x_2) \;=\; P_{X_1}(x_1)P_{X_2}(x_2). \tag{3.19}$$

Consequently, the conditional PDF becomes

$$P_{X_1|X_2}(x_1|x_2) \;=\; P_{X_1}(x_1). \tag{3.20}$$

## 3.2 Stochastic Processes

Our lucky friend steps out of the casino with his winnings, and is challenged by a stranger into a game involving tossing a coin. He is promised that each time he tosses a head he will win an amount of money which he will lose in case of a tail. Feeling lucky he accepts. The capital of our friend constitutes a *stochastic process*, that is, at any time his capital will depend on the number of tosses he has made so far. In particular, his capital after each toss depends on his capital prior to the toss and the random outcome of the toss. The process is a truly *Markovian* one.

### 3.2.1 Basic Concepts and Definitions

Following Ref. [141], in mathematical terms a stochastic processes $Y$ can be described by as time-dependent stochastic variable. Therefore, one can assume a hierarchy of joint PDFs,

$$P_n(y_1,t_1;\ldots;y_n,t_n), \tag{3.21}$$

that $Y$ attains the values $y_1, y_2 \ldots y_n$ at times $t_1, t_2, \ldots, t_n$, respectively. The definition of $P_n$ should be independent of the ordering of times and moreover one must require that

$$\int P_1(y_1,t_1)dy_1 \;=\; 1, \tag{3.22a}$$

$$\int P_n(y_1,t_1;\ldots;y_n,t_n)dy_n \;=\; P_n(y_1,t_1;\ldots;y_{n-1},t_{n-1}). \tag{3.22b}$$

Under these conditions the infinite hierarchy of $P_n$ $(n = 1, 2, \ldots)$ completely specifies the stochastic process [141]. In particular it enables one to compute averages as

$$\langle Y(t_1)\cdots Y(t_n)\rangle \;=\; \int y_1 \ldots y_n P_n(y_1,t_1;\ldots;y_n,t_n)dy_1\ldots dy_n. \tag{3.23}$$

Moreover, one can define the conditional PDFs in terms of $P_n$

$$P(y_1, t_1; \ldots; y_m; t_m | y'_1, t'_1; \ldots; y'_l; t'_l) \quad = \quad \frac{P_{m+l}(y_1, t_1; \ldots; y_m; t_m | y'_1, t'_1; \ldots; y'_l; t'_l)}{P_l(y'_1, t'_1; \ldots; y'_l; t'_l)}, \tag{3.24}$$

that is the PDF of $Y$ at times $t_i, (i = 1 \ldots m)$ having fixed the values of $Y$ at time $t'_i, (i = 1 \ldots l)$.

A key concept in that of a *stationary stochastic process*. These are processes whose statistical properties do not depend on time. One can express this mathematically, by allowing the hierarchy of $P_n$ to be unaffected by an arbitrary shift in time $\tau$:

$$P_n(y_1, t_1; \ldots; y_n, t_n) \quad = \quad P_n(y_1, t_1 + \tau; \ldots; y_n, t_n + \tau). \tag{3.25}$$

One can see that such a condition is met if $P_1$ is independent of time and all other $P_n$ depend solely on time differences $t_2 - t_1$, $t_3 - t_2$, etc.

The simplest case of a stochastic process, occurs when the value of $Y$ at different times are statistically independent to each other. Take, for example, the process defined by successively tossing of a die. The result of each toss is independent of any previous one. In the case of independence, $P_1$ suffice to describe the stochastic process since the hierarchy $P_n$ can be expressed as the product

$$P_n(y_1, t_1; \ldots; y_n, t_n) = \prod_{i=1}^{n} P_1(y_1, t_1). \tag{3.26}$$

Moreover, if $P_1$ is also independent of time (as in the case of tossing a die) the process is stationary. The next simplest case is known as a *Markov process*, in which the future is determined solely by the present.

### 3.2.2 Markov Processes

A Markov process, named after the Russian mathematician Andrei Andreyevich Markov, is a stochastic process in which the state at any time depends solely on the state in the immediate past and not on previous history. Mathematically the Markov property of a stochastic process $Y$ is formulated in terms of conditional PDFs, stating that for any set of successive time points $(t_1 < t_2 < \cdots < t_n)$ one has

$$P(y_n, t_n | y_1, t_t; \ldots; y_{n-1}, t_{n-1}) \quad = \quad P(y_n, t_n | y_{n-1}, t_{n-1}). \tag{3.27}$$

This property enables one to fully describe the evolution of the Markov process using the one-step conditional PDF and the PDF for the initial observation at $t_1$. In particular, the hierarchy $P_n$ can be expressed as:

$$\begin{aligned} P_n(y_1,t_1;y_2,t_2;\ldots;y_{n-1},t_{n-1};y_n,t_n) = \\ &= P_{n-1}(y_1,t_1;y_2,t_2;\ldots;y_{n-1},t_{n-1})P(y_n,t_n|y_1,t_1;y_2,t_2;\ldots;y_{n-1},t_{n-1}) \\ &= P_{n-1}(y_1,t_1;y_2,t_2;\ldots;y_{n-1},t_{n-1})P(y_n,t_n|y_{n-1},t_{n-1}) \\ &= \ldots \\ &= P_1(y_1,t_1)P(y_2,t_2|y_1,t_1)P(y_3,t_3|y_2,t_2)\cdots P(y_n,t_n|y_{n-1},t_{n-1}). \end{aligned} \tag{3.28}$$

provided the time ordering $t_1 < t_2 < t_3 < \cdots < t_{n-1} < t_n$.

### 3.2.3 Brownian Motion

The capital of our friend after successive tosses of the coin constitutes a *truly* Markovian process. For many physical systems, however, the Markovian property is merely an assumption made possible by the coarseness of our observations or description of the system [141]. To illustrate this point one usually appeals to the seminal paradigm of Brownian motion [50, 141]. This is the first application of a Markovian stochastic process for describing a physical phenomenon, in particular the motion of a light particle immersed in water. The motion of a Brownian particle is mainly driven by collisions with surrounding water molecules. The large number of water molecules and collisions occurring is prohibitive for a complete description of the system, however it allow for a statistical treatment. In particular, Einstein showed that the position of a Brownian particle at successive time points $t_0, t_0+\tau, t_0+2\tau, \ldots$ could accurately be captured by a Markovian process. However, this is merely allowed by the coarseness of our observations Choosing a sufficiently large time interval $\tau$ ensures a large number of collisions so that the net displacement of the particle will appear uncorrelated at different times. On this coarse-grained time-scale the process can be regarded as Markovian. Similarly, for most practical applications one seeks an appropriate time-scale $\tau$, such that changes of the system during $[t, t+\tau]$ depend on the state of the system at $t$ but not on any prior times.

### 3.2.4 The Master Equation

The Markovian property (or assumption) enables the characterisation of a stochastic process by means of a differential equation, most commonly known as the *Master equation*.

Below we present a sketch for deriving the Master equation stressing some crucial points and the implicit assumptions made when one directly writes down the Master equation for a system. For a more detailed treatment, however, the reader is referred to the literature (*e.g.*, see Refs. [50, 141]).

From the Markovian property follows that

$$P_3(y_1,t_1;y_2,t_2;y_3,t_3) \;=\; P_1(y_1,t_1)P(y_2,t_2|y_1,t_1)P(y_3,t_3|y_2,t_2). \tag{3.29}$$

Integrating over $y_2$ and dividing with $P(y_1,t_1)$ one obtains

$$P(y_3,t_3|y_1,t_1) \;=\; \int P(y_2,t_2|y_1,t_1)P(y_3,t_3|y_2,t_2)dy_2. \tag{3.30}$$

Equation 3.30 is called the *Chapman-Kolmogorov equation* and imposes a functional relationship between the conditional probabilities $P(y_i,t_i|y_j,t_j)$. The Master equation is a reformulation of the Chapman-Kolmogorov equation obtained in the limit of vanishingly small time differences, $t_3 - t_2 = \tau \to 0$.

To proceed any further, one has to write

$$\lim_{\Delta t\to 0}\frac{P(x,t+\Delta t|y,t)}{\Delta t} \;=\; W(x|y), \tag{3.31}$$

$W(x|y)$ assumes that the probability per unit time for a transition to occur between from state $x$ to state $y$ depends solely on theses states and is independent of time. This enables one to write $P(y_3,t_2+\tau|y_2,t_2)$ as

$$P(y_3,t_2+\tau|y_2,t_2) = \left[1-\tau\int W(y|y_2)dy\right]\delta(y_3-y_2) + \tau W(y_3|y_2) + \mathcal{O}(\tau^2). \tag{3.32}$$

Here, the first term in the above is the probability that no transition occurs during $\tau$. Substituting this in Eq. (3.30) yields:

$$\begin{aligned} P(y_3,t_2+\tau|y_1,t_1) \;=\;& P(y_3,t_2|y_1,t_1) - \tau\int W(y|y_3)dxP(y_3,t_2|y_1,t_1) \\ & +\tau\int W(y_3|y_2)P(y_2,t_2|y_1,t_1)dx_2 + \mathcal{O}(\tau^2) \end{aligned} \tag{3.33}$$

Finally, dividing by $\tau$ and letting $\tau \to 0$ one obtains the Master equation describing the

stochastic process

$$\frac{dP(y_3,t_2|y_1,t_1)}{dt_2} = \int \{W(y_3|y_2)P(y_2,t_2|y_1,t_1) - W(y_2|y_3)P(y_3,t_2|y_1,t_1)\}\, dy_2. \tag{3.34}$$

The Master equation is usually written in the simpler and more readable form,

$$\frac{dP(y,t)}{dt} = \int \{W(y|y')P(y',t) - W(y'|y)P(y,t)\}\, dy'. \tag{3.35}$$

In this form, however, it should be stressed that $P(y,t)$ does not stand for the marginal PDF $P_1(y,t)$ but for the one-step conditional PDF, *i.e.*, $P(y,t) \equiv P(y,t|y',t')$ for any choice of $y'$ and $t'$. To obtain $P_1(y,t)$ one uses the initial condition $P_1(y',0) = \delta(y'-y_0)$ to obtain

$$\begin{aligned} P_1(y,t) &= \int P(y,t|y',0)P_1(y',0)dy' \\ &= \int P(y,t|y',0)\delta(y'-y_0) \\ &= P(y,t|y_0,0) \end{aligned} \tag{3.36}$$

The Master equation can also be formulated for discrete processes, provided that one replaces the integral with a sum and interprets $P(x)$ as probability rather than probability density, *i.e.*,

$$\frac{dP(x,t)}{dt} = \sum_x \{W(x|x')P(x',t) - W(x'|x)P(x,t)\}\,. \tag{3.37}$$

The Master equation has a rather simple intuitive meaning. It describes the change in probability (or probability density) for observing any given state as the net outcome of gain and loss terms. In particular, the first term in Eq. (3.35) describes gain in the probability of observing $y$ due to transitions $y' \to y$, while the second term captures the loss due to transitions $y \to y'$.

One further remark is perhaps important at this point. In obtaining the Master equation we required the condition given by Eq. (3.31) to hold. $W(x|y)$ has units reciprocal to time and can intuitively be thought as the rate at which transitions $y \to x$ occur. As the notation implies, $W(x|y)$ does not depend on time and therefore implies that the process occurs homogeneously in time. This has important implications for the temporal dynamics of the stochastic process which remain implicit when one writes down the Master equation. To illustrate this point we consider a transition to state $y$ that occurs at time $t_0$. The

probability per unit time for any transition to occur from that point onwards is given by

$$\alpha(y) \;=\; \int W(x|y)dx. \tag{3.38}$$

Therefore, $w(t)$, the PDF for no transition to have occurred up to time $t_0 + t$ obeys

$$\begin{array}{rcll} w(t+\Delta t) & = & (1-\alpha\Delta t)w(t) & \Rightarrow \\ \dfrac{dw(t)}{dt} & = & -\alpha w(t) & \Rightarrow \\ w(t) & = & e^{-\alpha t}, & \end{array} \tag{3.39}$$

where at the last state we made use of the initial condition $w(0) = 1$. One readily sees that $w(t)$ is just the survival function of the probability density, $f(t)$, for the time to the next transition event, hence

$$f(t) \;=\; -w'(t) \;=\; \alpha e^{-\alpha t}. \tag{3.40}$$

It is clear the time needed for a transition to occur is exponentially distributed with mean $1/\alpha$. Transitions, therefore, proceed without memory and the process appear homogeneous in time. We shall return to this point when discussing methods for simulating continuous-time Markov processes.

## 3.3 One Step Processes

A special case of stochastic processes obeying the Markov property are the so called *one-step* or *birth-and-death* processes. Let us denote such a stochastic process by $N(t)$. At any time $N(t)$ attains values in the range of integers $n$ and the only permissible transitions are

$$\begin{array}{rcll} n & \rightarrow & n+1 & \text{(birth)}, \\ n & \rightarrow & n-1 & \text{(death)} \end{array}$$

These transitions occur with probabilities given by

$$P(n+1, t+dt|n,t) \;=\; g_n dt, \tag{3.41a}$$

$$P(n-1, t+dt|n,t) \;=\; r_n dt. \tag{3.41b}$$

Therefore, the total transition probability per unit time can be expressed concisely as

$$W(n'|n) \;=\; r_n \delta_{n',n-1} + g_n \delta_{n',n+1} \tag{3.42}$$

where we have introduced the *Kronecker's delta* (the discrete analog of the Dirac delta function) defined as

$$\delta_{i,j} = \begin{cases} 0 & , i \neq j \\ 1 & , i = j \end{cases} \tag{3.43}$$

Substituting $W(n'|n)$ in Eq. (3.37) one obtains the Master equation describing the one-step process:

$$\begin{aligned} \frac{dP(n,t)}{dt} &= r_{n+1}P(n+1,t) + g_{n-1}P(n-1,t) \\ &\quad -[g_n + r_n]P(n,t), \end{aligned} \tag{3.44}$$

subject to the initial condition $P(n,0) = \delta_{n,n_0}$. The first two terms on the right-hand side capture the gain in probability $P(n,t)$ due to transitions $n+1 \rightarrow n$ and $n-1 \rightarrow n$, respectively. Similarly the last term describes losses due to transitions $n \rightarrow n+1$ and $n \rightarrow n-1$.

At this point a few remarks concerning the application of one step processes to practical problems are perhaps essential. The Master equation given by Eq. 3.44 was defined over the range of all integers. However, in most cases, one-step processes with half-infinite ($n = 0, 1, \ldots$) or finite ($n = 0, 1, \ldots, N$) range suffice to capture the stochastic dynamics of real-world systems. Moreover, no specific form for $g_n$ and $r_n$ has been assumed; these rates can indeed be described by any collection of non-negative numbers. It is usually the case, however, that for most real-world applications $g_n$ are $r_n$ are given as some analytic function of variable $n$, *i.e.*,

$$r_n = r(n), \tag{3.45a}$$

$$g_n = g(n). \tag{3.45b}$$

In the simplest case $r(n)$ and $r(n)$ attain constant values for all $n$. This gives rise to, perhaps, the most well known examples of an one-step processes, the *nearest-neighbour random walks*. It turn out that in this case the Master equation can be solved completely, and an analytic form of $P(n,t)$ can be obtained [116]. If $g(n)$ and $g(n)$ are at most linear in $n$ one has a *linear one-step process*, for which the Master equation can also be solved [141]. Finally, the term *nonlinear one-step process* is reserved for processes with non-linear $g(n)$ and/or $r(n)$. Not surprisingly, time dependent solutions of the Master equation for nonlinear processes are in most cases not available.

Most often, one-step processes are used to describe the stochastic dynamics of systems consisting of a number of entities. Specific examples could be the growth of a bacterial colony, where individual bacteria duplicate and die with certain probabilities

per unit time, or the fluctuating levels of a specific protein within a cell, due to the random production and degradation of individual molecules. In such cases, a linear form of $g(n)$ and/or $r(n)$ merely states that individuals are independent of each other. This allows one to treat the random behaviour of each individual in isolation, as a separate stochastic process, and superimpose them to obtain the dynamics at the population level. Hence, the superposition principle (true for any linear system) provides the intuition for why one should expect linear one-step processes to be solvable. It also makes clear how nonlinearities introduce difficulties. In particular, non-linear terms in $g(n)$, $r(n)$ capture interactions between individuals that destroy independence and make the random history of each individual dependent on those of others. Intuitively, the system can no longer be broken up into independent components.

For most nonlinear one-step processes, therefore, one may either resort to approximation schemes and numerical methods for obtaining time dependent results or alternatively focus on the stationary distribution $P_s(n) \equiv P(n, t \to \infty)$. These topics will be subject of the following sections.

### 3.3.1 Boundary Conditions

When modelling the stochastic behaviour of a system, one often has to take into account certain physical restrictions concerning the range of values the systems' variables are allowed to take on and the behaviour of the system at the boundaries of these ranges. Take for example a population of bacteria dividing and dying or the arrivals and departures in a bank queue. Obviously, both the size of the population and the size of the queue ought to be positive at all times. However, a key difference exists between the two systems. When all bacteria have died the population becomes extinct. No individual can be born out of thin air and therefore the process is trapped in this state *ad infinitum*. On the other hand, an empty queue does in no way preclude the possibility of someone walking in and requesting to be served.

The above examples illustrate the two types of *boundary conditions* (BCs) one comes across when dealing with one step processes. The first type of boundary, referred to as *absorbing*, traps the process, whereas the second, referred to as *reflecting*, precludes the process from exiting a certain range of values. In most cases, boundaries are introduced naturally by the formulation of $g(n)$ and $r(n)$. For example, assuming that Eq. (3.44) is defined in the range $n = 0, 1 \ldots$, one can see that a reflecting boundary is introduced at

$n = 0$ if $r(0) = 0$. Thus, the Master equation takes the form:

$$\dot{P}(n,t) = r(n+1)P(n+1,t) + g(n-1)P(n-1,t) - [g(n) + r(n)]P(n,t), \tag{3.46}$$

for $n = 1, 2, \ldots$, and

$$\dot{P}(0,t) = r(1)P(1,t) - g(0)P(0,t). \tag{3.47}$$

Clearly, even if the processes was defined for $n = -\infty \cdots + \infty$, it would have been trapped in the region of positive integers (provided of course that it starts at this region) since the transition down to $n = -1$ is not allowed.

Similarly, $g(0) = 0$ imposes an absorbing boundary. Conventionally, the absorbing boundary is defined at $n = 1$ although state $n = 0$ is actually the absorbing state [116]. In this case the Master equation takes the form:

$$\dot{P}(n,t) = r(n+1)P(n+1,t) + g(n-1)P(n-1,t) - [g(n) + r(n)]P(n,t), \tag{3.48}$$

for $n = 2, 3, \ldots$ and

$$\dot{P}(1,t) = r(2)P(2,t) - [r(1) + g(1)]P(0,t) \tag{3.49a}$$

$$\dot{P}(0,t) = r(1)P(1,t). \tag{3.49b}$$

The absence of negative terms on the right-hand side of the last equation implies that state $n = 0$ acts as a probability sink. Once the process reaches that state it remains there.

In general, natural boundaries are introduced at all points $n = n_b$ where the form of the analytic functions $g$, $r$ dictate $g(n_b) = 0$ or $r(n_b) = 0$. However, in certain cases (as we shall see when discussing the first passage properties of one step processes) one is interested in erecting artificial boundaries so that the behaviour of the process can be studied within a given interval. Of course this can be accomplished by arbitrarily requiring certain transition probabilities to be zero. However, to preserve the the analytic form of $g(n)$ and $r(n)$ one usually resorts to a mathematically more convenient method of formulating boundary conditions. Consider the one step process described by Eq. (3.47): $r(0) = 0$ does not hold, nevertheless, one is interested in confining the process in the semi-infinite range ($n = 0, 1, \ldots$). By imposing the condition

$$r(0)P(0,t) = g(-1)P(-1,t), \tag{3.50}$$

one readily sees that for $n = 0$ Eq. (3.46) is retrieved. Therefore a reflecting boundary has been implemented by introducing the *fictitious* state $n = -1$ and requiring the above condition.

Similarly one can impose an absorbing boundary at $n = 0$, not by setting $g(0) = 0$ but by treating $n = 0$ as a *fictitious* state with the property

$$P(0,t) \;=\; 0 \tag{3.51}$$

The Master equation now defined for $n = 1, 2, \ldots$ reads

$$\dot{P}(n,t) \;=\; r(n+1)P(n+1,t) + g(n-1)P(n-1,t) - [g(n) + r(n)]P(n,t). \tag{3.52}$$

In this case it should be stressed that the total probability is not conserved $\sum_{n=1}^{\infty} P(n,t)$. Actually, probability is accumulated at $n = 0$, although for our convenience this is ignored by setting $P(0,t) = 0$.

The two equivalent formulations of reflecting and absorbing BCs, presented above for a boundary at $n = 0$, can be used for setting a boundary at any point. In particular, for a Master equation defined on the interval $[a, b]$ the BCs are summarised in the following table

| Boundary | Reflecting | Absorbing |
|---|---|---|
| $a$ | $r(a)P(a,t) = g(a-1)P(a-1,t)$<br>$r(a) = 0$ | $P(a-1,t) = 0$<br>$g(a-1) = 0$ |
| $b$ | $g(b)P(b,t) = r(b+1)P(b+1,t)$<br>$g(b) = 0$ | $P(b+1,t) = 0$<br>$r(a+1) = 0$ |

### 3.3.2 Stationary Solutions

In the long time limit all solutions of the Master equation [see Eq 3.44], $P(n,t)$ will tend to the *stationary solution*, $P_s(n)$. In other words as $t \to \infty$ the process becomes a stationary one and its statistical properties become time-independent. This is always the case for one-step processes with a finite state space, but can also be true for processes defined on an infinite range under certain conditions [141]

To obtain the stationary solution of a one-step process one has to set the derivative on the left hand-side of Eq. (3.44) equal to zero. After some rearrangement one obtains

$$0 \;=\; \{g(n+1)P_s(n+1) - r(n)P_s(n)\} + \{r(n-1)P_s(n-1) - g(n)P_s(n)\}\,. \tag{3.53}$$

The above is usually written in the form

$$0 \;=\; J(n+1) - J(n), \tag{3.54}$$

where we have defined

$$J(n) \;\equiv\; r(n)P_s(n) - g(n-1)P_s(n-1). \tag{3.55}$$

The quantity $J(n)$ describes the net probability flux between any two adjacent states $n$ and $n-1$.

To proceed any further one should also take under consideration the range of $n$ for which the process is defined. Let us first consider the case of a process bounded within some interval which, without loss of generality, we take to be $n = 0 \ldots N$. The reflecting boundary at the origin allows us to write $J(0) = 0$ and subsequently this gives rise to

$$\begin{array}{rcll} J(n) & = & 0 & \Rightarrow \\ r(n)P_s(n) & = & g(n-1)P_s(n-1), & \end{array} \tag{3.56}$$

for all $n$. To the physicist the above condition is reminiscent of the *detailed balance* condition met in equilibrium statistical mechanics [50, 141]. However, here, it merely states that for one step processes at the stationary state the net probability flow between any two states is zero. By repeatedly applying the above relationship, one ends up with

$$P_s(n) \;=\; \frac{1}{\mathcal{N}} \prod_{k=1}^{n} \frac{g(k-1)}{r(k)}. \tag{3.57}$$

where $1/\mathcal{N} = P_s(0)$. This prefactor can be obtained from the normalisation condition $\sum_{n=0}^{N} P_s(n) = 1$ as follows

$$\begin{array}{rcl} \displaystyle\sum_{n=0}^{N} P_s(n) & = & 1 \Rightarrow \\ \displaystyle P_s(0) + \sum_{n=1}^{N} P_s(n) & = & 1 \Rightarrow \\ \displaystyle \frac{1}{\mathcal{N}} \left( 1 + \sum_{n=1}^{N} \prod_{k=1}^{n} \frac{g(k-1)}{r(k)} \right) & = & 1 \Rightarrow \\ \displaystyle \mathcal{N} & = & \displaystyle \left( 1 + \sum_{n=1}^{N} \prod_{k=1}^{n} \frac{g(k-1)}{r(k)} \right). \end{array} \tag{3.58}$$

Equation (3.57) enables us to calculate the stationary solution of the Master equation even

in the case of nonlinear $g(n)$ and $r(n)$. However, special care is needed if the form of $r(n)$ allows for zeros within the range $n = 0, 1, \ldots, N$ (see unsolved exercise in Ref. [141], p. 141). The existence of points $n_i^*$, $i = 1, 2, \ldots, k$ with the property $g(n_i^*) = 0$ imposes a sequence of reflecting boundaries and in the long time limit the process will be confined within the region $n_k^*, \ldots, N$. The stationary distribution in this case will be given by

$$P_s(n) = \begin{cases} 0, & n < n_k^* \\ \dfrac{1}{\mathcal{N}}, & n = n_k^* \\ \dfrac{1}{\mathcal{N}} \prod_{k=n_k^*+1}^{n} \dfrac{g(k-1)}{r(k)}, & n_k^* < n \leq N \end{cases} . \tag{3.59}$$

subject to normalisation.

The above results also apply for the case of a half-infinite range ($n = 0, 1, 2 \ldots$), if one replaces $N$ with $\infty$. However, attention must be paid as one must make sure that the normalisation factor $\mathcal{N}$ in Eq. (3.57) does indeed converge (see unsolved exercise in Ref. [141] p. 142). A sufficient though not necessary condition of convergence is obtained by applying the ratio test on the infinite sum $\sum_{n=1}^{\infty} \prod_{k=1}^{n} \dfrac{g(k-1)}{r(k)}$ appearing in $\mathcal{N}$. One obtains

$$\lim_{n\to\infty} \frac{g(n-1)}{r(n)} < 1. \tag{3.60}$$

The above condition makes intuitive sense as it does not allow probability escape to $\infty$.

### 3.3.3 System Size Expansion

As stated above, time dependent solutions of the Master equation [Eq. (3.44)] are not generally possible in the case of nonlinear $g(n)$ and $r(n)$. One can, however, make use of approximation techniques provided that the system obeys certain conditions.

One-step processes capture the the stochastic dynamics of systems where only transitions of size $\pm 1$ are possible. In many cases such transitions are small compared to a characteristic quantity $\Omega$ describing the size of the system. The precise prescription of $\Omega$ will depend on the nature of the system considered and can for example be the total size of a bacterial population (assumed constant) or the volume of a reaction cube which is proportional to the total number of molecules present. The requirement $\Omega \gg 1$ sets a clear distinction between two scales: a microscopic one described by $n$ (*extensive variable*) and a macroscopic one described by $x = n/\Omega$ (*intensive variable*). This separation of scales allows one to perform a systematic expansion of the Master equation in terms of the small parameter $\Omega^{-1/2}$. Below we sketch the key steps involved in performing the

expansion [141].

One starts by noting that the transition probabilities $g(n)$ and $r(n)$ can be written in the form:

$$g(n) = f(\Omega)\left[g_0\left(\frac{n}{\Omega}\right) + \frac{1}{\Omega}g_1\left(\frac{n}{\Omega}\right) + \frac{1}{\Omega^2}g_2\left(\frac{n}{\Omega}\right) + \dots\right], \tag{3.61a}$$

$$r(n) = f(\Omega)\left[r_0\left(\frac{n}{\Omega}\right) + \frac{1}{\Omega}r_1\left(\frac{n}{\Omega}\right) + \frac{1}{\Omega^2}r_2\left(\frac{n}{\Omega}\right) + \dots\right], \tag{3.61b}$$

known as their *canonical* form [141]. In this form the transition probabilities become functions of the intensive variable $x = n/\Omega$ and depend on $\Omega$ only through the positive prefactor $f(\Omega)$. Of course, the existence of the canonical form is not guaranteed for any arbitrary function. Nonetheless it turns out that such a form can be written down for most of the cases one meets in practice [141]. Next one has to postulate that

$$\frac{n}{\Omega} = \phi(t) + \frac{\xi}{\sqrt{\Omega}}. \tag{3.62}$$

This is a key step, since the above *anstantz* imposes certain conditions on the time evolution of the stochastic process. In particular, Eq. (3.62) states that at all times our stochastic observable can be decomposed into two parts: a deterministic one, $\Omega\phi(t)$, and a fluctuating one, $\Omega^{1/2}\xi$. One can visualise, $P(n,t)$ therefore as a peak centered around $\Omega\phi(t)$ and of width proportional to $\Omega^{1/2}$. As we shall see, the $\Omega^{-1/2}$ scaling of the fluctuating term allows a purely deterministic description of the system as $\Omega \to \infty$; for finite system sizes it give rise to Gaussian noise around the deterministic value as a first approximation.

The transformation given by Eq. (3.62) yields

$$P(n,t) = \Pi(\xi,t), \tag{3.63a}$$

$$\frac{\partial \Pi}{\partial t} = \frac{\partial P}{\partial t} + \Omega^{1/2}\frac{d\phi}{dt}\frac{\partial \Pi}{\partial \xi}. \tag{3.63b}$$

Using Eq. 3.63 the above as well as the canonical forms of $g(n)$ and $r(n)$ one can trans-

form Eq. (3.44) into

$$\begin{aligned}
&\frac{\partial \Pi(\xi,t)}{\partial t} - \Omega^{1/2}\frac{d\phi}{dt}\frac{\partial \Pi}{\partial \xi} = \\
&f(\Omega)\left[r_0\left(\phi(t)+\frac{\xi+\Omega^{-1/2}}{\Omega^{1/2}}\right)+\frac{1}{\Omega}r_1\left(\phi(t)+\frac{\xi+\Omega^{-1/2}}{\Omega^{1/2}}\right)+\dots\right]\Pi(\xi+\Omega^{-1/2},t) \\
&+f(\Omega)\left[g_0\left(\phi(t)+\frac{\xi-\Omega^{-1/2}}{\Omega^{1/2}}\right)+\frac{1}{\Omega}g_1\left(\phi(t)+\frac{\xi-\Omega^{-1/2}}{\Omega^{1/2}}\right)+\dots\right]\Pi(\xi-\Omega^{-1/2},t) \\
&-f(\Omega)\left[r_0\left(\phi(t)+\frac{\xi}{\Omega^{1/2}}\right)+\frac{1}{\Omega}r_1\left(\phi(t)+\frac{\xi}{\Omega^{1/2}}\right)+\dots\right]\Pi(\xi,t) \\
&-f(\Omega)\left[g_0\left(\phi(t)+\frac{\xi}{\Omega^{1/2}}\right)+\frac{1}{\Omega}g_1\left(\phi(t)+\frac{\xi}{\Omega^{1/2}}\right)+\dots\right]\Pi(\xi,t).
\end{aligned} \tag{3.64}$$

Finally, by Taylor expanding one has (writing up to first order terms)

$$\begin{aligned}
\frac{\partial \Pi(\xi,t)}{\partial \tau} - \Omega^{1/2}\frac{d\phi}{d\tau}\frac{\partial \Pi}{\partial \xi} &= \Omega^{1/2}\left[r_0(\phi(t))-g_0(\phi(t))\right]\frac{\partial \Pi}{\partial \xi} \\
&\quad +\left[r_0'(\phi(t))-g_0'(\phi(t))\right]\frac{\partial(\xi\Pi)}{\partial \xi} \\
&\quad +\frac{1}{2}\left[r_0'(\phi(t))+g_0'(\phi(t))\right]\frac{\partial^2 \Pi}{\partial \xi^2}+\mathcal{O}(\Omega^{-1/2}).
\end{aligned} \tag{3.65}$$

where $\tau = f(\Omega)t$.

So far $\phi$ has been an arbitrary function of time. At this point however one has to choose $\phi$ so as to make the $\Omega^{1/2}$ terms vanish. In particular, one has

$$\frac{d\phi}{d\tau} = g_0(\phi)-r_0(\phi), \tag{3.66}$$

which gives the *macroscopic* behaviour of the system. Along with the initial condition $\phi(0)=x_0=n_0/\Omega$ it completely describes the system in the limit $\Omega\to\infty$ and provides the macroscopic part of the solutions in the case of finite yet large $\Omega$. It should be noted that for nonlinear $g(n)$ and $r(n)$ Eq. (3.66) is an nonlinear ordinary differential equation. There is no guarantee that it can be solved explicitly not even for its stationary solutions $\phi_s$, *i.e.*, roots of the equation

$$g_0(\phi_s)-r_0(\phi_s) = 0. \tag{3.67}$$

Nevertheless, in the case the Master equation describes some physical system, one expects that Eq. 3.66 possesses at least one stable stationary solution, which the time-dependent solutions $\phi(t)$ will approach as $t \to \infty$. For the sake of brevity, in the rest we just assume that such a stable stationary solution $\phi_s$ exist and is unique. In particular, we require the following stability conditions to hold

$$g_0(\phi_s) - r_0(\phi_s) = 0 \quad \text{for a unique } \phi_s, \tag{3.68a}$$

$$g_0'(\phi) - r_0'(\phi) < 0 \quad \text{for all } \phi(t). \tag{3.68b}$$

Terms of order $\Omega^0$ give rise to

$$\frac{\partial \Pi(\xi,t)}{\partial t'} = -\left[g_0'(\phi) - r_0'(\phi)\right] \frac{\partial(\xi\Pi)}{\partial \xi} + \frac{1}{2}\left[r_0'(\phi) + g_0'(\phi)\right] \frac{\partial^2 \Pi}{\partial \xi^2} \tag{3.69}$$

describing the time evolution of the fluctuating part $\xi$. This is a linear Fokker-Planck equation describing a *Ornstein-Uhlenbeck* process [118], that is a process involving diffusion (second term) and linear drift (first term). The solution to any linear Fokker-Planck equation is be found to be Gaussian, so the first moments $\langle\xi\rangle$ and $\langle\xi^\rangle$ suffice to describe the process. By multiplying Eq. (3.69) by $\xi$ and $\xi^2$ and integrating one obtains

$$\frac{d\langle\xi\rangle}{d\tau} = (g_0'(\phi) - r_0'(\phi))\langle\xi\rangle \tag{3.70a}$$

$$\frac{d\langle\xi^2\rangle}{d\tau} = 2(g_0'(\phi) - r_0'(\phi))\langle\xi^2\rangle + [r_0'(\phi) + g_0'(\phi)], \tag{3.70b}$$

subject to the initial conditions $\langle\xi(0)\rangle = \langle\xi^2(0)\rangle = 0$ From the equations above one can directly see why the stability condition $g_0'(\phi) - r_0'(\phi) < 0$ is required. It prevents the moments from growing without bounds and therefore allows for a stationary distribution.

From the above one readily finds that in the stationary state

$$\langle\xi\rangle_s = 0, \tag{3.71a}$$

$$\langle\xi^2\rangle_s = \frac{r_0'(\phi) + g_0'(\phi_s)}{2[g_0'(\phi) - r_0'(\phi_s)]}, \tag{3.71b}$$

where $\phi_s$ is the stable steady state of Eq. (3.66). Finally, the stationary autocorrelation function is given by [141]

$$\langle\xi(0)\xi(\tau)\rangle_s = \langle\xi^2\rangle_s \exp\left[-(g_0'(\phi) - r_0'(\phi_s))\tau\right] \tag{3.72}$$

The results of the system size expansion presented above, namely Equations (3.66), (3.70a) and (3.70b) give to a first approximation the picture of the time dependent and stationary properties of the process for finite $\Omega$. At this point the reader should be referred to reference [141] for a more detailed discussion of the system size expansion as well as appropriate discussion of specific situations where the stability conditions given by Eq. 3.68 are violated. The reader should also be referred to Chapter 7 of this thesis where such a case is treated.

### 3.3.4 Numerical Methods

Numerical methods constitute an alternative approach for dealing with Master equations where time-dependent solutions are not available. Perhaps the simplest and most widely method used is the *Gillespie algorithm* (or *kinetic Monte Carlo method*), originally proposed by Dan Gillespie for simulating systems of chemical reactions [52]. It generates stochastic trajectories of the system that are in exact agreement with the formulation of the Master equation. In this respect, it should be considered an exact method, that is one that does not introduce any errors as for example Euler's method for numerically solving differential equations.

The algorithm is summarised as follows [52]

1. **Initialisation step:**
   (a) Initialise system variables $n \rightarrow n_0$.
   (b) Initialise time $t \rightarrow t_0$.
2. **MonteCarlo step:**
   (a) For each possible transition $i$ $(1, \ldots, k)$ calculate the quantity

$$r_i \;=\; \frac{\sum_{j=1}^{j=i} a_j}{\sum_{j=1}^{k} a_j}, \tag{3.73}$$

   where $a_i$ is the probability per unit time transition $i$ has to occur.
   (b) Generate a uniformly random number $p$ in the interval $[0, 1]$
   (c) Choose the first transition $i$ for which the following condition holds

$$p \;\leq\; r_i. \tag{3.74}$$

   (d) Save the change this transition yields to the system variables $n_s$.

(e) Generate a random number $t_s$ obeying an exponential distribution with rate parameter

$$\lambda \;=\; \sum_{i=1}^{N} a_1. \tag{3.75}$$

3. **Update step:**

   (a) Update system variables $n \to n + n_s$.

   (b) Update time $t \to t + t_s$.

4. **Iteration step:** If the time limit has been exceed or an absorbing boundary has been reached terminate otherwise go to step 2.

The Monte Carlo step is the key step of the algorithm. The idea behind it is a simple one, complying with our formulation of the Master equation. In particular, at each step one chooses a *single* transition to occur with probability that is proportional to the its propensity function $a_i$. Furthermore, the time need for a transition to occur is *exponentially distributed* with mean $1/\sum_i a_i$. These two considerations are identical to the ones we made when deriving the Master equation. Therefore one expects that the Gillespie algorithm yields trajectories that are statistically correct as far as the formulation of the Master equation is concerned.

Each run of the Gillespie algorithm provides one sample trajectory from the infinitely many implied by the Master equation. The method, however, does not assume a constant time-step and therefore to obtain time dependent properties of $P(n,t)$ one must proceed with caution. In particular, one has to run the algorithm a considerable number of times so that adequate statistics are gathered for any time interval $[t, t+\delta t]$ as $\delta t \to 0$. For stationary solutions, one usually runs the algorithm allowing the system to reach its steady state. This can be ensured, by using results obtained form the system size expansion presented above. For example, initialising the system at steady state and allowing the algorithm to run for times much longer than the autocorrelation time will suffice. One can therefore run the algorithm repeatedly and calculate the properties of $P_s(n)$ with arbitrary precision. Alternatively, one long run of the algorithm can be performed. By sampling this single trajectory at times much longer that the autocorrelation time one can obtain the stationary properties of the process. This is ensured by the ergodicity of stationary processes, that is, time averaging is equivalent to ensemble averaging. Summarising, when using the Gillespie algorithm one must pay special attention to errors introduced during sampling. Such errors are unavoidable since one cannot sample the whole space of possible trajectories. One is pacified, however, by the fact that the Gillespie algorithm is an otherwise exact method.

Several other numerical methods for solving the Master equation exist in the literature. Some of these methods can be considered as extensions to the Gillespie algorithm: they allow for more efficient simulations when the system size is large or consists of many variables, whilst remaining exact. Others, compromise exactness by making certain assumptions which allow for faster computation times.

## 3.4 First Passage Processes

Our friend is engaged in his game with the stranger. He has already lost half of his initial capital and he starts thinking whether he should withdraw. After some more thought he decides to continue playing until he regains the amount he has lost or loose everything. Will he break even? For how many more tosses will he have to wait until he breaks even or looses everything? Such questions illustrate the concepts of *a first-passage probability* and *first-passage times*, that is, the probability and time for a stochastic processes to reach some state.

Consider the Master equation for a general one-step process given in Eq. (3.44) defined for in some interval $n = L, \ldots, R$. One wants to know the time $\mathcal{T}_{R,m}$ it takes for the system to reach site $n = R$ for the first time having started from some arbitrary point within the interval $m$ ($R < m < L$). Of course, $\mathcal{T}_{R,m}$ is not a fixed quantity but a stochastic variable obeying the PDF $f_{\mathcal{T}_{R,m}}(t)$, *i.e.*,

$$\text{Prob}(t < \mathcal{T}_{R,m} < t + dt) \;=\; f_{\mathcal{T}_{R,m}}(t)dt. \tag{3.76}$$

Writing down the Master Equation with a reflecting boundary at $L$ and an absorbing one at $R$ one has

$$\begin{aligned}
\dot{P}(L,t) &= r(L+1)P(L+1,t) - g(L)P(n,t), && \text{(3.77a)}\\
\dot{P}(n,t) &= r(n+1)P(n+1,t) + g(n-1)P(n-1,t) \\
&\quad -[g(n)+r(n)]P(n,t), && \text{(3.77b)}\\
\dot{P}(R-1,t) &= g(R-2)P(R-2,t) \\
&\quad -[g(R-1)+r(R-1)]P(R-1,t), && \text{(3.77c)}
\end{aligned}$$

subject to the initial condition $P(n,0) = \delta_{n,m}$. Boundary $R$ acts as a probability sink,

therefore, the probability $S(t)$ that the system at time $t$ has not yet reached $R$ is given by

$$S(t) \;=\; \sum_{n=L}^{R-1} P(n,t). \tag{3.78}$$

$S(t)$ is merely the survival PDF of $\mathcal{T}_{R,m}$ linked to $f_{\mathcal{T}_{R,m}}(t)$ via the relationship

$$f_{R,m}(t) \;=\; -\frac{d}{dt}S(t) \;=\; -\sum_{n=L}^{R-1}\frac{d}{dt}P(n,t) \;=\; g(R-1)P(R-1,t) \tag{3.79}$$

where the last step was performed by summing the Master equation over all permissible $n$.

Similar considerations allow us to calculate the PDFs of $\mathcal{T}_{R,m}$ and $\mathcal{T}_{L,m}$, the time needed for the process to reach either state $R$ of $L$. One has to write the Master equation with two absorbing boundaries present at $L$ and $R$ and obtains

$$f_{\mathcal{T}_{R,m}}(t) \;=\; g(R-1)P(R-1,t) \tag{3.80a}$$
$$f_{\mathcal{T}_{L,m}}(t) \;=\; g(R-1)P(L-1,t) \tag{3.80b}$$

The probabilities of arriving first to either absorbing boundary are given by

$$\pi_{R,m} \;=\; \int_0^\infty f_{\mathcal{T}_{R,m}}(t)dt, \tag{3.81a}$$
$$\pi_{L,m} \;=\; \int_0^\infty f_{\mathcal{T}_{L,m}}(t)dt. \tag{3.81b}$$

These are referred to in the literature of first passage processes as *splitting probabilities* [116] and must of course obey

$$\pi_{R,m} + \pi_{L,m} \;=\; 1. \tag{3.82}$$

Finally using the above one also can obtain the *conditional mean first passage times*, $\langle\mathcal{T}_{R,m}\rangle$ and $\langle\mathcal{T}_{L,m}\rangle$ as well as the *unconditional mean first passage time* (to either boundary) $\langle\mathcal{T}_m\rangle$:

$$\langle\mathcal{T}_{R,m}\rangle \;=\; \frac{1}{\pi_{R,m}}\int_0^\infty t f_{\mathcal{T}_{R,m}}(t)dt, \tag{3.83a}$$
$$\langle\mathcal{T}_{L,m}\rangle \;=\; \frac{1}{\pi_{L,m}}\int_0^\infty t f_{\mathcal{T}_{L,m}}(t)dt. \tag{3.83b}$$
$$\langle\mathcal{T}_m\rangle \;=\; \tau_{R,m} + \tau_{L,m} \tag{3.83c}$$

### 3.4.1 Solving the Master Equation in a Bounded Interval

Therefore, for both cases presented above (reflecting/absorbing and absorbing/absorbing boundaries) the problem of obtaining the PDFs of the first passage times boils down to solving the Master equation in the interval $[L, R]$. In particular, in the case of reflecting/absorbing boundaries on seeks an expression for $P(R-1,t)$ while in the absorbing/absorbing case one seeks expressions for both $P(R-1,t)$ and $P(L-1,t)$.

A straightforward yet laborious technique for solving the Master equation in a bounded interval involves using some integral transform of $P(n,t)$. Most often the *Laplace transform* is chosen:

$$\tilde{P}(n,s) \;=\; \mathcal{L}\{P(n,t)\} \;=\; \int_0^\infty e^{-st} P(n,s)dt. \tag{3.84}$$

Under this transformation, time $t$ is mapped into a new variable $s$ having units of $1/[\text{time}]$. Therefore, the $s$-domain is customarily interpreted as the frequency domain. Nothing is lost under such a transformation and convert back to the time domain using the *inverse Laplace transform*

$$P(n,t) \;=\; \mathcal{L}^{-1}\{\tilde{P}(n,s)\} \;=\; \frac{1}{2\pi i}\lim_{T\to\infty}\int_{\gamma-iT}^{\gamma+iT} e^{st}\tilde{P}(n,s)ds. \tag{3.85}$$

where $i^2 = -1$ and $\gamma$ some real number appropriately chosen (greater than the the real part of all singularities of $\tilde{P}(n,s)$). $\tilde{P}(n,s)$ is particularly useful due to the following property

$$\mathcal{L}\left\{\frac{dP(n,t)}{dt}\right\} \;=\; s\tilde{P}(n,s) - P(n,0). \tag{3.86}$$

Using this property one can transform the Master equation into an algebraic set of difference equations. For example in the case of a reflecting boundary at $L$ and an absorbing at $R$ applying and the Laplace transform on the Master equation yields:

$$\begin{aligned}
s\tilde{P}(L,s) &= r(L+1)\tilde{P}(L+1,s) - \tilde{P}P(n,s), && \text{(3.87a)}\\
s\tilde{P}(n,s) - \delta_{n,m} &= r(n+1)\tilde{P}(n+1,s) + g(n-1)\tilde{P}(n-1,s) \\
&\quad -[g(n)+r(n)]\tilde{P}(n,s), && \text{(3.87b)}\\
s\tilde{P}(R-1,s) &= g(R-2)\tilde{P}(R-2,s) - [g(R)+r(R)]\tilde{P}(R,s), && \text{(3.87c)}
\end{aligned}$$

where he have also made use of the initial condition $P(n,0) = \delta_{n,m}$ The above system consists of $L-R$ equations with $L-R$ unknowns (viz. $\tilde{P}(n,t), n = L,\ldots,R-1$) and

can therefore be solved. Subsequently, $f_{\mathcal{T}_{R,m}}(t)$ can be obtained as

$$f_{\mathcal{T}_{R,m}}(t) \;=\; g(R-1)P(R-1,t) \;=\; g(R-1)\mathcal{L}^{-1}\{\tilde{P}(R-1,s)\} \tag{3.88}$$

Moreover, all integer moments of $f_{\mathcal{T}_{R,m}}$ can be obtained without performing the inverse Laplace transform. This is accomplished by noticing that $\tilde{f}_{\mathcal{T}_{R,m}}(s) = g(R-1)\tilde{P}(R-1,s)$ is the moment generating function of $f_{\mathcal{T}_{R,m}}(t)$ containing all integer moments as coefficients of it power expansion in $s$:

$$\begin{aligned}
\tilde{f}_{R,m}(s) &= g(R-1)\tilde{P}(R-1,s) \\
&= \int_0^\infty g(R-1)e^{-st}P(R-1,t)dt \\
&= \int_0^\infty g(R-1)\left[1-st+\frac{(st)^2}{2}-\dots\right]P(R-1,t)dt \\
&= \int_0^\infty \left[f_{R,m}(t)-stf_{\mathcal{T}_{R,m}}(t)+\frac{(st)^2}{2}f_{\mathcal{T}_{R,m}}(t)-\dots\right]dt \\
&= 1-s\langle\mathcal{T}_{R,m}\rangle+\frac{(s)^2}{2}\langle\mathcal{T}_{R,m}^2\rangle-\dots
\end{aligned} \tag{3.89}$$

The above described method can be easily be extended for the case two absorbing boundaries are present at $L$ and $R$ [116].

### 3.4.2 The Backward Master Equation

A particularly useful tool for solving first passage problems is the *backward* or adjoint Master equation that describes the time evolution of a process backward in time. The master equation defined by Eq. (3.44) describes the time evolution of $P(n,t) \equiv P(n,t|n_0,t_0)$ the probability density of finding the system at state $n$ at time $t$ given that it was initially prepared at state $m$. In this respect, $P(n,t|m,t_0)$ is to be considered as a function of $(n,t)$ while holding $(m,t_0)$ fixed. One can, alternatively also regard $P(n,t|m,t_0)$, as a function of $(m,t_0)$ holding $(n,t)$ fixed, in this case it describes the probability of the initial value $m$ given the system is observed at state $n$ at time $t$. It turns out that the time evolution of $P(n,t|m,t_0)$ obeys an equation similar to the Master equation, dubbed as backward Master equation

$$\begin{aligned}
\frac{dP(n,t|m,t_0)}{dt_0} &= g_m P(n,t|m+1,t_0)+r_m P(n,t|m-1,t_0) \\
&\quad -[g_m+r_m]P(n,t|m,t_0).
\end{aligned} \tag{3.90}$$

Also by noting that for homogeneous processes

$$P(n,t|m,t_0) \;=\; P(n,t-t_0|m,0) \;=\; P(n,t'|m,0) \tag{3.91}$$

one can rewrite the backward Master equation as

$$\begin{aligned} -\frac{dP(n,t'|m,0)}{dt'} \;&=\; g_m P(n,t'|m+1,0) + r_m P(n,t'|m-1,0) \\ &\quad -[g_m + r_m]P(n,t'|m,0). \end{aligned} \tag{3.92}$$

Let us assume in the rest that the process is confined in the interval $n = L,\ldots,R.$ with an reflecting boundary at $L$ (implemented in the backward equation by setting $P(n,t'|L-1,0) = P(n,t'|L,0)$ ) and an absorbing boundary at $R$ ($P(n,t'|R+1,0)$) [1]. We once again focus on the stochastic quantity $\mathcal{T}_{R,m}$, the first passage time to $R$ given the process started at state $m$, which obeys the PDF $f_{\mathcal{T}_{R,m}}(t)$. The survival probability $S(t,m)$ that the process has not yet reached the absorbing boundary is

$$S(t,m) \;=\; \sum_{n=L}^{R-1} P(n,t'|m,0). \tag{3.93}$$

where we explicitly stated that the survival probability is also a function of the initial state $m$. The mean first passage time to $R$ is given by

$$\begin{aligned} \mathcal{T}(m) \;&\equiv\; \langle \mathcal{T}_{R,m} \rangle \\ &=\; \textstyle\int_0^\infty t f_{\mathcal{T}_{R,m}}(t) \\ &=\; -\textstyle\int_0^\infty t \partial_t S(t,m) dt \\ &=\; -\textstyle\int_0^\infty t \partial_t S(t,m) dt \\ &=\; -\textstyle\int_0^\infty S(t,m) dt. \end{aligned} \tag{3.94}$$

where in the last term we have used integration by parts and the fact that $G(\infty,0) = 0$ and $G(0,m) = 1$. Summing Eq. (3.92) over $n = L,\ldots,R-1$ yields an equation for $S(t,m)$. In particular, one has

$$\frac{-dS(t',m)}{dt'} \;=\; g_m S(t,m+1) + r_m S(t,m-1) - [g_m + r_m]S(t,m). \tag{3.95}$$

Now, by integrating over time and making use of the relationship $\mathcal{T}(m) = \int_0^\infty S(t,m)dt$

[1] Note the introduction of the fictitious state $L-1$ and $R+1$

obtained above yields an equation for the mean first passage time

$$\begin{aligned} -\frac{dS(t',m)}{dt'} &= g_m\mathcal{T}(m+1) + r_m\mathcal{T}(m-1) - [g_m + r_m]\mathcal{T}(m) \quad \Rightarrow \\ -[G(\infty,m) - G(0,m)] &= g_m\mathcal{T}(m+1) + r_m\mathcal{T}(m-1) - [g_m + r_m]\mathcal{T}(m) \quad \Rightarrow \\ 1 &= g_m\mathcal{T}(m+1) + r_m\mathcal{T}(m-1) - [g_m + r_m]\mathcal{T}(m). \end{aligned} \tag{3.96}$$

subject to the boundary conditions $\mathcal{T}(R-1) = \mathcal{T}(R)$ and $\mathcal{T}(L+1) = 0$ The above set of difference equations can easily be solved for $\mathcal{T}(m)$ yielding [116]

$$\mathcal{T}(m) = \sum_{i=m}^{R} A(i) \sum_{k=L}^{i} \frac{1}{g_k A(k)}. \tag{3.97}$$

where

$$A(n) = \prod_{i=L+1}^{n} \frac{r_i}{g_i}. \tag{3.98}$$

The above result can be used to obtain the mean first passage times to any point $R$ for an one-step process defined on the range of positive integers $(0, 1, \dots)$. The result can be written in terms of the stationary solution $P_s(n)$ (see unsolved exercise in [141], p. 3201) as

$$\begin{aligned} \mathcal{T}(m) &= \sum_{i=m}^{R} A(i) \sum_{k=L}^{i} \frac{1}{g_k A(k)} \\ &= \sum_{i=m}^{R} \frac{g_0 P_s(0)}{g_i P_s(i)} \sum_{k=0}^{i} \frac{P_s(k)}{g_0 P_s(0)} \\ &= \sum_{i=m}^{R} \frac{1}{g_i P_s(i)} \sum_{k=0}^{i} P_s(k). \end{aligned} \tag{3.99}$$

Finally, multiplying Eq. 3.95 by $t'$ and integrating over $t'$ one obtains

$$\begin{aligned} -2\mathcal{T}(m) &= g_m\mathcal{T}_2(m+1) + r_m\mathcal{T}_2(m-1) - (g_m + r_m)\mathcal{T}_2(m) \\ &= g_m(\mathcal{T}_2(m+1) - \mathcal{T}_2(m)) + r_m(\mathcal{T}_2(m-1) - \mathcal{T}_2(m)), \end{aligned} \tag{3.100}$$

This equation relates the mean first passage time $\mathcal{T}(m)$ to the second moment $\mathcal{T}_2(m) \equiv \langle \mathcal{T}_{R,m}^2 \rangle$. Having already obtained an expression for $\mathcal{T}(m)$ the above equation can be solved recursively yielding a result for $\mathcal{T}_2(m)$. Similarly, successive moments of the first passage probability can be obtained from equation Eq. 3.95 by multiplying with higher powers of $t'$.

## 3.5 Summary

In this Chapter, we presented a brief introduction to the theory of stochastic processes. The aim was to provide the general reader with sufficient background knowledge to understand and appreciate the work presented in subsequent chapters. As the acquainted reader might have noticed, in certain occasions the material presented lacks mathematical rigour and generality and should therefore not be considered as sufficient or complete. The literature, however, on stochastic processes is vast including many comprehensive and coherent introductory books and manuscripts. Refs. [50, 116, 141] are just a few, particularly tailored for interdisciplinary audiences, and upon which the presentation of this Chapter was based.

# Chapter 4

# Single Molecule Level: The Dynamics of a Transcribing RNA Polymerase

As described in Chapter 2, transcriptional pauses disrupt the processive synthesis of RNA and can play a profound role in regulating gene expression. A particular class of pauses is induced by backtracking, a phenomenon that involves the backward translocation of the TEC along the DNA template. In this Chapter, motivated by recent single molecule studies, we present a stochastic model of the transcription elongation phase incorporating backtracking dynamics. Using the model we study the statistics of elongation pauses induced by RNAP backtracking, as well as the effect of these pauses on the statistics of the elongation phase. Our results indicate that pauses due to RNAP backtracking obey a heavy tailed distribution and can significantly alter the statistics of the total elongation times.

## 4.1 Introduction

DNA Transcription constitutes a vital life process through which genetic information stored in DNA is expressed into RNA. The ability of cells to carry out their genetically prescribed function and behaviour crucially relies on the regulation of this process. For example, it has long been know that transcription initiation poses a key step of regulation; enabling cells to modulate the levels of gene expression and hence synchronise their inter-

nal workings or adapt to environmental changes [93,112]. More recently, the regulation of the transcription elongation phase has also become widely appreciated. Regulation at this level is often mediated by transcriptional pauses, which allow specific proteins to interact with poised RNAP molecules and exert their regulatory function [119]. The implication of transcriptional pausing with regard to gene regulation has attracted lately much interest in the dynamics of the elongation phase [32, 53, 63].

A more thorough understanding of DNA transcription has become possible with the *in-vitro* study of the process using single molecule manipulation techniques [63]. In particular, the usage of optical traps has enabled one to track the motion of the transcribing RNAP molecule along the DNA template with near base-pair resolution. (see Fig. 4.1(A)), shedding light on the dynamics of transcription. Such single molecule studies have, for example, showed how RNAP molecules harness thermal noise to translocate along the DNA template, achieving polymerisation rates up to $25$ nt/sec [1]. More importantly, they have revealed that RNAP does not transcribe the template at a constant rate. Rather transcription is frequently interrupted by pauses obeying a wide temporal distribution and lasting up to several minutes (see Fig. 4.1(B-C)). In many cases, pausing is induced by the backward motion of the RNAP on the DNA template, a phenomenon dubbed *backtracking* [58]. During backtracking the RNAP looses grip of the $3'$ end of the RNA, and the transcription elongation complex (TEC) slides backwards along the DNA. The process from there on is diffusional; that is the RNAP is kicked back and forth along the DNA template by thermal noise until the active site reattains its initial position and polymerisation is resumed (see Fig. 4.1(D)). Although backtracking has only been observed *in-vitro*, there is ample evidence concerning its biological significance. In particular, the existence of DNA sequences that promote backtracking indicate that this phenomenon can also play a significant role in the regulation of the elongation phase [7]. Furthermore, backtracking has been directly implicated in transcriptional error correction [124, 136], suggesting that backtracking is also relevant for *in-vivo* transcription.

In this Chapter we aim to quantitatively understand backtracking and its effect on the temporal dynamics of the elongation phase. The remainder of this Chapter is organised as follows. We first present a stochastic model of the transcription elongation phase. The model incorporates polymerisation and depolymerisation of the nascent RNA as well as backtracking. Unlike previous modelling attempts [10, 60, 71, 137], we use the model to provide a quantitative characterisation of transcriptional pausing based on the underlying mechanistic details of backtracking. Our results show that pause lifetimes should obey a wide distribution, and are consistent with experimental findings [47, 65, 92, 99, 124]. Next, we study how pauses affect the statistic of the total elongation time. Our results indicate

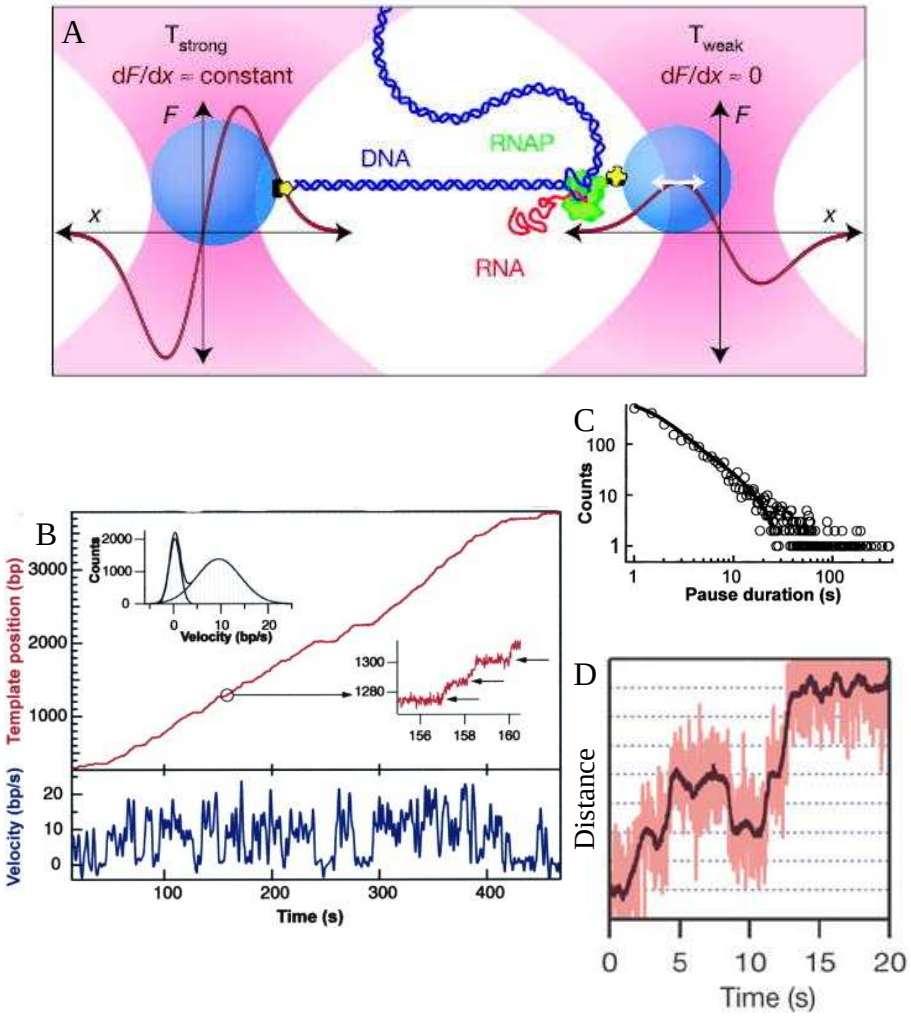



Figure 4.1: Experimental findings from single molecule studies of DNA transcription demonstrating the prevalence of pauses. (A) Schematic illustration of an optical method used for in single molecule studies of DNA transcription. Two beads are held in separate optical traps. A single RNAP molecule is bound to one of the beads while the other one is bound to the downstream end of the DNA. As the RNAP transcribed the DNA, the beads are pulled together. The motion of the RNAP along the DNA template is registered as a displacement of the right bead, which is held by a weaker optical trap. Reprinted by permission from Macmillan Publishers Ltd: E. A. Abbondanzieri *et al.*, Nature, **438** (2005), copyright (2005). (B) Representative trace of the RNAP position along the DNA template. Transcription is interrupted by frequent pauses lasting from $\sim 1$ (right inset, arrows) to several seconds. Reprinted by permission from Elsevier: K. C. Neuman *et al.*, Cell, **115** (2003) Copyright(2003). (C) Distribution of pause lifetimes. Transcriptional pausing occurs on multiple timescales. Here, the distribution is fitted by a sum of two exponentials (solid line) with lifetimes of $1.20.1$ s and $6.00.4$. Reprinted by permission from Elsevier: K. C. Neuman *et al.*, Cell, **115** (2003) Copyright(2003). (D) Backtracking motion of the RNAP molecule along the DNA template. Horizontal lines denote $0.34$ nm spacing (nucleotide length-scale).

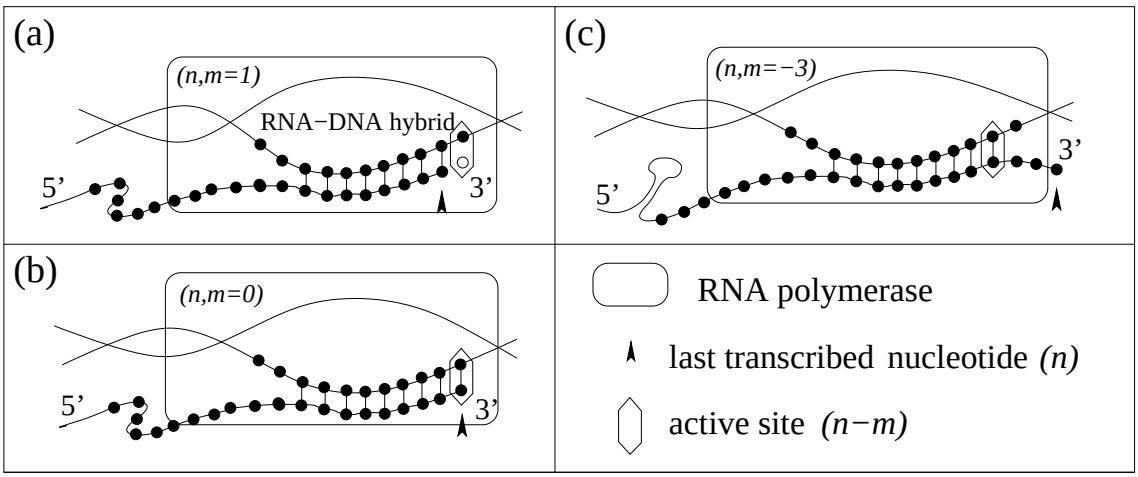



Figure 4.2: Schematic illustration of the transcription elongation complex (TEC) in different translocation states: (a) Post-translocated state at $(n, m = 1)$, (b) pre-translocated state $(n, m = 0)$ and (c) backtracked state $(n, m = -3)$. The position of the last transcribed nucleotide is denoted by $n$. The physical position of the TEC along the DNA template is marked by $m$, the position of the active site relative to $n$.

that backtracking pauses can dramatically affect the temporal statistics of the process, giving rise to a heavy-tailed distribution of elongation times.

## 4.2 A Stochastic Model of the Elongation Phase

In this section we present a stochastic model of the elongation dynamics. The model, motivated by recent experimental findings, incorporates polymerisation and depolymerisation of the nascent RNA as well as backtracking of the RNAP. The basic notation is first introduced and polymerisation/depolymerisation and backtracking dynamics are explained in detail. Finally, key assumptions underlying our modelling attempt are discussed and justified.

### 4.2.1 Basic Notation

A simple model that captures the essence of the elongation phase can be described in terms of two discrete variables $n$ and $m$. Variable $n$ denotes the size of the nascent RNA or equivalently the position of the last transcribed DNA nucleotide. We should note that these two definitions will be used interchangeably throughout the Chapter depending on whether emphasis is wished to be given to the position of the TEC along the DNA or to the length of the RNA. Since our model does not capture transcription initiation, $n$

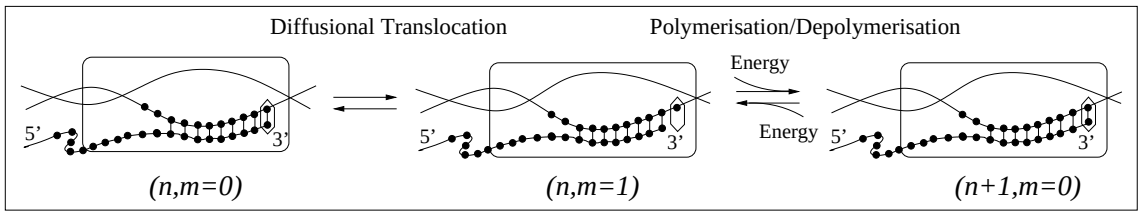



Figure 4.3: Schematic illustration of the state transitions leading to polymerisation and depolymerisation.

is not defined relative to the actual transcriptional starting point (TSP). Rather, $n = 0$ corresponds to the position at which the elongation phase is entered by the formation of the stable TEC usually a few $(8 - 10)$ nucleotides downstream of the TSP [see Chapter 2 (2.2)]. Finally, $n = N$ denotes the end of the transcriptional unit, where the process terminates. The second variable $m$ denotes the position of the polymerase's active site relative to $n$ and ranges from $-n$ to 1. In particular, states $m = 0$ and $m = 1$ correspond to the pre-translocated and post-translocated states of the TEC, respectively, while $m < 0$ denotes backtracked states (see Fig. 4.2).

In summary, $n$ marks the overall progress of the process and is hence affected only by polymerisation and depolymerisation events. On the other hand, $m$ indicates the physical position of the TEC along the DNA template relative to $n$. Alternatively, one could use the absolute position the RNAP active site on the DNA template, *i.e.*, $x = m + n$.

### 4.2.2 Polymerisation/Depolymerisation Dynamics

Our model of the elongation phase starts with the TEC occupying state $(n = 0, m = 0)$. The only transition possible from this state is to the post-translocated state $(n = 0, m = 1)$, from which the TEC can translocate back to $(n = 0, m = 0)$ or proceed with polymerisation $(n = 1, m = 0)$. In general, nucleotide polymerisation can only proceed from the post-translocated state. Thus, with the TEC occupying the pre-translocated state $(n, m = 0)$, polymerisation of a single nucleotide to the nascent RNA chain requires two steps: (1) the TEC sliding forward to the post-translocated state $(n, m = 1)$ and (2) the extension of mRNA by one nucleotide, which leaves the TEC in the next pre-translocated state $(n + 1, m = 0)$. Conversely, the reverse process of depolymerisation can only proceed from the pre-translocated state and leaves the TEC in the previous post-translocated state $(n - 1, m = 1)$. A schematic diagram of state transitions leading to polymerisation/depolymerisation of the nascent RNA is given in Fig. 4.3.

The above described state transitions capture the dynamics of the RNAP as it moves

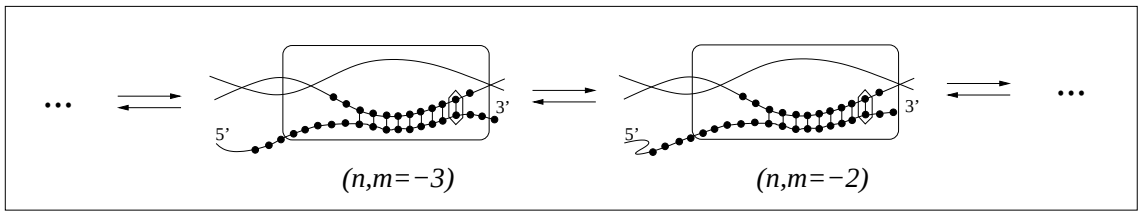



Figure 4.4: Schematic illustration of the state transitions capturing backtracking dynamics. The process involves the diffusional translocation of the TEC along the DNA template, while the length of the RNA transcript remains constant.

along the DNA polymerising the RNA chain, and which are reminiscent of a *Brownian ratchet* [1]. For any given template position $n$, therefore, the TEC owning to thermal noise freely moves back and forth between the pre-translocated $(n, m = 0)$ and the post-translocated $(n, m = 1)$ states. From the post-translocation state $(n, m = 1)$ polymerisation of the next nucleotide is possible. Polymerisation dissipates energy and marks the transition to $(n + 1, m = 0)$. Once polymerisation has occurred going back to $(n, m = 1)$ requires further energy dissipation (to break the phosphodiester bond). Such a depolymerisation event is of course always possible, but on a much longer time-scale than that needed for thermal noise to push the TEC into the post-translocated state $(n + 1, m = 1)$ and enable it to carry on with polymerisation.

### 4.2.3 Backtracking Dynamics

Inclusion of backtracking in the model provides an additional pathway, as the TEC can now hop from the pre-translocated state $(n, m = 0)$ into the first back-tracked state $(n, m = -1)$. Subsequent translocation events, driven by thermal noise, shift the TEC's active site back and forth along the DNA template (see Fig. 4.4). In some cases, backtracking will end as the TEC reattains the pre-translocated state $(n, m = 0)$ (allowing polymerisation/depolymerisation to resume). In other instances, backtracking is interrupted (so called *transcriptional arrest*) and the TEC stalls at some state $(n, m = m^*)$ [58]. In such a scenario accessory proteins[1] can induce cleavage of the exposed $3'$ RNA end, bringing the TEC once again in the pre-translocated state $(n - m^*, m = 0)$.

In theory, backtracking can move the TEC as far back as $(n, m = -n)$ [58]. However, backtracking is often restricted up to a few nucleotides from the last transcribed nucleotide. This restriction stems mainly from interactions between the TEC and the nascent RNA [58]. As the $5'$ end of the RNA exits the TEC, it is free to fold upon itself

[1] such as the Gre/TFIIS cleavage factors [20, 45]

and form stable structures such as RNA hairpins. Subsequently, when the TEC backtracks and slides backwards it interacts with these RNA structures, which preclude extensive backtracking and can even lead to transcriptional arrest. To accommodate the above, in our model we impose a backtracking boundary $M$ so that backtracking is restricted up to $m = -M < -n$ and $m = -n$ when $n \leq M$.

### 4.2.4 Some Key Notes on the Model

The model, presented above, provides a simplified physical picture of the transcription elongation dynamics, particularly relevant to the questions regarding the temporal dynamics of the process that we seek to answer. Here, we discuss and justify key simplifications and assumptions that underlie our modelling attempt.

In our model the TEC is pictured as a rigid body moving along the DNA. Such a simplification allows us to follow the motion of TEC by just using the position of the RNAP's active site as a marker. As far as our model is concerned, all other structural characteristics of the TEC such as the length of RNA-DNA hybrid or the size of the melted DNA region, remain unchanged during its motion. This is approximately valid since large scale conformational changes of TEC have not been observed during its motion and the picture of inchworm-like motion has been abandoned [1, 63].

Furthermore, we picture DNA as a linear chain of sites, which denote the position of nucleotides. Translocation events are assumed to reposition the RNAP's active site by one nucleotide along the DNA chain, either forwards or backwards. In this manner, our model only allows for a finite number of translocation states. These states effectively corresponds to minima in the energy landscape that transiently trap the motion of the TEC.

For the backtracking dynamics, we have assumed that a boundary exists at $m = M$. As discussed above, this boundary captures interactions between the TEC and the nascent RNA that restrict extensive backtracking. However, it should be noted that the distance the TEC is allowed to backtrack is not in general constant but depends on the specific sequence of the RNA and fluctuates owning to the stochasticity with which RNA structures appear and disappear. The fast RNA dynamics however render variations in $M$ rather small of the order of a few nucleotides. Therefore treating $M$ as constant, is not expected to significantly alter the dynamics and constitutes a valid approximation.

So far the model has been presented in its most general form: state transitions capturing the polymerisation/depolymerisation of the RNA and the translocation of the TEC have been defined, however, no rates have been associated with any of these transitions.

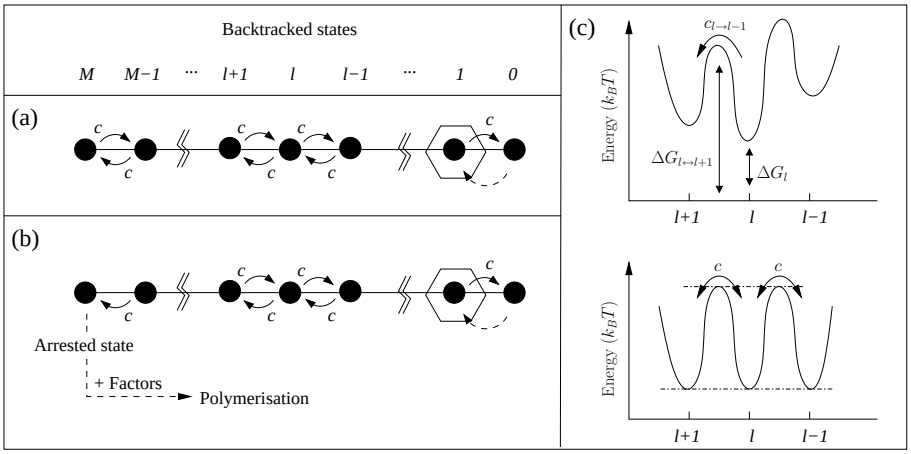



Figure 4.5: (a-b) Schematic illustration of the two cases of backtracking: (a) restricted backtracking and (b) backtracking leading to transcriptional arrest. Variable $l$ denotes the number of nucleotides that the TEC has translocated backwards. Translocation is possible up to $l = M$. A backtracking pause commences with the TEC at state $l = 1$ (dotted arrow) and terminates when state $l = 0$ has been reached. In the second case, the TEC is arrested at state $l = M$, and of elongation factors are necessary to regain a polymerisation competent state (dotted arrow). (c) Schematic illustration of the free-energy landscape during backtracking. According to Kramer's rate theory the rate of hopping depends on the difference between the height of the activation barrier and the free-energy of the current state. Assuming that energetic variations due to sequence inhomogeneities are negligible, an isoenergetic landscape (bottom) is obtained giving rise to equal rates of hopping.

This is left for the subsequent sections where rates are introduced and further assumptions regarding their dependence on the underlying sequence are made.

## 4.3 Backtracking and Elongation Pauses

We first treat the dynamics of RNAP backtracking in isolation from the rest of the process. This enables us to ask the question: what is the lifetime of a single pause induced by backtracking? By formulating the question as a simple first passage problem we are able to obtain analytic results for the distribution of the pause durations. Our results indicate that pauses induced by RNAP backtracking obey a heavy-tailed distribution, which is in agreement with experimental observations [47, 124].

### 4.3.1 Mathematical Formulation

As described in section 4.2.3 during backtracking the TEC hops between consecutive translocation states denoted by $m < 0$. Here, however, to avoid negative integers we shall use the notation $l \equiv -m$. Backtracked states correspond to minima in the free energy landscape that transiently trap the diffusional motion of the TEC. The rate ($c_{l\to l\pm1}$) at which transition between consecutive states ($l \to l \pm 1$) occur will depend on the free energy landscape and according to Kramer's rate theory [69] is given by

$$c_{l\to l\pm1} = c_0 \exp\left[-(\Delta G_{l\leftrightarrow l\pm1} - \Delta G_l)/k_B T\right], \tag{4.1}$$

where $c_0$ is a prefactor, $k_B$ is the Boltzmann constant, $T$ is the absolute temperature, and $\Delta G_l$ and $\Delta G_{l\leftrightarrow l+1}$ denote the free energy of the current state and the height of the activation barrier, respectively [see Fig. 4.5(c)].

Initially, the TEC is considered to attain state $l = 1$. From there, the dynamics of $P(l,t)$, the PDF of finding the TEC in state $l$ at time $t$ given it was in state $l = 1$ at $t = 0$, are described by the Master equation:

$$\frac{\partial P(l,t)}{\partial t} = c_{l-1\to l}P(l-1,t) + c_{l+1\to l}P(l+1,t) - (c_{l\to l+1} + c_{l\to l-1})P(l,t) \tag{4.2}$$

Backtracking terminates when the TEC slides back to state $l = 0$, therefore we impose on Eq. (4.2) the boundary condition $P(0,t) = 0$. Furthermore we consider two biologically relevant scenarios (discussed in section 4.2.3) corresponding to different boundary conditions imposed on state $l = M$:

1. *Restricted backtracking* – no translocation is possible beyond state $l = M$ (reflecting boundary)

2. *Backtracking leading to transcriptional arrest* – the TEC gets trapped at state $l = M$ (absorbing boundary)

The free-energy landscape that dictates the rates of hopping between contiguous states is shaped mainly by the length of the RNA-DNA hybrid, which is the major contributor to the stability of the TEC [58]. Additional contributions come from the actual sequence of hybrid as well as from nonspecific interactions between the RNAP, the DNA and the transcript [58]. Since we have assumed that the length of the hybrid and all other structural properties of the TEC remain relatively unchanged, we can neglect energetic variations due to changes in the sequence,[2] and regard the TEC as moving in a periodic free-energy

[2]We assume that energetic variations due to sequence inhomogeneity are averaged out over the length

landscape [see Fig. 4.5(c)]. This enables us to treat backtracking as purely diffusional process (unbiased random walk) with a constant rate $c$. Equation (4.2) then becomes:

$$\frac{\partial P\left(l,t\right)}{\partial t} = cP\left(l-1,t\right)+cP\left(l+1,t\right)-2cP\left(l,t\right), \tag{4.3}$$

subject to the same boundary conditions as above.

### 4.3.2 Case I – *Restricted Backtracking*

In this case backward translocation beyond state $l = M$ is blocked, owing to interactions between structural elements of the transcript and the TEC. The corresponding boundary conditions for Eq. (4.3) are: $P(0,t) = 0$ (absorbing) and $cP(M,t) = cP(M+1,t)$ (reflecting).

We are interested in the statistics of the pause lifetime $\mathcal{T}_0$, or alternatively the first passage time to state $l = 0$. As we have seen in Chapter 3 (3.4) the PDF of $\mathcal{T}_0$ is given by the probability flux to state $l = 0$, *i.e.*, $\mathcal{P}_{\mathcal{T}_0}(t) \equiv cP(1,t)$, which can be obtained using the Laplace transform method [116]. In particular, using the Laplace transform $\tilde{p}(l,s) = \int_0^\infty P(l,t)e^{-st}dt$, we can eliminate the time derivative in Eq. (4.3) and obtain a set of algebraic difference equations:

$$s\tilde{p}\left(l,s\right)-\delta_{l,1} = c\tilde{p}\left(l-1,s\right)+c\tilde{p}\left(l+1,s\right)-2c\tilde{p}\left(l,s\right), \tag{4.4}$$

where $\delta_{l,1}$ is the Kronecker delta. The corresponding boundary conditions in the Laplace domain are $\tilde{p}(0,s) = 0$ and $c\tilde{p}(M,s) = c\tilde{p}(M+1,s)$. We solve Eq. (4.4) recursively to obtain a closed formula for $\tilde{\mathcal{P}}_{\mathcal{T}_0}(s) \equiv c\tilde{p}(1,s)$, the Laplace transform of the probability flux to state $l = 0$:

$$\tilde{\mathcal{P}}_{\mathcal{T}_0}(s) = \frac{\sinh\left[M\phi(s)\right]-\sinh\left[(M-1)\phi(s)\right]}{\sinh\left[(M+1)\phi(s)\right]-\sinh\left[M\phi(s)\right]}, \tag{4.5}$$

where $\tanh\left[\phi(s)\right] = \sqrt{1-\frac{1}{(s/2c+1)^2}}$.

**Moments of $\mathcal{P}_{\mathcal{T}_0}(t)$**

Equation 4.5 is an exact result as far as our model of backtracking is concerned as $\tilde{\mathcal{P}}_{\mathcal{T}_0}(s)$ is the moment generating function of the PDF we seek, $\mathcal{P}_{\mathcal{T}_0}(t)$. In particular, $\tilde{\mathcal{P}}_{\mathcal{T}_0}(s=0)$ yields the probability of eventually hitting state $l = 0$. This quantity can be trivially

of the RNA-DNA hybrid.

calculated to be 1, that is, the TEC will eventually exit the pause and resume elongation. Furthermore, the coefficients of the Taylor expansion of $\tilde{\mathcal{P}}_{\mathcal{T}_0}(s)$ around $s = 0$ yield the raw moments of the distribution [116]:

$$\begin{aligned}\tilde{\mathcal{P}}_{\mathcal{T}_0}(s) &= \tilde{\mathcal{P}}(s=0) + \frac{s}{1!}\frac{d\tilde{\mathcal{P}}_{\mathcal{T}_0}(s)}{ds}\Big|_{s=0} + \frac{s^2}{2!}\frac{d^2\tilde{\mathcal{P}}_{\mathcal{T}_0}(s)}{ds^2}\Big|_{s=0} + \mathcal{O}(s^3) \\ &= 1 + \langle \mathcal{T}_0\rangle s + \frac{\langle \mathcal{T}_0^2\rangle s^2}{2} + \mathcal{O}(s^3). \end{aligned} \tag{4.6}$$

Some straightforward calculations lead to expressions for the mean pause duration $\langle \mathcal{T}_0\rangle$ and variance $\sigma^2_{\mathcal{T}_0}$

$$\langle \mathcal{T}_0\rangle = \frac{M}{c}, \tag{4.7a}$$

$$\sigma^2_{\mathcal{T}_0} = \langle \mathcal{T}_0^2\rangle - \langle \mathcal{T}_0\rangle^2 = \frac{M + 2M^3}{3c^2}. \tag{4.7b}$$

**Approximate Result for $\mathcal{P}_{\mathcal{T}_0}(t)$**

From Eq. (4.5), using the addition theorem for the hypebolic sine[3] and taking the limit $s/c \ll 1$ (corresponding to $t \gg 1/c$), one obtains an approximate result for $\tilde{\mathcal{P}}_{\mathcal{T}_0}(s)$:

$$\begin{aligned}\tilde{\mathcal{P}}_{\mathcal{T}_0}(s) &= \frac{\sinh[M\phi(s)] - \sinh[M\phi(s)]\cosh[-\phi(s)] - \cosh[M\phi(s)]\sinh[-\phi(s)]}{\sinh[(M+1)\phi(s)] - \sinh[(M+1)\phi(s)]\cosh[-\phi(s)] - \cosh[(M+1)\phi(s)]\sinh[-\phi(s)]} \\ &\approx \frac{\sinh\left[M\sqrt{s/c}\right] - \sinh\left[M\sqrt{s/c}\right]\cosh\left[-\sqrt{s/c}\right] - \cosh\left[M\sqrt{s/c}\right]\sinh\left[-\sqrt{s/c}\right]}{\sinh\left[(M+1)\sqrt{s/c}\right] - \sinh\left[(M+1)\sqrt{s/c}\right]\cosh\left[-\sqrt{s/c}\right] - \cosh\left[(M+1)\sqrt{s/c}\right]\sinh\left[-\sqrt{s/c}\right]} \\ &= \frac{\sinh\left[M\sqrt{s/c}\right] - \sinh\left[M\sqrt{s/c}\right](1+\dots) - \cosh\left[M\sqrt{s/c}\right]\left(-\sqrt{s/c}+\dots\right)}{\sinh\left[(M+1)\sqrt{s/c}\right] - \sinh\left[(M+1)\sqrt{s/c}\right](1+\dots) - \cosh\left[(M+1)\sqrt{s/c}\right]\left(-\sqrt{s/c}+\dots\right)} \\ &\approx \frac{\cosh\left[M\sqrt{s/c}\right]}{\cosh\left[(M+1)\sqrt{s/c}\right]}. \end{aligned}$$

The above result can be readily transformed back to the time domain, yielding an approximation for $\mathcal{P}_{\mathcal{T}_0}(t)$ valid for times much longer than the average stepping time, $t \gg 1/c$. In terms of the Jacobi $\theta_1$ the inversion yields [105]

$$\mathcal{P}_{\mathcal{T}_0}(t) \approx a^{-1}\frac{\partial}{\partial \nu}\theta_1\left(\frac{1}{2}\nu a^{-1}\Big| ta^{-2}\right), \tag{4.8}$$

[3] $\sinh(x+y) = \cosh(x)\sinh(y) + \sinh(x)\cosh(y)$

where $\nu = M/\sqrt{c}$, $a = (M+1)/\sqrt{c}$ and $\theta_1(z|q)$ can be expressed in series as [105]

$$\theta_1(z|q) = \frac{1}{\sqrt{\pi t}} \sum_{n=-\infty}^{\infty} (-1)^n \exp\left[-(z+n-1/2)^2/q\right]. \tag{4.9}$$

Simpler expressions for $\mathcal{P}_{\mathcal{T}_0}(t)$, exemplifying the behaviour of the process, can be obtained in the limits $t \ll M^2/c$ and $t \gg M^2/c$

$$\mathcal{P}_{\mathcal{T}_0}(t) \approx \begin{cases} \dfrac{t^{-3/2}}{2\sqrt{\pi c}}, & \dfrac{1}{c} \ll t \ll \frac{M^2}{c}, \\ \dfrac{\pi c}{(1+M)^2} \sin\left(\dfrac{\pi}{2(M+1)}\right) \exp\left[-\dfrac{c\pi^2}{4(1+M)^2} t\right], & t \gg \dfrac{M^2}{c}. \end{cases} \tag{4.10}$$

The picture obtained from the above form of $\mathcal{P}_{\mathcal{T}_0}(t)$ is a rather simple one. For times short compared to the time scale of diffusion to the reflecting state $l = M$ (*i.e.*, $t \ll M^2/c$), $\mathcal{P}_{\mathcal{T}_0}(t)$ scales as $t^{-3/2}$, as expected for the first passage probability of a random walker in a semi-infinite, one-dimensional domain [116]. Conversely, for times much longer than $M^2/c$, the effect of the reflecting boundary becomes apparent, altering the asymptotics of $\mathcal{P}_{\mathcal{T}_0}(t)$ and imposing a rapid exponential decay. The two different asymptotic behaviours are illustrated in Fig. 4.6, where the analytic result [Eq. (4.8)] have been plotted together with the data obtained from stochastic simulations of the model.

### 4.3.3 Case II – *Backtracking Leading to Transcriptional Arrest*

In this case the TEC initially occupies state $l = 1$ and can resume polymerisation when state $l = 0$ has been reached. However, here, state $l = M$ signals the entrance into an arrested state, form which the TEC can only escape with the aid of accessory elongation factors [20, 45]. Hence, the boundary conditions imposed on Eq. (4.3) are absorbing at both ends: $P(0,t) = P(M,t) = 0$.

The existence of two absorbing boundaries introduces only minor differences from Case I. Here, we are interested in both the PDF of the recovery time $\mathcal{T}_0$, $\mathcal{P}_{\mathcal{T}_0}(t) \equiv cP(1,t)$, and the PDF of time to arrest $\mathcal{T}_M$, $\mathcal{P}_{\mathcal{T}_M}(t) \equiv cP(M-1,t)$. Following a similar treatment as in Case I we obtain an exact analytic result for the moment generating functions of the

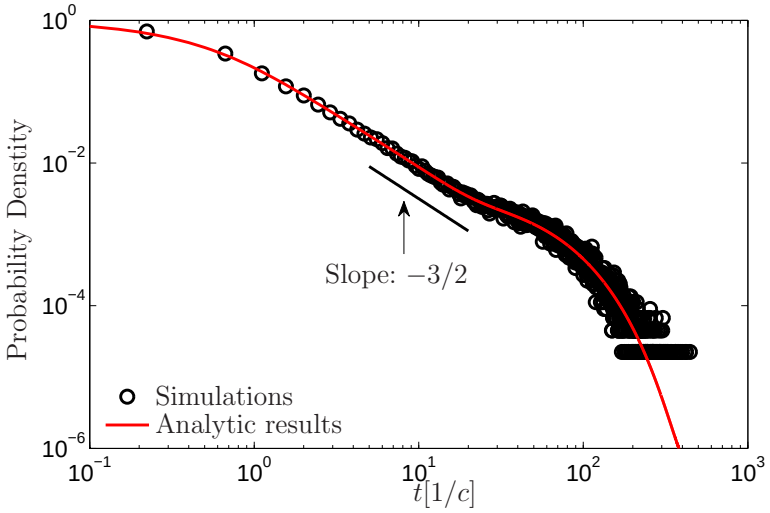



Figure 4.6: The probability density function of pause lifetimes ($\mathcal{P}_{\mathcal{T}_0}(t)$) in the case of restricted backtracking ($M = 10$). Plotted are the analytic result [Eq. (4.8)] (solid line) and the results of stochastic simulations of the model (circles). The PDF $\mathcal{P}_{\mathcal{T}_0}(t)$ exhibits a power law decay ($1/c \ll t \ll M^2/c$), followed by an exponential cutoff in long time limit ($t \gg M^2/c$).

two probability distributions

$$\tilde{\mathcal{P}}_{\mathcal{T}_0}(s) = \frac{\sinh\left[(M-1)\phi(s)\right]}{\sinh\left[M\phi(s)\right]}, \tag{4.11a}$$

$$\tilde{\mathcal{P}}_{\mathcal{T}_M}(s) = \frac{\sinh\left[(\phi(s)\right]}{\sinh\left[M\phi(s)\right]}, \tag{4.11b}$$

where $\tanh\left[\phi(s)\right] = \sqrt{1 - \frac{1}{(s/2c+1)^2}}$.

**Moments of $\mathcal{P}_{\mathcal{T}_0}(t)$ and $\mathcal{P}_{\mathcal{T}_M}(t)$**

As before by evaluating the above equations at $s = 0$ yields the probability of eventual recovery, $\pi_0$ and eventual arrest $\pi_M$, which should sum to 1:

$$\pi_0 = 1 - \frac{1}{M}; \quad \pi_M = 1 - \pi_0 \tag{4.12}$$

Furthermore one can obtain the conditional mean times for each event by evaluating the $s$ derivative of $\tilde{\mathcal{P}}_{\mathcal{T}_0}(s)$ and $\tilde{\mathcal{P}}_{\mathcal{T}_M}(s)$ at $s = 0$

$$\langle \mathcal{T}_0 \rangle = \frac{2M-1}{6c}, \tag{4.13a}$$

$$\langle \mathcal{T}_M \rangle = \frac{M^2-1}{6c}. \tag{4.13b}$$

Higher moments $\langle \mathcal{T}_0^k \rangle$ and $\langle \mathcal{T}_M^k \rangle$ can be obtained by evaluating the $k^{\text{th}}$ derivative of $\tilde{\mathcal{P}}_{\mathcal{T}_0}(s)$ and $\tilde{\mathcal{P}}_{\mathcal{T}_M}(s)$ at $s = 0$, respectively.

In the presence of accessory factors, such as the bacterial Gre proteins, the arrested transcript is cleaved and the TEC returns to a polymerisation competent state. If we assume that the accessory factors act on a relatively fast time-scale (as compared with $\langle \mathcal{T}_0 \rangle$ and $\langle \mathcal{T}_M \rangle$), then the overall mean pause duration is just the weighted sum of $\langle \mathcal{T}_0 \rangle$ and $\langle \mathcal{T}_M \rangle$

$$\langle \mathcal{T} \rangle = \pi_0 \langle \mathcal{T}_0 \rangle + \pi_M \langle \mathcal{T}_M \rangle = \frac{M-1}{2c} \tag{4.14}$$

**Approximate Result for $\mathcal{P}_{\mathcal{T}_0}(t)$ and $\mathcal{P}_{\mathcal{T}_M}(t)$**

Moreover, in the limit $t \gg 1/c$, approximate analytic expression can be obtained for $\mathcal{P}_{\mathcal{T}_0}(t)$ and $\mathcal{P}_{\mathcal{T}_M}(t)$ by inverting the Laplace transforms given in Eq. (4.11) The inversion, in terms of the Jacobi $\theta_4$ function, yields [105]

$$\mathcal{P}_{\mathcal{T}_0}(t) \approx a_0^{-1} \frac{\partial}{\partial \nu_0} \theta_4 \left( \frac{1}{2} \nu_0 a_0^{-1} \middle| t a_0^{-2} \right), \tag{4.15a}$$

$$\mathcal{P}_{\mathcal{T}_{M0}}(t) \approx a_M^{-1} \frac{\partial}{\partial \nu_M} \theta_4 \left( \frac{1}{2} \nu_M a_M^{-1} \middle| t a_M^{-2} \right), \tag{4.15b}$$

where $\nu_0 = (M-1)/\sqrt{c}$, $\nu_M = 1/\sqrt{c}$, $a_0 = a_M = M/\sqrt{c}$, and $\theta_4(z|q)$ can be expressed in series as [105]

$$\theta_4(z|q) = \frac{1}{\sqrt{\pi t}} \sum_{n=-\infty}^{\infty} (-1)^n \exp\left[ -(z+n+1/2)^2/q \right]. \tag{4.16}$$

Compact expressions for $\mathcal{P}_{\mathcal{T}_0}(t)$ are obtained in the limits $t \ll 1/c$ and $t \gg M^2/c$:

$$\mathcal{P}_{\mathcal{T}_0}(t) \approx \begin{cases} \dfrac{t^{-3/2}}{2\sqrt{\pi c}}, & \dfrac{1}{c} \ll t \ll \dfrac{M^2}{c}, \\ \dfrac{2\pi c}{M^2} \sin\left(\dfrac{\pi}{M}\right) \exp\left(-\dfrac{\pi^2 c}{M^2} t\right), & t \gg \dfrac{M^2}{c}. \end{cases} \tag{4.17}$$

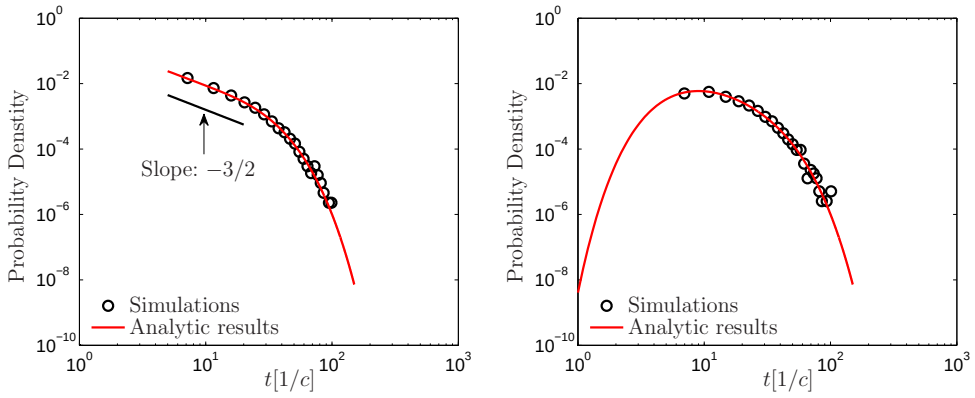



Figure 4.7: Results obtained for backtracking leading to transcriptional arrest (Case II) with $M = 10$: (left) distribution of self-recovered pauses, $\mathcal{P}_{\mathcal{T}_0}(t)$, and (right) distribution of time to arrest, $\mathcal{P}_{\mathcal{T}_M}(t)$. Plotted are the analytic results [Eq. (4.15a) and (4.15b) respectively] as solid lines and the results of stochastic simulations as circles. $\mathcal{P}_{\mathcal{T}_0}(t)$ exhibit a power law decay for $1/c \ll t \ll M^2/c$, followed by an exponential cutoff in long time limit ($t \gg M^2/c$).

Once again, the PDF demonstrates a power law decay for $1/c \ll t \ll M^2/c$, followed by an exponential cutoff. For sufficiently long times, $t \gg M^2/c$, that allow diffusion to the boundary $l = M$, the PDF of the time to arrest decays exponentially and is given by

$$\mathcal{P}_{\mathcal{T}_M}(t) \approx \frac{2\pi c}{M^2} \sin\left(\frac{\pi}{M}\right) \exp\left(-\frac{\pi^2 c}{M^2} t\right), t \gg \frac{M^2}{c}. \tag{4.18}$$

The different asymptotic behaviours are illustrated in Fig. 4.7, where the analytic results have been plotted together with the data obtained from stochastic simulations of the model.

### 4.3.4 The Effect of Applied Force

A key characteristic of the single molecule techniques used to study the dynamics of the elongation phase is that they allow the application of loads on the RNAP as it transcribes the DNA. Studying the effect that external forces have on the elongation dynamics is of key importance, since RNAP molecules continuously have to overcome transcriptional roadblocks or forces due to the coiling of the DNA molecule [47]. In this section we briefly discuss the effect of forcing on backtracking dynamics.

So far the TEC has been assumed to diffuse on periodic free-energy landscape where minima correspond to distinct backtracked states that are separated by the length-scale of a nucleotide, $\delta x = 3.4$Å. External forcing tilts this energy landscape resulting in a

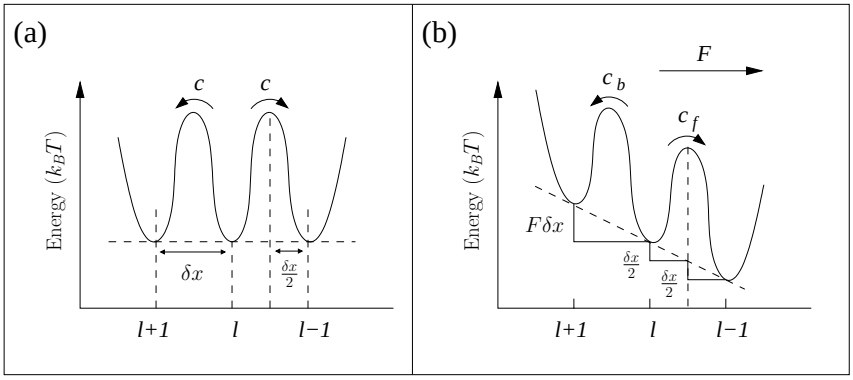



Figure 4.8: Schematic illustration of the free-energy landscape during backtracking with and without external forcing. As a result of an external force with magnitude $+F$ (assisting) the free-energy landscape is tilted by a factor of $F\delta x$ per backtracked state.

translocation bias. In particular, the application of an external force on the RNAP tilts the free-energy landscape by a factor of $F\delta x$ per translocation state, where $F$ is the magnitude of the force in the direction of movement (see Fig. 4.8). That is, the energy of successive backtracked states differs by $F\delta x$ [69]. According to Kramer's rate theory, the forward and backward translocation rates become

$$c_f = c\exp\left[\frac{F\delta x}{2\ k_BT}\right], \tag{4.19a}$$

$$c_b = c\exp\left[-\frac{F\delta x}{2\ k_BT}\right], \tag{4.19b}$$

where $c$ is the translocation rate in the absence of any forcing.

The Master master equation describing the backtracking dynamics in the presence of force is given by

$$\frac{\partial P\left(l,t\right)}{\partial t} = c_bP\left(l-1,t\right)+c_fP\left(l+1,t\right)-\left(c_f+c_b\right)P\left(l,t\right). \tag{4.20}$$

subject to the same boundary conditions discussed in preceding sections: $P(0,t)=0$, $c_bP(M,t)=c_fP(M+1,t)$ for Case I (restricted backtracking) and $P(0,t)=0$,$P(M,t)=0$ for Case II (backtracking leading to transcriptional arrest). In what follows we focus on the case of restricted backtracking. However, similar results can be be obtained for the second scenario as well.

As before an exact result can be obtained for the Laplace transform of $\mathcal{P}_{\mathcal{T}_0}(t)$, the PDF

of the pause lifetimes $\mathcal{T}_0$:

$$\tilde{\mathcal{P}}_{\mathcal{T}_0}(s) = \sqrt{r}\,\frac{\sqrt{r}\sinh\left[M\phi(s)\right]-\sinh\left[(M-1)\phi(s)\right]}{\sqrt{r}\sinh\left[(M+1)\phi(s)\right]-\sinh\left[M\phi(s)\right]}, \tag{4.21}$$

where $\tanh\phi(s)=\sqrt{1-\frac{4R}{(s/c_b+r+1)^2}}$, and $r\equiv\dfrac{c_f}{c_b}=\exp\left[\dfrac{F\delta x}{k_BT}\right]$. Parameter $r$ quantifies the effect of the force, with $r>1$ indicating an assisting force and $r<1$ an opposing one. In the absence of external forcing ($r=1$) one can easily verify that the above equation reduces to Eq. (4.5). The expression found for $\tilde{\mathcal{P}}_{\mathcal{T}_0}(s)$ can be used to obtain analytic results for the moments of the probability distribution $\mathcal{P}_{\mathcal{T}_0}(t)$. In particular, the mean pause duration $\langle\mathcal{T}_0\rangle$ and variance $\sigma_T$ take the form

$$\langle\mathcal{T}_0\rangle = \frac{1-1/r^M}{c_f\,(1-1/r)} \tag{4.22a}$$

$$\sigma^2_{\mathcal{T}_0} = \frac{1}{c_f^2\,(1-1/r)^2}\left[(1+r)\,\frac{1-1/r^M}{1-1/r}-\frac{4\,M}{r^M}\right]. \tag{4.22b}$$

Note that once again taking the limit $R\to 1$ yields the results obtained for the symmetric case [Eq. (4.7)].

As it stands Eq. (4.21) cannot be easily inverted back into the time domain. Instead, numerical methods are used to obtain an estimate of $\mathcal{P}_{\mathcal{T}_0}(t)$ (see section 4.5). Figure 4.9 illustrates distribution of the pause lifetimes for different magnitudes of external forces (assisting or opposing the forward motion of the RNAP). In particular, the heavy-tailed characteristics of the pause distribution, seen in the symmetric case, are still evident for assisting forces up to $F\sim k_BT/\delta x\approx 10pN$ (at room temperature $T=300$K).

## 4.4 The Statistics of the Elongation Phase

Having studied the statistics of backtracking pausing in detail, in this section we use the model of the elongation phase to assess the effect of the transcriptional pauses on the overall dynamics of the process. We particularly focus on the statistics of the elongation times, *i.e.*, the time needed for the TEC to reach position $(n=N,m=0)$ having started from state $(n=0,m=0)$. Two variants of model are considered. First in a model without backtracking (Model A), we show that elongation times scales linearly with the DNA template size. Second in a model that incorporates backtracking (Model B) we find that elongation pauses can dominate the process and give rise to a heavy-tailed distribution of the elongation times.

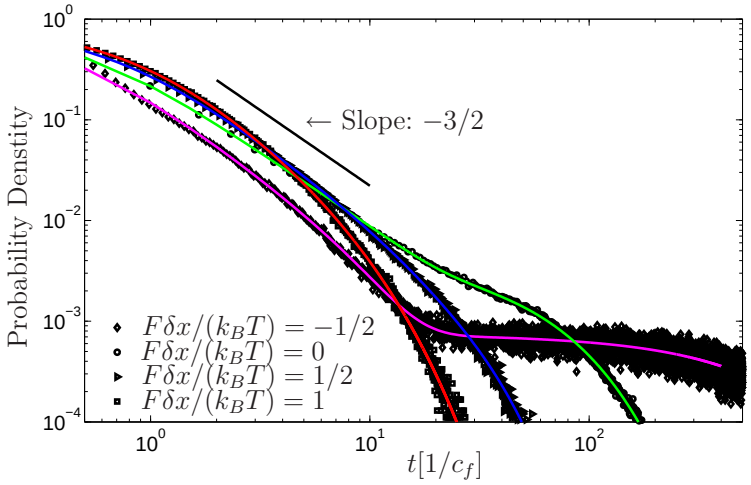



Figure 4.9: Results obtained for restricted backtracking (Case I) in the presence of external forcing and $M = 10$. Solid lines were obtained by numerically inverting $\tilde{\mathcal{P}}_{\mathcal{T}_0}(s)$ [Eq. (4.21)], while markers correspond to results obtained from stochastic simulations of the model [Eq. (4.20) with boundary conditions $P(0,t) = 0$ and $c_b P(M,t) = c_f P(M+1,t)$].

### 4.4.1 Model A – Translocation Limited Polymerisation

In this variant of the model backtracked states are ignored, and at each template position $n$ only two translocation states are possible: $m = 1$ and $m = 0$, which allow transcript polymerisation and depolymerisation, respectively. The rates of polymerisation and depolymerisation are given by $k_+$ and $k_-$, while $a$ and $b$ is the translocation rate from $m = 0$ to $m = 1$ and $b$ the reverse rate from $m = 1$ to $m = 0$. (See typical values in Table 4.1).

The dynamics of $P_{n,m}(t)$, the probability of finding the TEC in state $(n, m)$ at time $t$, are described by the Master equation [50, 141]:

$$\frac{\partial P_{n,0}(t)}{\partial t} = k_+ P_{n-1,1} + bP_{n,1} - (k_- + a)P_{n,0}, \tag{4.23a}$$

$$\frac{\partial P_{n,1}(t)}{\partial t} = k_- P_{n+1,0} + aP_{n,0} - (k_+ + b)P_{n,1}, \tag{4.23b}$$

where $n$ varies from 0 to $N - 1$. We assume that depolymerisation is impossible from $(n = 0, m = 0)$ and that the process is terminated when state $(n = N, m = 0)$ has been reached. Consequently, the boundary conditions imposed on Eq. (4.23) are reflecting at $(n = 0, m = 0)$ and absorbing at $(n = N, m = 0)$. As discussed in Chapter 3 (3.3.1), reflecting boundaries can be implemented by defining a fictitious state $n = -1$ and setting $k_- P_{0,0} = k_+ P_{-1,1}$. On the other hand, to obtain the absorbing boundary, it suffices to set

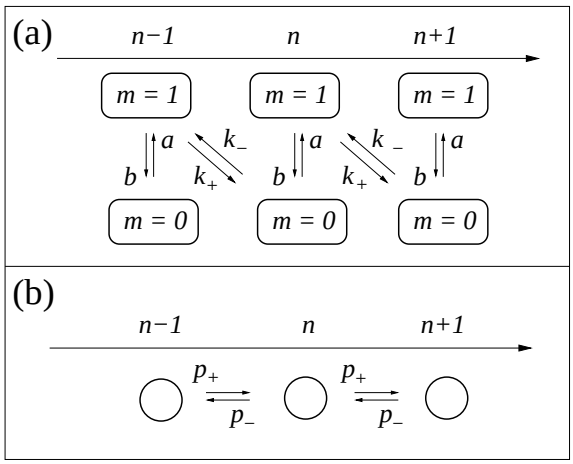



Figure 4.10: (a) Schematic illustration of Model A, including polymerisation depolymerisation, and transitions between the pre- and post-translocated states. (b) A mean field approximation of Model A, yielding a biased random walk, is obtained in the limit of fast translocation dynamics

$P_{N,0} = 0$ [50], which is equivalent to setting the transition rate from $(N, 0)$ to $(N - 1, 1)$ equal to zero.

If we assume that translocation occurs at a much faster time-scale that polymerisation/depolymerisation, *i.e.*, $k_+, k_- \ll a, b$ [10, 60], a *mean-field* (or *quasi-steady state*) approximation is obtained, equivalent to a biased random walk. In this limit, at each position $n$ equilibrium between the two translocation states ($m = 0, 1$) is established rapidly; hence we can write

$$P_{n,1}(t) \approx \frac{a}{a+b} P_n(t), \quad P_{n,0}(t) = P_n(t) - P_{n,1}(t). \tag{4.24}$$

Summing Eq. (4.23) over $m$ and using the above relationship one obtains the Master equation describing the the dynamics of $P_n(t) = P_{n,0}(t) + P_{n,1}(t)$, the probability of finding the TEC at position $n$:

$$\frac{\partial P_n}{\partial t} = p_+ P_{n-1} + p_- P_{n+1} - (p_- + p_+) P_n, \tag{4.25}$$

where the *effective* polymerisation and depolymerisation rates are given by:

$$p_+ \approx \frac{k_+ \, a}{a+b}, \quad p_- \approx \frac{k_- \, b}{a+b}. \tag{4.26}$$

We focus on the total elongation time, $\mathcal{T}_N$, *i.e.*, the time it takes the TEC to arrive

| Parameter | Value | References |
|---|---|---|
| $b/a$ | $0.8$ | [74, 156] |
| $k_b/k_f$ | $0.01$ | [60] |
| $k_f$ | $36s^{-1}$ | [121] |

Table 4.1: Typical values for the rates of polymerisation, depolymerisation and translocation between the post- and pre-translocated states.

at $(n = N, m = 0)$ starting position from $(n = 0, m = 0)$. Using the method of the backward Master equation [see Chapter 3 ( 3.4.2)] we calculate the mean $(\mu \equiv \langle \mathcal{T}_N \rangle)$ and variance $(\sigma^2 \equiv \langle \mathcal{T}_N^2 \rangle - \langle \mathcal{T}_N \rangle^2)$ of $\mathcal{T}_N$:

$$\mu = \frac{1}{p_+ (1-K)} \left[ N - \frac{K\left(1-K^N\right)}{1-K} \right], \tag{4.27a}$$

$$\sigma^2 = \frac{\left(1+K+K^{1+N}\right)}{p_+^2 (1-K)^3} \left[ N - \frac{K\left(1-K^N\right)\left(4+K+K^{1+N}\right)}{(1-K)\left(1+K+4K^{1+N}\right)} \right], \tag{4.27b}$$

where $K = p_-/p_+$.

Figure 4.11 shows results obtained from stochastic simulations of model A [Eq. (4.23)], along with the analytic results obtained in the mean field approximation, for different values of $N$ and $K$. In the small $K$ regime and for small values of $N$, the elongation times are approximately Gamma distributed:

$$P_{\mathcal{T}_N}(t) = t^{\alpha-1} \frac{e^{-t\beta} \beta^\alpha}{\Gamma(\alpha)}, \tag{4.28}$$

where $\Gamma$ denotes the Gamma function and $\alpha = \mu^2/\sigma^2$, $\beta = \sigma^2/\mu$ are the shape and scale parameters of the distribution, respectively. As $N$ is increased the distribution approaches a Gaussian, in agreement with the Central Limit Theorem, with mean and variance given by Eq. (4.27).

Under normal conditions, one expects polymerisation to be overwhelmingly favoured over depolymerisation [58], *i.e.*, $K = p_-/p_+ \ll 1$. Taylor expanding $\mu$ and $\sigma^2$ around $K = 0$ yields

$$\mu = \frac{N}{p_+} + K\frac{(N-1)}{p_+} + \mathcal{O}\left(K^2\right), \tag{4.29a}$$

$$\sigma^2 = \frac{N}{p_+^2} + K\frac{(4N-4)}{p_+^2} + \mathcal{O}\left(K^2\right). \tag{4.29b}$$

Hence, in the limit $K \to 0$ both $\mu$ and $\sigma^2$ scale linearly with the template length $N$,

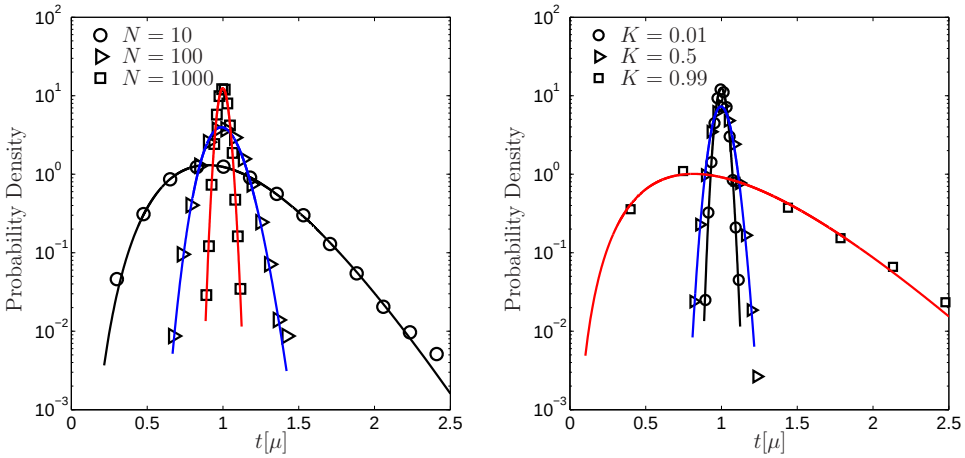



Figure 4.11: The probability density function of the elongation times in the absence of backtracking. Marker denote results obtained from stochastic simulations of the model [see Eq. (4.23)] and are fitted with either a Gamma ($N = 10$) or a Gaussian distribution ($N = 10^2, 10^3$) with mean and variance given by Eq. (4.27). (a) Results for $K = 0.01$, $p_+ = 20$ s$^{-1}$ and different template lengths $N = 10$, $10^2$, $10^3$ bp. (b) Results for $N = 10^3$ bp, $p_+ = 20$ s$^{-1}$ and different polymerisation biases $K = 0.01, 0.5, 0.99$.

and consequently fluctuations around the mean are of the order $1/\sqrt{N}$. In other words, the distribution of the elongation times becomes narrowly peaked around the mean as $N$ is increased, and in the limit $N \to \infty$, where fluctuations tend to zero, the process becomes essentially deterministic. Conversely, in the $K \to 1$ limit, polymerisation and depolymerisation tend to play equal roles, leading to fluctuations in the transcription time that do not vanish as $N$ is increased (see Fig. 4.12).

### 4.4.2 Model B – Elongation with Backtracking

In this case, in addition to polymerisation/depolymerisation and transitions between the the pre-translocated ($m = 0$) and post-translocated ($m = 1$) states, the TEC is allowed to backtrack. In particular the TEC hops from the pre-translocated state ($n, m = 0$) into the first back-tracked state ($n, m = 1$) with rate $k_b$. Subsequent translocation events can randomly shift the TECs active site back and forth, with rate $c$ up to some limit ($n, m = M$). Furthermore, we focus on the case of restricted backtracking, i.e active polymerisation/depolymerisation resumes when the TEC reattains the active states ($m = 0, 1$) and no transcriptional arrest is possible.

The dynamics of $P_{n,m}(t)$, the probability of finding the TEC in state $(n, m)$ at time $t$,

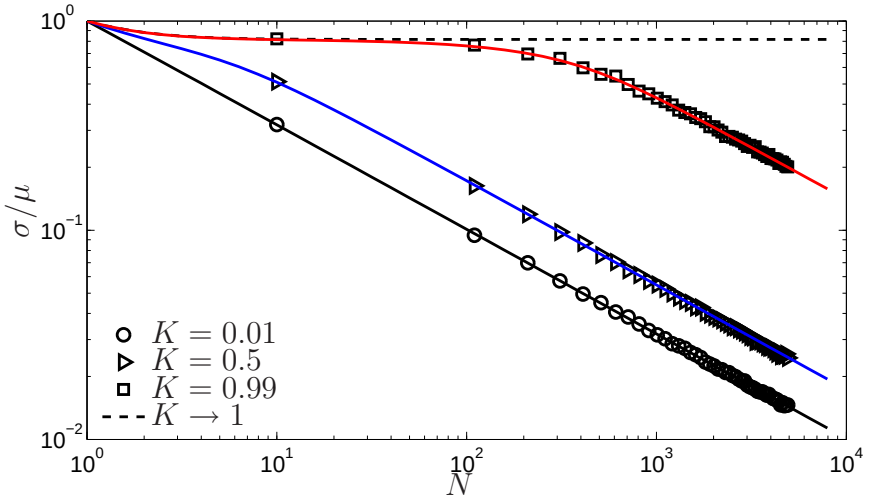



Figure 4.12: Coefficient of variation ($\sigma/\mu$) for the elongation times in the absence of backtracking as a function of the template length $N$ and for different values of $K$. As expected the width of the distribution scales as $1/\sqrt{N}$.

are described by:

$$\frac{\partial P_{n,1}}{\partial t} = k_- P_{n+1,0} + a P_{n,0} - (k_+ + b) P_{n,1}, \tag{4.30a}$$

$$\frac{\partial P_{n,0}}{\partial t} = k_+ P_{n-1,1} + b P_{n,1} + c P_{n,-1} - (k_- + a + k_b) P_{n,0}, \tag{4.30b}$$

$$\frac{\partial P_{n,-1}}{\partial t} = k_b P_{n,0} + c P_{n,-2} - 2c P_{n,-1} \tag{4.30c}$$

$$\vdots \tag{4.30d}$$

$$\frac{\partial P_{n,-M}}{\partial t} = c P_{n,-M+1} - c P_{n,-M} \tag{4.30e}$$

with boundary conditions $k_- P_{0,0} = k_+ P_{-1,1}$ (reflecting) and $P_{N,0} = 0$ (absorbing).

Once again, assuming that the pre-translocated and post-translocated states are in equilibrium one obtains

$$\frac{\partial P_{n,*}}{\partial t} = cPn, -1 + p_+ P_{n-1,*} + p_- P_{n+1,*} - (p_- + p_- + p_b) P_{n,0}, \tag{4.31a}$$

$$\frac{\partial P_{n,-1}}{\partial t} = p_b P_{n,*} + c P_{n,-2} - 2c P_{n,-1} \tag{4.31b}$$

$$\vdots \tag{4.31c}$$

$$\frac{\partial P_{n,-M}}{\partial t} = c P_{n,-M+1} - c P_{n,-M} \tag{4.31d}$$

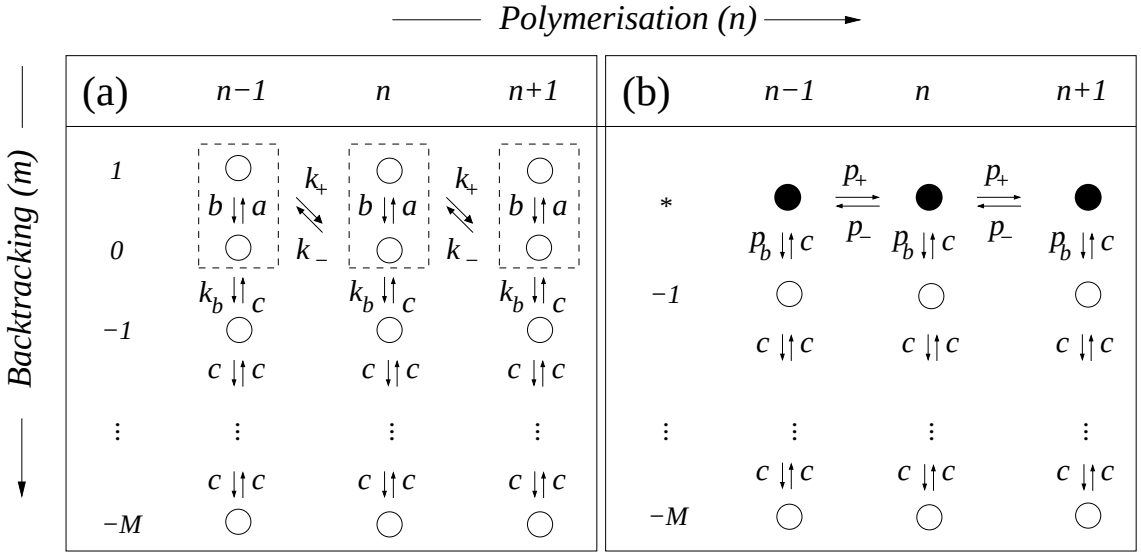



Figure 4.13: Schematic illustration of Model B, involving polymerisation/depolymerisation dynamics and backtracking depolymerisation and backtracking. Model B allows backtracking as far as $m = -M$, with $M \ll N$. If $n < M$, backward translocation is permitted up to state $m = -n$ (not shown).

where $P_{n,*} = P_{n,0} + P_{n,1}$ and the effective rates are given by

$$p_+ \approx \frac{k_+ a}{a+b}, \quad p_- \approx \frac{k_- b}{a+b}, \quad \text{and} \quad p_b \approx \frac{k_b b}{a+b} \tag{4.32}$$

Having characterised backtracking statistics, we use stochastic simulations of the model given by Eq. (4.31) to examine the effects of backtracking on the total elongation time. In particular, the macroscopic (observable) properties that we consider are:

1. the number of pauses $\delta$ over a DNA template of length $N$
2. the aggregate lifetime of all the pauses, $\tau_p$ relative to the time spent on active polymerisation $\tau_a$.

As we shall see these properties are linked to the microscopic parameters $p_b$, $p_+$ and $c$ and will be varied in our stochastic simulations to assess the contribution of pauses to the total elongation time.

Since at every site backtracking is kinetic competition with polymerisation and depolymerisation, one expects that for large templates the number of pauses $\delta$ observed should obey:

$$\frac{\delta}{N} = \frac{p_b'}{p_b + p_+ + p_-} \approx \frac{p_b}{p_+}, \tag{4.33}$$

where in taking the last step we have assumed that the rate of polymerisation is the fastest one, *i.e.*, $p_+ \gg p_b, p_-$.

Moreover, as seen in section 4.3.2 the mean pause duration is $M/c$. Hence, an estimate of the aggregate pause duration is given by

$$\tau_p \;=\; \delta\frac{M}{c} \;\approx\; N\frac{p_b}{p_+}\cdot\frac{M}{c}. \tag{4.34}$$

On the other hand, the time spent on active polymerisation is the one obtained in our treatment of Model A, *i.e.*,

$$\tau_a \;\approx\; \frac{N}{p_+}. \tag{4.35}$$

The ratio of these two time-scales is therefore,

$$R \;\equiv\; \frac{\tau_p}{\tau_a} \;\approx\; p_b\frac{M}{c}, \tag{4.36}$$

which is a dimensionless measure of pausing, quantifying its relative contribution to the elongation time.

Figures 4.14 and 4.15 illustrate the results of the stochastic simulations of Model B [Eq. (4.31)] for different values of $R$ and keeping the frequency of pauses $\delta/N$ constant. As expected, for $R \to 0$ the polymerisation-only model (Model A) is recovered. In particular, the width of the distribution scales like $1/\sqrt{N}$ (see Fig. 4.15) and the distribution of elongation times demonstrates a high peak around the mean elongation time predicted by Model A (see Fig. 4.14 left panel), indicating that either no pauses or only brief ones occur. As $R$ is increased, rare pauses with prolonged durations ($\gg M^2/c$) start to have a significant contribution to the overall elongation time. This effect is clearly illustrated by the heavy-tailed distribution of elongation times seen in Fig. 4.14 (left panel) for $R = 0.1$. In particular, the exponential tail resembles the one found for individual pause lifetimes (see Fig. 4.6) indicating that the total elongation time is often dominated by single rather long-lived pauses. For even higher values of $R$ the elongation phase is dictated by backtracking dynamics and the distribution of elongation times illustrates quasi-exponential characteristics (see Fig. 4.15). For increasing pause frequency (higher $\delta/N$) the effect on the total elongation time is clearly more profound; the distribution becomes broader and exhibits a general shift towards longer elongation times [see Fig. 4.14(b)].

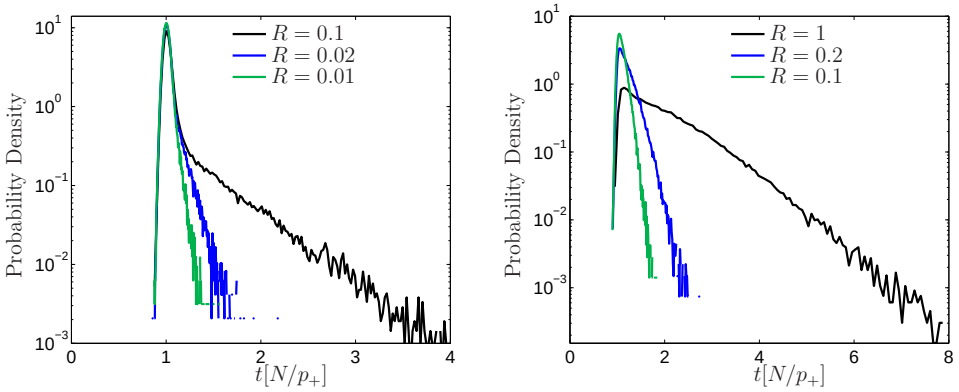



Figure 4.14: Distributions of dimensionless elongation times (scaled by $N/p_+$) in the presence of backtracking (Model B) for different values of the control parameter $R$. The distributions were obtained using stochastic simulations of the model [Eq. (4.31)]. Parameters used: (a) $N = 4$ kb, $M = 10$ bp, $p_+ = 10$ s$^{-1}$, $K = 0.01$ and $p_b$ chosen to yield $\delta/N \approx p_b/p_+ = 1$ pauses/kb (Refs. [46, 124]). (b) $N = 1$ kb, $M = 10$ bp, $p_+ = 10$ s$^{-1}$, $K = 0.01$ and $p_b$ chosen to yield $\delta/N \approx p_b/p_+ = 10$ pauses/kb.

## 4.5 Numerical Methods

In this section we give an overview of the computational tools and numerical methods used to obtain the various results presented.

### 4.5.1 Models of Backtracking

To verify the validity of the analytic results obtained for the statistics of the backtracking pauses (see section 4.3) stochastic simulations of the model [Eq. (4.3)] were performed using the Gillespie algorithm [52]. In particular, the state of the system was monitored using

- a variable $m$ denoting the translocation state of the TEC,
- a timer $t$.

The system was initialised with $m = 0$ and $t = 0$. At each step of the algorithm, all permissible state transitions were calculated based on current translocation state. Then one transition was chosen with probability proportional to the corresponding transition probability and the state of the system was updated [see Chapter 3 (3.3.4)]. The simulation was terminated when an absorbing boundary had been reached and the value of the timer $t$ was saved for analysis. The code was implemented in ANSI-C. Each of the data sets used in Figures 4.6, 4.7, and 4.9 was generated by $10^5$ independent simulation runs. Finally, for

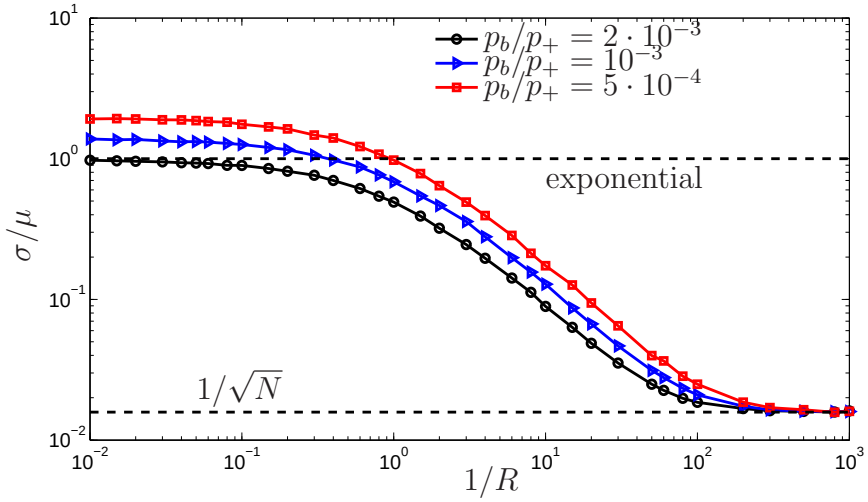



Figure 4.15: Coefficient of variation ($\sigma/\mu$) of the elongation times for Model B as a function of the control parameter $1/R$ and for different values of the of pause frequencies ($p_b/p_+$). As $1/R \to 0$, pauses become more significant and the distribution of elongation times becomes broader. In the case of frequent pausing ($p_b/p_+ = 2 \cdot 10^{-3}$) the distribution exhibits exponential characteristics ($\sigma/\mu = 1$). As $1/R \to \infty$ the effect of pauses vanishes and Model B approaches Model A, where $\sigma/\mu \approx 1/\sqrt{N}$. Results were obtained using stochastic simulations of the model [Eq. (4.31)]. Parameters used: $N = 4$ kb, $M = 10$ bp, $p_b = 0.01$ s$^{-1}$, $K = 0.01$ and $p_+ = 2$, 10 and 20 s$^{-1}$.

the numerical inversion of the Laplace transform in Eq. (4.20) (see Fig 4.9) a MATLAB implementation of the Gaver-Stehfest algorithm[4] was used.

### 4.5.2 Models of Elongation Phase

All data presented in section 4.4 were generated using stochastic simulations of the models given by Equations 4.25 and 4.31. For the simulations the Gillespie algorithm [52] was used. In particular the state of the system was monitored using

- a two variables $(n, m)$ denoting the translocation state of the TEC,
- a timer $t$.

The system was initialised with $(n = 0, m = 0)$ and $t = 0$. At each step of the algorithm, all permissible state transitions were calculated based on current translocation state. Then one transition was chosen with probability proportional to the corresponding transition probability and the state of the system was updated [see Chapter 3 (3.3.4)]. The simulation

[4] freely available from http://www.mathworks.com/matlabcentral/fileexchange/9987

was terminated when the absorbing boundary $(n = N, m = 0)$ had been reached and the value of the timer $t$ was saved for analysis. The code was implemented in ANSI-C and $10^5$ independent simulation runs were performed generating the data used in Figures 4.11, 4.12, 4.14, and 4.15 (see captions for the numerical values of the parameters).

## 4.6 Summary and Discussion

In this Chapter, motivated by recent experimental studies [1, 99, 124], we presented a stochastic model of the single molecule dynamics during the transcription elongation phase. The model incorporates polymerisation and depolymerisation of the nascent RNA as well as the backward translocation of the TEC along the DNA template, a phenomenon dubbed backtracking [58]. Unlike previous modelling attempts [10, 60, 71, 137], our main focus was to to provide a quantitative picture of the temporal dynamics of the process.

Special emphasis was given on the quantitative characterisation of the transcriptional pauses induced via backtracking. Two biologically relevant scenarios were considered; backtracking pauses that end with the TEC sliding back into an elongation competent state and pauses that can potentially lead to transcriptional arrest. For both scenarios we obtained analytic results for the distribution of the pause duration, which we verified with computer simulations. Our results show that transcriptional pausing induced via backtracking obeys a broad distribution, with a power law decay ($t^{-3/2}$) followed by an exponential cutoff. Furthermore, the wide temporal distribution is maintained even in the presence of moderate external loads acting on the RNAP molecule.

Interestingly, our findings are consistent with the non-exponential, heavy-tailed distribution of pause lifetimes observed in single molecule studies of bacterial transcription [99, 124]. Indeed, re-analysis of the data indicates that the pauses are well-fitted by a model similar to the one presented here [35]. More recently, a power law ($t^{-3/2}$) in the distribution of pauses has been also observed for eukaryotic transcription (see Fig. 4.16) [47] . This result was independently explained by the authors using a continuous analog of the model of backtracking present here. In this model during backtracking the TEC is allowed to diffuse continuously on the DNA template, rather than by taking discrete steps (as allowed in our model). The two models become equivalent, however, as long as the length-scale of the stepping in our model is much smaller than the length-scale by which the TEC is allowed to backtrack (*i.e.,* $M \gg 1$). It should also be stressed that the spatial resolution of the experiment did not allow the direct observation of backtracking for all pauses. This leaves open the possibility that shorter pauses did not involve backtracking but were induced through a different mechanism – perhaps similar to the one suggested

for ubiquitous pausing [see Chapter 2 (2.2)] observed in bacterial transcription [63, 99]. In summary, further experiments and more thorough analysis of the data seem to be necessary before a final conclusion could be drawn regarding the dynamics of backtracking and transcriptional pausing.

We also used the model to study how backtracking pauses affect the overall elongation dynamics. In particular, by means of mean field theory and stochastic simulations we obtained results regarding the mean elongation time and its variance. Our key results are particularly instructive in two limits: (i) when pauses cause a weak perturbation to elongation dynamics and (ii) when they significantly affect it. In the first case, elongation times follow a narrow Gaussian distribution with fluctuations around the mean scaling like $1/\sqrt{N}$, where $N$ is the length of the DNA template. In the second regime, when there is a significant number of backtracking pauses whose duration is comparable to the active polymerisation time, there is a dramatic change in the distribution of transcriptional times. In particular, the distribution becomes broader and demonstrates quasi-exponential characteristics

The existence of specific DNA sequences inducing backtracking pauses as well as the presence of accessory proteins assisting their recovery indicate that backtracking plays an important role in the regulation of the elongation phase [7]. To this end, our results have direct implications regarding the simple birth and death models used to interpret the stochastic nature of RNA production and its implication regarding cell behaviour and fate [27,55,114]. In these models, DNA transcription is assumed to obey Poisson statistics under the assumption that the initiation phase constitutes the rate limiting step of the process – an assumption that allows one to disregard elongation dynamics. In general, however, the frequency of transcription initiation has a wide dynamical range *in vivo* [85], and *in vitro* studies have shown that initiation times can be as fast as a few seconds [15, 89, 127, 160]. Hence, rapid initiation times can be significantly shorter than the time needed for elongation, which as we have seen demonstrates features (*i.e.,* pauses) that could dominate the overall rate of transcription [119]. In such cases, simple Poisson models of transcription might need to be revised to incorporate the intrinsic fluctuations of the elongation phase (see Chapter 6).

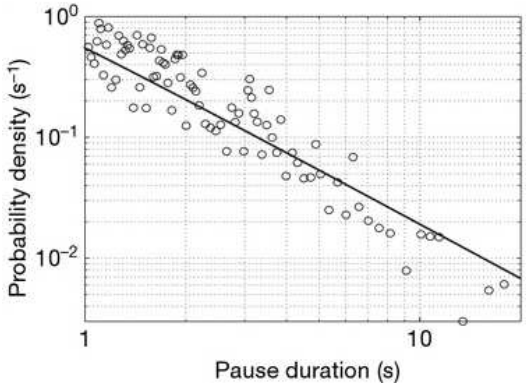



Figure 4.16: Distribution of measured pause durations in single molecule experiments. Data are fitted with a $t^{-3/2}$ power law [47]. Reprinted by permission from Macmillan Publishers Ltd: E. A. Galburt *et al.*, Nature (London) **446**, 820 (2007), copyright (2007)

# Chapter 5

# Transcriptional Error Correction

Life crucially relies on the accuracy with which RNA sequences are transcribed from DNA. To ensure the required levels of fidelity in the face of high spontaneous error rates, DNA transcription relies on error correction mechanisms. A proposed mechanism of transcriptional error correction involves backtracking of the RNA polymerase and mRNA cleavage. In this Chapter we present and study a microscopic model of this editing process. The model offers a quantitative understanding of transcriptional error correction by linking the observed error rate directly to the microscopic rate parameters of the process. Our results indicate that transcriptional error correction via backtracking and RNA cleavage is consistent with a multistep kinetic proofreading scheme. Furthermore, we show that such a mechanism can significantly enhance the fidelity of DNA transcription, yielding error frequencies that are in agreement with *in-vivo* observations.

## 5.1 Introduction

DNA transcription constitutes a vital life process. As discussed in Chapter 2, RNA molecules that are transcribed from the DNA are subsequently used as templates for protein synthesis or can have key roles in various other cellular processes, such as gene regulation and DNA replication. For all these functions to be carried out properly the accuracy of RNA sequences is a crucial requirement. Indeed, errors introduced as the genetic information is transferred into RNA can have far-reaching implications, leading

to the production or non-functional or even malfunctioning proteins and compromising the robust function of the cell [30].

The importance of accuracy during DNA transcription becomes even more profound if one takes into consideration the scale at which the process takes place and the underlying physics. During transcription the RNA polymerase (RNAP) moves along the DNA adding nucleotides to the RNA chain. Let us for the sake of the argument assume that the RNA nucleotides are picked solely on the basis of how well they basepair with the corresponding DNA nucleotide. As we have seen in Chapter 2, basepairing between complementary nucleotides is enabled by hydrogen bonds, which keep the two nucleotides together. In particular, two hydrogen bonds are involved in the formation of a adenine-uracil (A-U) base-pair while three in the case of a guanine-cytosine (G-C) base-pair. This difference of one hydrogen bond is the basis of nucleotide complementarity and is indeed a a very subtle one. Since the energetic contribution of one hydrogen bond is relatively small, in the order of a few $k_BT$ [18], thermal fluctuations dominating the cellular environment are expected to frequently force basepairing between non-complementary nucleotides. More specifically, simple thermodynamics arguments predict that during transcription passive errors should occur at a rate of $10^{-2} - 10^{-3}$ errors/nucleotide [18].

Such high error rates are prohibiting for the survival and perpetuation of life. This is exemplified by the fact that transcriptional error rates observed *in-vivo* are order of magnitudes lower ($10^{-5}$ errors/nt) [18]. Therefore, error correction mechanisms must exist that enhance the discriminatory power of the RNAP and enable it to transcribe RNA chains more accurately than expected from the simple basepairing rule. In particular, experimental evidence is in support of two proofreading mechanisms: one acting at the level of nucleotide addition [143] and the other one mediated through RNAP backtracking and subsequent cleavage of the RNA [124, 147, 159]. The existence of these different proofreading mechanisms raises interesting questions regarding their relative roles in enhancing transcriptional fidelity. These can be answered by the construction of predictive models able to discriminate between the different processes.

In this Chapter we present a theoretical study of the error correcting mechanism mediated by RNAP backtracking and RNA cleavage, hereafter referred to as *nucleolytic proofreading*. Our effort is particularly motivated by recent single molecule studies of DNA transcription that shed light on the microscopic details of backtracking [47, 124] [see also Chapter 4 (4.2.3)] The remainder of this Chapter is organised as follows. We embark by discussing the general problem of biological accuracy and how cellular processes accomplish reduced error rates and increased specificity. We then turn to DNA transcription and present the model of the elongation dynamics involving polymeriza-

tion of correct/incorrect nucleotides, backtracking and RNA cleavage. Using this model we study the role on nucleolytic proofreading in enhancing transcriptional fidelity. Our key results link the observed error rate directly to the microscopic rate parameters of the process and make specific predictions which can be experimentally tested.

## 5.2 Kinetic Proofreading

The question of how cellular processes achieve astonishingly low error rates despite the inherently stochastic environment in which they occur had been puzzling the physics and biology community for quite some time. The motivation had mainly been from DNA replication, where error rates as low as $10^{-9}$ errors/nt had been observed. Although the ability of the DNA polymerase (the enzyme that carries out DNA replication) to cleave nucleotides was a well established fact, the question of how the enzyme was distinguishing between correct and incorrect nucleotides still remained open [96].

Breakthrough finally came around the mid-$70$'s through the seminal work of J. J. Hopfield and J. Ninio [68,101]. Their work proposed an elegant phenomenological framework for explaining how the discriminatory power of enzymes could be enhanced due to differences between the kinetic rates for incorporation and catalysis of correct and incorrect substrates. This now well established framework, known as *kinetic proofreading* (KP) or *kinetic amplification* (KA), provides the fundamental mechanism of accuracy in many diverse biological processes. Examples found in the literature include the antigen recognition by T-cell receptors [90], the disentanglement of DNA by topoisomerases [155], signal transduction [134] and gene expression [19].

The conventional description of KP involves the enzymatic catalysis of two substrates, $S_c$ and $S_w$, obeying Michaelis-Menten kinetics [68]:

$$\begin{aligned} E + S_c &\underset{k_c}{\overset{k_c'}{\rightleftharpoons}} ES_c \overset{\alpha_c}{\rightarrow} E + P_c \quad \text{Correct product,} \\ E + S_w &\underset{k_w}{\overset{k_w'}{\rightleftharpoons}} ES_w \overset{\alpha_w}{\rightarrow} E + P_w \quad \text{Wrong product,} \end{aligned}$$

where $E$ is the enzyme carrying out the catalysis, $ES_c$, $ES_w$ denote the intermediate species and $P_c$, $P_w$ the end products. To quantify the discriminatory power of the enzyme, we define the error fraction $\mathcal{E}$ as

$$\mathcal{E} = \frac{\text{rate of } P_w \text{ formation}}{\text{rate of } P_c \text{ formation}}. \tag{5.1}$$

Assuming that both $S_c$ and $S_w$ are present in equal concentrations and that their discrim-

ination is based on the "off" reaction rates, *i.e.*, $k'_c = k'_w$, $\alpha_c = \alpha_w = \alpha$, and $k_c < k_w$, it can be shown [68] that the limiting error fraction $\mathcal{E}_0$ is attained in the limit $\alpha \ll k_c, k_w$ and is given by

$$\mathcal{E}_0 \equiv \frac{k_c}{k_w} = \exp\left[-\Delta G/(k_B T)\right], \tag{5.2}$$

where $\Delta G$ is the free energy difference between the intermediates ($ES_w$ and $ES_c$), $T$ the absolute temperature, and $k$ is the Boltzmann constant. Therefore, the fidelity of the process is limited by the energy difference between the two intermediate species.

Kinetic proofreading captures the essence of error correction by stipulating the existence of one or more non-equilibrium (irreversible) intermediate steps in the catalytic process. These steps dissipate energy and act as fidelity checkpoints, enhancing the discriminatory power of the catalytic enzyme and resulting in reduced error rates. The simple reaction scheme, treated above, with the inclusion of an additional irreversible step takes the form:

$$\begin{array}{ccccccc} E + S_c & \underset{k_c}{\overset{k'_c}{\rightleftharpoons}} & ES_c & \overset{\alpha_c}{\rightarrow} & ES_c^* & \overset{\beta_c}{\rightarrow} & E + P_c \\ & & & & \downarrow l_c & & \\ & & & & E + S_c & & \\ E + S_w & \underset{k_w}{\overset{k'_w}{\rightleftharpoons}} & ES_w & \overset{\alpha_w}{\rightarrow} & ES_w^* & \overset{\beta_w}{\rightarrow} & E + P_w \\ & & & & \downarrow l_w & & \\ & & & & E + S_w & & \end{array}$$

Once again, all corresponding rates being equal except $k_c < k_w$ and $l_c < l_w$, it can be shown that in the limit $\alpha, \beta \ll k_c, k_w, l_c, l_w$ the error fraction is given by

$$\mathcal{E} = \frac{k_c}{k_w}\frac{l_c}{l_w} = \mathcal{E}_0^2, \tag{5.3}$$

where for the sake of simplicity we have assumed that the free energy difference between the two intermediates $ES_c^*$ and $ES_w^*$ is also $\Delta G$. Therefore, as far as the fidelity of the process is concerned the incorporation of a single irreversible step in the catalytic pathway is equivalent to doubling the energy difference $\Delta G$ in the original catalytic scheme. More generally, the inclusion of $m$ irreversible steps can reduce the error fraction up to $\mathcal{E}_0^{m+1}$. However, it should be noted that the enhancement in the fidelity of the process does not come without a cost. In particular, the time-scale separation $\alpha, \beta \ll k_c, k_w, l_c, l_w$ means that substrate catalysis (even in the case of the correct substrate) undergoes several cycles before the end product is achieved. The energy dissipated in each of these cycles is the price paid for the enhanced accuracy of the process.

Because of its remarkable generality, KP is regarded as a guiding principle for under-

standing how biological processes accomplish the necessary levels of accuracy. However, to complement this general level of description, quantitative and predictive models that incorporate detailed information about specific biological processes are needed. With this in mind, in the remainder of this Chapter we focus on DNA transcription and on nucleolytic proofreading Recent experimental studies on DNA transcription have shed light on the microscopic dynamics of backtracking [47, 124, 147, 159] enabling the construction of predictive models of transcriptional error correction mechanism.

## 5.3 Mechanism of Transcriptional Error Correction

The low error rates ($10^{-5}$ errors/nt) accomplished by the RNAP can be attributed to at least two distinct proofreading mechanisms. The first mechanisms acts at the level of nucleotide addition and is similar to the mechanism of kinetic proofreading discussed in the preceding section. In particular, the mechanism relies on the existence of a high energy intermediate along the polymerization pathway, which acts as a fidelity checkpoint and enhances the discriminatory power of the RNAP [143]. We shall refer to this mechanism as *classical proofreading* (CP).

*Nucleolytic proofreading* (NP) on the other hand is mediated through RNAP backtracking and the nuclease character of the RNAP [4, 58], *i.e.,* the ability of the active site of the polymerase to induce cleavage of the nascent RNA [58]. As we have seen in Chapter 4 (4.2.3), during backtracking the transcription elongation complex slides backwards along the DNA template [58]. Being relocated away from the $3'$ end of the nascent RNA, the active site can now exert its nucleolytic function and cleave the RNA chain. In general, different RNA pols demonstrate different endonuclease activities [138] and in certain cases accessory proteins (such as Gre, TFIIS) are necessary to stimulate RNA cleavage [45]. However, how does the RNAP manage to cleave at the right place, achieving discrimination between correct and incorrect nucleotides? We propose that the answer lies in the different translocation rates that are imposed by the presence or absence of an erroneous nucleotide. In particular, the presence of an error will cause the the RNAP to stagger making the catalysis of RNA cleavage and therefore excision of the erroneous nucleotide more probable.

## 5.4 Model of Nucleolytic Proofreading

In this section we present and study a stochastic model of the transcription elongation phase involving polymerization of correct and incorrect nucleotides, backtracking, and

RNA cleavage. The model is an extension of the one presented in Chapter 4 (4.2), and aims at capturing the essence of NP.

### 5.4.1 Basic Notation

Transcription elongation can be described in terms of two variables, $n$ and $m$. Variable $n = 0, \ldots, N$ denotes the template position of the last transcribed nucleotide or equivalently the length of the mRNA transcript up to a small offset. In particular, we define $n = 0$ to be the position at which the elongation phase is entered, by the formation of the TEC a few (8-10) nucleotides downstream of the actual transcriptional starting point. Position $n = N$ corresponds to the terminal position up to this offset.

On the other hand, variable $m = 0, \ldots, M$ marks the physical position of the TEC along the DNA template, and in particular the position of the polymerase active site relative to $n$. Here, $m = 0$ indicates that the TEC is in the active state, [1] where polymerization of the next nucleotide can occur, while $m > 0$ indicates that the TEC is in a backtracked state [see Fig. 5.1(a)]. Since extensive backtracking is often blocked by hairpins or other secondary RNA structures that are formed as the RNA exits the TEC [58], we assume that backtracking is restricted to a fixed distance $m = M$, which we take to be independent of $n$.[2] The process starts with the TEC at $(n = 0, m = 0)$ and terminates upon reaching state $(n = N, m = 0)$.

Given a TEC in an active state $(n, m = 0)$, the TEC can either backtrack to state $(n, m = 1)$ with rate $k_b$ or polymerize the next nucleotide $(n + 1, m = 0)$. Polymerization of correct nucleotides occurs with rate $k_p$, while incorrect nucleotides are polymerized with rate $\bar{k}_p$. We use $\epsilon$ to denote the *spontaneous* error fraction, *i.e.*, the fraction of thermally induced errors

$$\epsilon \;=\; \frac{\bar{k}_p}{k_p} \;\Rightarrow\; \bar{k}_p \;=\; \epsilon k_p. \tag{5.4}$$

Once backtracked the TEC hops between contiguous translocation states, $(n, 0 < m \leq M)$ with rate $c$, except when the TEC hops into an error site $m = l$ from a deeper backtracked state $l + 1$ which occurs with a reduced rate $\bar{c}$ (see Fig. 5.1). Finally, from each backtracked state, $(n, m = m^* > 0)$, cleavage occurs at rate $k_c$, removing the last $m^* - 1$ nucleotides from the RNA chain and leaving the TEC in state $(n - m^*, m = 0)$.

[1] Unlike the model presented in Chapter 4, the model here does not consider pre- and post-translocated states. Rather, for the sake of simplicity, these two states have been lumped together into a single state under the assumption that equilibrium is readily achieved between them.

[2] For positions $n < M$ we assume that backtracking is restricted to $m = n$.

Therefore, given an erroneous nucleotide at some position $n-l$ ($l \geq 0$), cleavage from any state $(n, m > l)$ ensures its removal. Note that the difference in the hopping rates ($\bar{c} < c$) at an error site is the key ingredient of error correction since it increases the likelihood of cleavage at states $(n, m > l)$. A schematic diagram of state transitions for the model is given in Fig. 5.1(b).

### 5.4.2 Physical Picture

As discussed in Chapter 4 (4.2.4) backtracked states correspond to wells in the free-energy landscape that transiently trap the diffusional motion of the TEC along the DNA template. The depth of these wells is dictated by the interactions between the RNAP, the DNA and the RNA transcript that contribute to the structural stability of the TEC, with the RNA-DNA hybrid being a major contributor. In the absence of any errors along the RNA-DNA hybrid, our model assumes a periodic free-energy landscape that gives rise to a constant hopping rate $c$ [see Fig. 5.1(c), right panel]. On the other hand, the presence of an erroneous nucleotide along the RNA-DNA hybrid partially destabilises the TEC, *i.e.*, increases the free-energy. This increase in the free-energy, $\Delta G$, is due to the mispairing between the erroneous RNA nucleotide and its corresponding DNA nucleotide and should, therefore, also be approximately equal to the free-energy difference dictating the spontaneous error fraction $\epsilon$. As the TEC backtracks past the error site the erroneous nucleotide diassociates from the RNA-DNA hybrid and therefore the hybrid becomes error-free once again. This leads to a drop in the free-energy by $\Delta G$. Now, to reincorporate the erroneous nucleotide into the RNA-DNA hybrid the TEC needs to overcome an enhanced energetic barrier, which gives rise to a slower hopping rate $\bar{c}$. Other than this decrease (increase) in the free-energy as the erroneous nucleotide exits (re-enters) the RNA-DNA hybrid we assume that free-energy landscape remains qualitatively unchanged, that is remains periodic [see Fig. 5.1(c), left panel]. According to Kramer's rate theory [141] the ratio of the two hopping rates is given by

$$\frac{\bar{c}}{c} \approx \exp\left[-\Delta G/k_B T\right] \approx \epsilon. \tag{5.5}$$

As we have seen in Section 5.2 kinetic proofreading captures the essence of error correction by stipulating the existence of one or more non-equilibrium (irreversible) intermediate steps in the catalytic process. In our model these intermediate steps that dissipate energy are the successive polymerisation events that add nucleotides on the RNA chain and push already incorporated ones pass the backtracking limit $M$ (where cleavage is no longer possible).

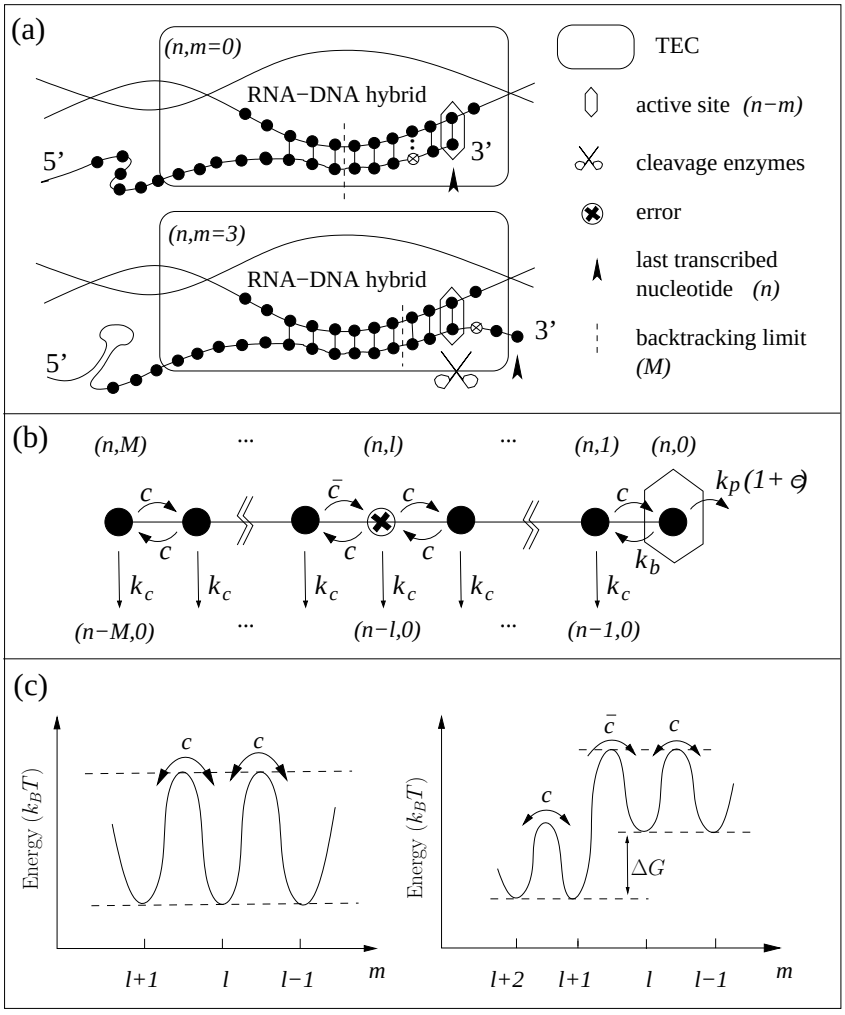



Figure 5.1: Schematic illustration of the model of transcriptional error correction. (a) Schematic illustration of the model. Variable $n$ denotes the position of the last transcribed nucleotide, whereas variable $m$ denotes the position of the polymerase's active site relative to $n$. The RNA is marked by $3'$ and $5'$. The transcription elongation complex (TEC) is depicted in the active $(n, m = 0)$ (top) and in a backtracked $(n, m = 3)$ (bottom) state. (b) Schematic illustration of the TEC dynamics at a given position $n$. The TEC will eventually polymerize forward or cleave from one of the backtracked states. The slow rate of hopping $\bar{c}$ into the error state $(n.m = l)$ increases the likelihood of cleavage from states $(n, m > l)$ and therefore the removal of the error. (c) Schematic illustration of the energy landscape driving backtracking dynamics in the presence or absence of an erroneous nucleotide. The presence of an error results in an increase of the free-energy by $\Delta G$.

### 5.4.3 Dynamics at the Single Nucleotide Level

For the analytic treatment of the model we first consider the dynamics of the process at a fixed template position $n$. Results obtained in this section we later on be used to construct an effective model of the full elongation process.

The stochastic dynamics of the TEC at a fixed position $n$ are described by the Master equation

$$\dot{\mathbf{P}}(t) \;=\; \mathbf{W}^{(s)} \cdot \mathbf{P}(t). \tag{5.6}$$

Here, $\mathbf{P}$ is a column vector of size $(M+1)$:

$$\mathbf{P}(t) \;=\; (P(m=0,t), P(m=1,t), \ldots, P(m=M,t))', \tag{5.7}$$

where prime ($'$) denotes transposition. Each element, $P(m,t)$, of the vector corresponds to the PDF of finding the TEC at translocation state $m$ at time $t$ having started from $m=0$ at $t=0$. $\mathbf{W}^{(s)}$ is the $(M+1)\times(M+1)$ transition matrix. Superscript $s$ denotes the dependence of the matrix on the sequence of the correct and incorrect nucleotides along the transcript. In particular, $s \in S^n$ with $S \equiv \{0,1\}$, is a binary list of 0s and 1s, which represent correct and erroneous nucleotides respectively:

$$s \;=\; \underbrace{\{0,\;\; 1,\;\; \ldots,\;\; 0\}}_{n \text{ elements}}$$

The general tridiagonal structure of $\mathbf{W}^{(s)}$, is given by

$$\begin{bmatrix}
-[(1+\epsilon)k_p+k_b] & c+s_n(\bar{c}-c) & 0 & \cdots & & & \\
k_b & -[2c+s_n(\bar{c}-c)+k_c] & & & & & \\
0 & c & \ddots & c+s_{n-j+2}(\bar{c}-c) & & & \\
\vdots & & & -[2c+s_{n-j+2}(\bar{c}-c)+k_c] & & & \\
 & & & c & \ddots & c+s_{n-M+1}(\bar{c}-c) & \\
 & & & & & -[c+s_{n-M+1}(\bar{c}-c)+k_c]
\end{bmatrix}.$$

Specifically, off-diagonal elements of the matrix correspond to transition probabilities between the different translocation states. In particular, the element of the matrix at the $k^{\text{th}}$ row and $j^{\text{th}}$ column, $W^{(s)}_{kj}$ ($k \neq j$), yields the transition rate from translocation state $m=j$ to $m=k$. On the other hand, elements in the diagonal of the matrix correspond to the total transition probabilities out of a state. Note that $\mathbf{W}^{(s)}$ depends only on the last

$M$ elements of $s$, *i.e.*, $s_n, s_{n-1}, \ldots, s_{n-M+1}$.

The model allows only transitions between contiguous translocation states, hence the matrix has non-zero elements only along the main diagonal and the first diagonals above and below the main one. However, the columns of $\mathbf{W}^{(s)}$ do not sum up to 1. This indicates that probability is lost to some absorbing boundaries. In fact, the above formulation of $\mathbf{W}^{(s)}$ implies $M+1$ absorbing boundaries, through which the TEC can leave template position $n$:

- Boundary $i = 0$: polymerization of the next nucleotide ($n \to n+1$) occurring from the active state $m = 0$.
- Boundaries $i = 1, \ldots, M+1$: cleavage of the transcript ($n \to n-i$) occurring from each backtrack state $m = i$.

As described in Chapter 3 (3.4) , by applying the Laplace transform $\tilde{\mathbf{P}}(z) = \int_0^\infty e^{-zt}\mathbf{P}(t)dt$ to Eq. (5.6), we obtain a system of algebraic equations, which can be solved for all $\tilde{P}(m,z)$ ($m = 0, \ldots, M$). Subsequently, the splitting probabilities $p_i$ for eventually hitting boundary $i$ as well as the corresponding conditional mean exit times, $\tau_i$ can be obtained using the Laplace transform of the probability fluxes to each boundary [116]:

$$p_0 = (1+\epsilon)k_p\tilde{P}(0, z=0); \quad \tau_0 = (1+\epsilon)k_p\frac{\tilde{P}'(0, z=0)}{\tilde{P}(0, z=0)}, \tag{5.8a}$$

$$p_i = k_c\tilde{P}(i, z=0); \quad \tau_i = k_c\frac{\tilde{P}'(i, z=0)}{\tilde{P}(i, z=0)}, \quad i = 1, \ldots, M \tag{5.8b}$$

Note that $p_i$ and $t_i$ will depend on the sequence of correct and incorrect nucleotides, $s$. In the following the notation $p_i(s)$ and $\tau_i(s)$ will be used to make this dependence explicit.

### 5.4.4 Effective Model of the Elongation Dynamics

So far we have formulated the stochastic dynamics of the TEC at fixed nucleotide position $n$. Here, we present how an *effective* model of overall elongation dynamics can be constructed for times $\tau$ much longer than the typical dwell time at each position, *i.e.*, $\tau \gg \tau_i (0 \leq i \leq M)$. [3]

At the coarse-grained time-scale $\tau$ one observes the TEC polymerising and cleaving the RNA transcript at rates which are proportional to the splitting probabilities $p_i$ obtained above. Let $\Pi(n, s, \tau)$ be the probability of finding the TEC at position $n$ at time $t$ having

[3] We note that all results obtained below do not depend on the exact definition of $\tau$.

produced a transcript $s \in S^n$. From each position $n$, the TEC can either polymerize or cleave the RNA transcript with rates $r_i$ $(i = 0, \ldots, M)$ given by

$$r_0(s) \;=\; p_0(s)/\tau : \;\; n \to n+1 \quad \text{(polymerisation)}, \tag{5.9a}$$

$$r_i(s) \;=\; p_i(s)/\tau : \;\; n \to n-i \quad \text{(cleavage)}. \tag{5.9b}$$

Summing $\Pi(n, s, t)$ over all possible configurations of $s$, one obtains the probability of finding the TEC at position $n$ at time $t$, irrespective of the transcript sequence:

$$\Pi(n, t) \;=\; \sum_{s \in S^n} \Pi(n, s, t). \tag{5.10}$$

The stochastic dynamics of $\Pi(n, t)$ can therefore be expressed as

$$\frac{d\Pi(n,t)}{dt} \;=\; J(n-1|n) - J(n|n+1) + \sum_{i=1}^{M} \left[ J(n+i|n) - J(n|n-i) \right], \tag{5.11}$$

where $J(n_1|n_2)$ denotes the probability flux from $n_1$ to $n_2$. In particular one has

$$lllJ(n_1|n_1+1) \;=\; \sum_{s \in S^{n_1}} r_0(s)\Pi(n_1, s, t), \tag{5.12a}$$

$$J(n_1+i|n_1) \;=\; \sum_{s \in S^{n_1+i}} r_i(s)\Pi(n_1 - i, s, t). \tag{5.12b}$$

The process starts at $n = 0$ and is terminated when position $n = N$ has been reached. We therefore impose the boundary conditions $J(0|-1) = J(-1|0)$ (reflecting) and $J(N|N-1) = 0$ (absorbing).

In the following, Eq. (5.11) will be used to obtain an expression for $\mathcal{P}_n$, $\bar{\mathcal{P}}_n$, the probability of reaching the terminal position ($n = N$), having incorporated a correct or an incorrect nucleotide at position $n$, and irrespective of the rest of the sequence. We use $\mathcal{P}_n$ and $\bar{\mathcal{P}}_n$ to quantify the transcriptional fidelity in terms of the error fraction, defined at each position $n$ as [68, 101]:

$$\mathcal{E}_n \;\equiv\; \frac{\bar{\mathcal{P}}_n}{\mathcal{P}_n}. \tag{5.13}$$

### 5.4.5 Analytic Results

Here, for the sake of simplicity, we present a detailed treatment of $M = 1$ case. The generalised results for $M > 1$ are then presented and discussed (for a detailed derivation

see Appendix A).

Most of the results presented below are given in terms of following dimensionless quantities, which characterise the competing processes in the model:

- $\alpha_1 \equiv k_c/c$ captures the efficiency of cleavage for correct nucleotides
- $\alpha_2 \equiv k_c/\bar{c} = \alpha_1/\epsilon$ captures the efficiency of cleavage for incorrect nucleotides
- $K \equiv k_p/k_b$ captures the tendency of the TEC to backtrack

$M = 1$ **case**

In this case the TEC can backtrack by only one nucleotide. Therefore, the transition matrix, $\mathbf{W}^{(s)}$ in Eq. (5.6) will depend solely on whether the last nucleotide has been correctly or incorrectly transcribed. In particular one has

$$\mathbf{W}^{(s^c)} = \begin{bmatrix} -[(1+\epsilon)k_p + k_b] & c \\ k_b & -(c + k_c) \end{bmatrix}, \tag{5.14a}$$

$$\mathbf{W}^{(s^w)} = \begin{bmatrix} -[(1+\epsilon)k_p + k_b] & \bar{c} \\ k_b & -(\bar{c} + k_c) \end{bmatrix}. \tag{5.14b}$$

where we have used the notation $s^c = (\ldots, 0)$ and $s^w = (\ldots, 1)$ to denote transcripts whose last nucleotide has been correctly and incorrectly transcribed, respectively. Applying the Laplace transform, $\tilde{\mathbf{P}}(z) = \int_0^\infty e^{-zt}\mathbf{P}(t)dt$, on Eq. (5.6) and evaluating at $z = 0$ one can obtain the splitting probabilities $p_i \equiv p_i(s^c)$ and $\bar{p}_i \equiv p_i(s^w)$:

$$p_0 = \frac{K(1+\epsilon)(1+\alpha_1)}{K(1+\epsilon)(1+\alpha_1)+\alpha_1}; \quad p_1 = 1 - p_0, \tag{5.15a}$$

$$\bar{p}_0 = \frac{K(1+\epsilon)(1+\alpha_2)}{K(1+\epsilon)(1+\alpha_2)+\alpha_2}; \quad \bar{p}_1 = 1 - \bar{p}_0, \tag{5.15b}$$

where $p_0$, $\bar{p}_0$ correspond to the polymerisation and $p_1$, $\bar{p}_1$ to cleavage.

The splitting probabilities divided by $\tau$ yield the effective rates, $r_i$ and $\bar{r}_i$ ($i = 0, 1$), in Eq. (5.11) (for $M = 1$). The process starts at position $n = 0$ and is terminated when state $n = N$ has been reached. To calculate the probability that the terminal position $n = N$ is reached with a correct or incorrect nucleotide incorporated at position $n = n'$ we break the domain of the process into 3 regions, namely

- Region $R_-$: $n = 0, \ldots, n' - 1$,
- Region $R_0$: $n = n'$,

- Region $R_+$: $n = n' + 1, \ldots, N - 1$.

The process enters region $R_0$ when a nucleotide is polymerised at position $n = n'$. In particular, the probability flux from $R_-$ to $R_0$ is given by

$$J(R_-|R_0) = \sum_{s \in S^{n-1}} r_0(s)\Pi(n-1, s, t). \tag{5.16}$$

This polymerisation event will result in either a correct or an incorrect nucleotide at position $n'$, This gives rise to two independent branches in the process, the "correct" and the "erroneous" one. Hence, the reverse probability flux, from $R_0$ to $R_-$, will be through both of these branches, *i.e.*

$$\begin{aligned} J(R_0|R_-) &= r_1\Pi(n, s^c, t) + \bar{r}_1\Pi(n, s^w, t), \\ &\equiv J^c(R_0|R_-) + J^w(R_0|R_-). \end{aligned} \tag{5.17}$$

The two branches evolve independently of one another and will lead to probability flowing into region $R_+$:

$$\begin{aligned} J(R_0|R_+) &= r_0\Pi(n, s^c, t) + \bar{r}_0\Pi(n, s^w, t), \\ &\equiv J^c(R_0|R_+) + J^w(R_0|R_+). \end{aligned} \tag{5.18}$$

Of course when the process enters region $R_+$ it branches once again. However, the total probability entering $R_+$ should be conserved, either flowing back to $R_0$ or to the absorbing boundary $n = N$. This allows us to write

$$J(R_+|R_0) = J^c(R_+|R_0) + J^w(R_+|R_0), \tag{5.19a}$$

$$J(R_+|N) = J^c(R_+|N) + J^w(R_+|N). \tag{5.19b}$$

In the long time limit $t \to \infty$ the fluxes in and out of the different regions will balance and a steady probability flow towards the terminal position $n = N$ will be achieved. Applying the Laplace transform $\tilde{\Pi}(n, s, z) = \int_0^\infty e^{-zt}\Pi(n, s, t)dt$ on Eq. (5.11), summing over the three regions of interest ($R_-, R_0, R_+$) and evaluating at $z = 0$ one can obtains a

system of equations relating the Laplace transform of the aforementioned fluxes:

$$\begin{aligned}
\tilde{J}^c(R_0|R_-) + \tilde{J}^w(R_0|R_-) - \tilde{J}(R_-|R_0) + 1 &= 0, && \text{(5.20a)}\\
\frac{\epsilon}{\epsilon+1}\tilde{J}(R_-|R_0) - \tilde{J}^w(R_0|R_-) + \tilde{J}^w(R_+|R_0) - \tilde{J}^w(R_0|R_+) &= 0, && \text{(5.20b)}\\
\frac{1}{\epsilon+1}\tilde{J}(R_-|R_0) - \tilde{J}^c(R_0|R_-) + \tilde{J}^c(R_+|R_0) - \tilde{J}^c(R_0|R_+) &= 0, && \text{(5.20c)}\\
\tilde{J}^w(R_0|R_+) - \tilde{J}^w(R_+|R_0) - \tilde{J}^w(R_+|N) &= 0, && \text{(5.20d)}\\
\tilde{J}^c(R_0|R_+) - \tilde{J}^c(R_+|R_0) - \tilde{J}^c(R_+|N) &= 0, && \text{(5.20e)}
\end{aligned}$$

where the notation $\tilde{J}$ is used to denote the Laplace transform of the corresponding probability flux evaluated at $z = 0$. All of these quantities have probability status [116]. Note, for example that in the last line terms $\tilde{J}^c(R_+|R_N)$ and $\tilde{J}^c(R_+|R_0)$ up to division by $\tilde{J}^c(R_0|R_+)$ can be interpreted as splitting probabilities; some probability $\tilde{J}^c(R_0|R_+)$ is injected into $R_+$ (through the "correct" branch) and subsequently divided among the 2 boundaries, $n = N$ and $n = n'$. More importantly, the division does not depend through which of the two branches the probability ends up in region $R_+$. This consideration allows us to write

$$\begin{aligned}
\tilde{J}^c(R_+|N) &= A_T\tilde{J}^c(R_0|R_+) &&= A_T r_0 \Pi(n, s^c, t),\\
\tilde{J}^c(R_+|R_0) &= A_{n'}\tilde{J}^c(R_0|R_+) &&= A_{n'} r_1 \Pi(n, s^c, t),\\
\tilde{J}^w(R_+|N) &= A_T\tilde{J}^w(R_0|R_+) &&= A_T \bar{r}_0 \Pi(n, s^w, t),\\
\tilde{J}^w(R_+|R_0) &= A_{n'}\tilde{J}^w(R_0|R_+) &&= A_{n'} \bar{r}_1 \Pi(n, s^w, t),
\end{aligned} \qquad \text{(5.21)}$$

subject to the condition

$$A_T + A_{n'} = 1. \qquad \text{(5.22)}$$

Substituting the relationships given by Eq. 5.17, 5.18, and 5.21 into the system of equations one can obtain an expression for the probabilities of interest: $\mathcal{P}_{n'}$ and $\bar{\mathcal{P}}_{n'}$:

$$\mathcal{P}_{n'} = \tilde{J}^c(R_+|N) = \frac{1}{\mathcal{N}}\frac{p_0}{1 - A_{n'}p_0}, \qquad \text{(5.23a)}$$

$$\bar{\mathcal{P}}_{n'} = \tilde{J}^w(R_+|N) = \frac{\epsilon}{\mathcal{N}}\frac{\bar{p}_0}{1 - A_{n'}\bar{p}_0}. \qquad \text{(5.23b)}$$

Here, $\mathcal{N}$ can be obtained from the normalisation condition $\mathcal{P}_n + \bar{\mathcal{P}}_n = 1$ and $A_{n'}$ corresponds to the probability that starting from $n = n' + 1$ cleavage to position $n = n'$ will occur prior to termination. An expression for $A_n$ can be obtained by initialising the process at $n = n' + 1$, and regarding the process bounded in $R_+$, with $R_0$ and $n = N$ being

absorbing boundaries [see Chapter 3 (3.3.1]. This yield the general recursion formula

$$A_n = \frac{\epsilon}{(\epsilon+1)}\frac{\bar{p}_1}{(1-\bar{p}_0 A_{n+1})} + \frac{1}{(\epsilon+1)}\frac{p_1}{(1-p_0 A_{n+1})}, \tag{5.24}$$

with boundary condition $A_N = 0$. In the limit $\epsilon \to 0$, the above reduces to

$$A_n \approx \frac{\beta(\beta^{N-n}-1)}{(\beta^{N-n+1}-1)}, \tag{5.25}$$

where $\beta = p_1/p_0$.

Finally, the error fraction at any position $n$ for $M = 1$ is given by

$$\mathcal{E}_n \equiv \frac{\bar{\mathcal{P}}_n}{\mathcal{P}_n} = \frac{\epsilon\bar{p}_0}{p_0}\frac{1-A_n p_0}{1-A_n\bar{p}_0}. \tag{5.26}$$

Fig. 5.2 (top panel) shows $\mathcal{E}_n$ as a function of $K$, for different positions along the template.

Let us now consider two limits where $\mathcal{E}$ attains a constant value independent of position $n$. First we examine the limit $K \gg 1$, where polymerization is overwhelmingly favored over cleavage ($p_0 \to 1$ and $\bar{p}_0 \to 1$). As expected, in this limit Eq. (5.26) reduces to $\mathcal{E} \approx \epsilon$. On the other hand, in the limit $K \ll \alpha_1 \ll \epsilon$, cleavage events dominate the process. In this regime Eq. (5.26) reduces to $\mathcal{E} \approx \epsilon\bar{p}_0/p_0$, or, in terms of the microscopic rate parameters

$$\mathcal{E} \approx \epsilon \cdot \frac{\bar{c}}{c}. \tag{5.27}$$

Hence, in this limit the error fraction depends only on $\epsilon$ and the ratio of hopping rates. Since we take these two quantities to be approximately equal, we have $\mathcal{E} \approx \epsilon^2$.

### $M > 1$ case

For the more general case $M \geq 1$ similar results can be obtained in the limit $\epsilon \ll 1/M$, *i.e.*, at most one error can occur in a region of $M$ nucleotides. In particular, it can be shown that in the same limit ($K \ll \alpha_1 \ll \epsilon$) the error fraction is given by

$$\mathcal{E} \approx \epsilon^{M+1} \cdot \frac{M^M}{\Gamma(M+1)}, \tag{5.28}$$

where $\Gamma$ denotes the Gamma function. Thus, the combined action of backtracking and cleavage can result in error rates that scale exponentially with $M$, the maximum backtracking distance. We note that the error fraction attained by KP with $M$ intermediate

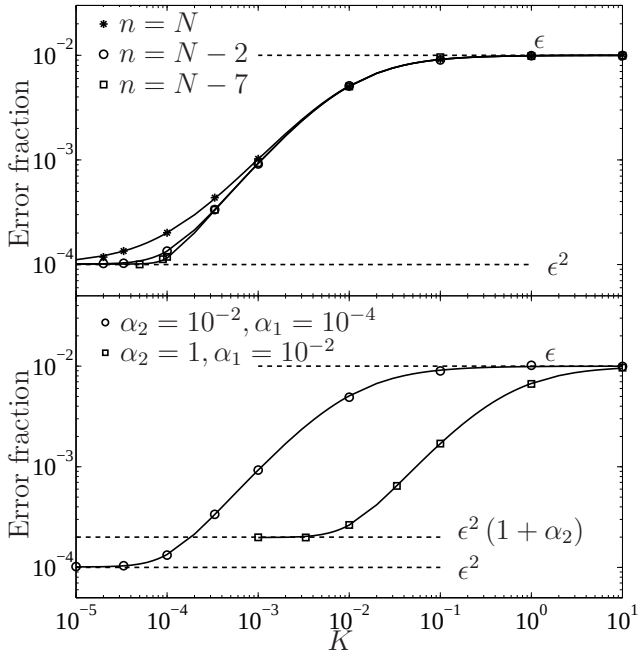



Figure 5.2: Error fraction ($\mathcal{E}_n$) as a function of $K$ for $M = 1$. Analytic results [Eq. (5.26)] are plotted as solid lines and markers show results obtained from stochastic simulations: (top) $\mathcal{E}_n$ at different positions for $\alpha_1 = 10^{-4}$, $\alpha_2 = 10^{-2}$, $\epsilon = 10^{-2}$ and $N = 9$. (bottom) $\mathcal{E}_n$ for different cleavage efficiencies, $\alpha_1$ and $\alpha_2$ at position $n = N - 2$, with $\epsilon = 10^{-2}$ and $N = 4$. Dashed lines show limits discussed in text.

steps has a similar $M$ dependence [68]. The two limits discussed above are illustrated in Fig. 5.2 (bottom panel). Numerical data were generated using stochastic simulations [52] of the full transcription elongation model.

### 5.4.6 An Estimating the Error Fraction

Estimates of the error fractions predicted by our model can be obtained by taking into account information from experimental studies. First of all, the spontaneous error fraction $\epsilon$ can be calculated from the free energy difference due to a misincorporated nucleotide ($\Delta G \approx 4 - 7 k_B T$), *i.e.*, $\epsilon \approx e^{-\Delta G/k_B T} \approx 10^{-2} - 10^{-3}$ [18]. The cleavage rate, $k_c$, for bacterial RNAP was measured $k_c \approx 0.1 - 1\text{s}^{-1}$ in the presence of saturating concentrations of accessory cleavage factors [128]. Moreover, single molecule experiments have suggested that the TEC hops between backtrack states with rate $c \approx 1 - 10\ \text{s}^{-1}$ [47, 124]. Using estimates of the maximum spontaneous error fraction $\epsilon = 0.01$, slowest cleavage rate $k_c = 0.1\text{s}^{-1}$ and fastest hopping rate $c = 1\text{s}^{-1}$ we can obtain estimates of the lower bounds on cleavage efficiencies $\alpha_1 \approx 10^{-2}$ and $\alpha_2 \approx 1$. These estimates yield error fractions comparable to the ones observed *in vivo* ($10^{-4} - 10^{-5}$), even for $M = 1$ but

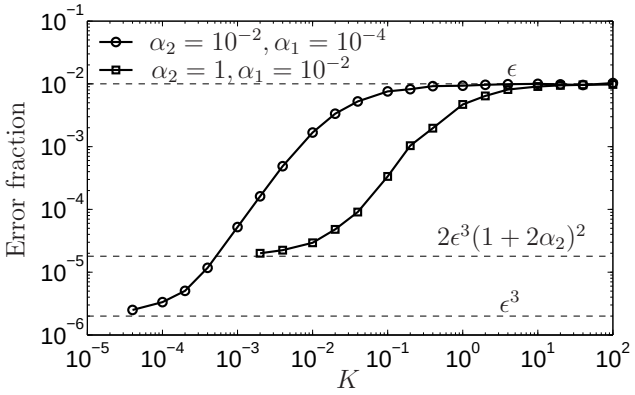



Figure 5.3: Error fraction as a function of $K$ for $M = 2$. Results were obtained using stochastic simulations of the model for $N = 4$, $\epsilon = 10^{-2}$, and $\alpha_1 = 10^{-2}, 10^{-4}$.

sufficiently low values of $K$ (see Fig. 5.2 bottom panel). Most importantly, however, low error fractions can be obtained in our model even well away from the limiting regime with small $M$ (see Fig. 5.3 for $M = 2$ case).

### 5.4.7 Some Key Notes on the Model

We should note that certain simplifications were made in the model that do not however alter the essence of the results. In particular, depolymerisation as well as the dependence of the microscopic rates on the sequence composition were neglected. Interestingly, sequence heterogeneity can affect transcriptional fidelity. For example, GC rich domains can lead to slower backtracking rates (due to the increased stability of the RNA-DNA hybrid) [5]. Our model then predicts that the slower backtracking dynamics imposed by the sequence will slightly reduce the efficiency of the error correction.

Also, alternative formulations of the model are possible depending on which step is assumed to provide the discriminatory power to the process. In our current formulation, discrimination between correct and incorrect nucleotides is solely provided during backtracking, where hopping back into an error site occurs at a much slower rate, $\bar{c} \ll c$. A more general formulation (see Fig. 5.4), which yields however quantitatively similar results, involves:

1. a fast rate of backtracking $\bar{k}_b$ in the presence of an misincorporated nucleotide at position $n$ as compared to $k_b$ in the presence of a correct one.

2. a fast hopping rate $c_f$ ($c < c_f$) from state $(n, m = l)$ (error site) into state $(n, m = l + 1)$

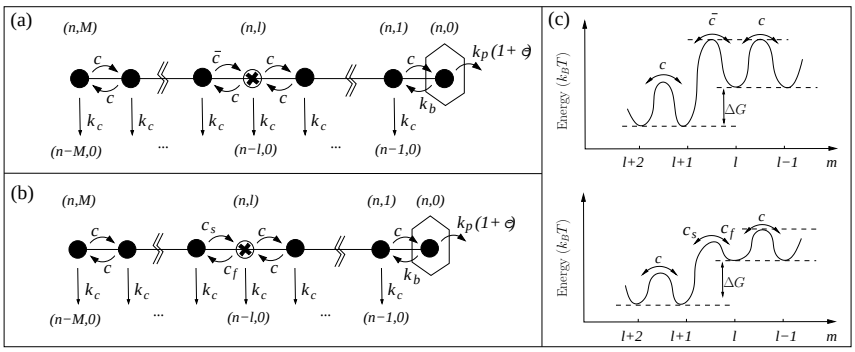



Figure 5.4: Schematic illustration of an alternative formulation of the error correction model. In the alternative formulation of the model discrimination between correct and incorrect nucleotides is achieved not only during backtracking, owing to the differential hopping rates, but also at the active state, where TECs with an misincorporated nucleotide at register would tend to backtrack more often.

3. a slow hopping rate $c_s$ ($c > c_s$) from state $(n, m = l + 1)$ into state $(n, m = l)$ (error site)

As before, one expects that the ratio of the above rates would be approximately equal to the spontaneous error fraction $\epsilon$ since all processes are driven by approximately the same free energy difference $\Delta G$

$$\frac{c_s}{c_f} \approx \frac{k_b}{\bar{k}_b}\frac{c_s}{c} \approx \epsilon \approx \exp\left[-\Delta G/(k_B T)\right]. \tag{5.29}$$

In this scenario discrimination would not only occur during backtracking, owing to the differential hopping rates, but also at the active state, where TECs with an misincorporated nucleotide at register would tend to backtrack more often.

### 5.4.8 Numerical Methods

To validate the analytic results obtained, we performed stochastic simulations of the full transcription elongation model using the Gillespie algorithm [17,52]. Each simulation run started with the TEC in state $(n = 0, m = 0)$ and terminated when state $(n = N, m = 0)$ had been reached. From each state, the next state was randomly selected among all accessible states with a probability proportional to the corresponding transition rate. The sequence $s$ was implemented as a binary list and used at each step of the algorithm to assign the correct transition rate to each accessible state. Whereas polymerization corresponds to the addition of an element (0 or 1) to the list, cleavage, $(n, m = l) \rightarrow (n - l, m = 0)$, corresponds to the removal of the last $l$ elements of the list. The sequence $s$ was saved at

the end of each simulation run.

For each set of parameters, simulations of the model were repeated until at least $100$ errors had been observed for each position. The error fraction at each position was then calculated as

$$\text{Error fraction} \quad = \quad \frac{\#\text{ incorrect nucleotides } (1s)}{\#\text{ correct nucleotides } (0s)}. \tag{5.30}$$

## 5.5 Summary and Discussion

In this Chapter we presented and studied a microscopic model of a transcriptional error correction mechanism involving RNAP backtracking and RNA cleavage. Our model incorporates polymerisation of correct and incorrect nucleotides, RNAP backtracking and RNA cleavage. In analogy with kinetic proofreading, in our model backtracking provides a multiple-checking reaction, which probes the fidelity of the last few nucleotides several times before the next polymerization occurs. In fact, the greater the delay introduced by this mechanism, the greater the accuracy of the process [68, 101]. Consistent with this picture we find a minimum error fraction, which scales exponentially with the maximum backtracking distance $M$, in the limit where backtracking and cleavage dominate the process.

Recent experiments have provided support for at least two mechanisms of transcriptional error correction. The first one involves a fidelity checkpoint during the nucleotide addition cycle [143], whereas the second involves backtracking of the RNAP and RNA cleavage [4, 124, 143, 147, 159]. Our model suggests experiments that would provide the quantitative details required to discriminate between these mechanisms and elucidate their relative roles in transcriptional proofreading.

A particular prediction of our model is the strong dependence of transcriptional fidelity on the translocation rates. For example, GC rich domains that lead to lower backtracking rates (due to the increased stability of the RNA-DNA hybrid) [5] should reduce the efficiency of error correction. More importantly, single molecule manipulation techniques can be used to vary backtracking rates in a controlled manner and validate our model. In particular, applying a load is expected to strongly affect nucleolytic proofreading since the TEC moves at least a distance $\sim \delta x$ (where $\delta x = 3.4$Å) during the backtracking phase. In contrast, minor effects are expected for proofreading mechanisms along the polymerization pathway, since they should only involve small movements ($\ll \delta x$) of the enzyme.

Our model also predicts that RNAP species with a greater tendency to backtrack should accomplish lower error rates. Experimental studies have already revealed that

specific mutations in the sequence of RNAP can have profound effects on transcriptional fidelity [66]. It is therefore particularly interesting to study exactly how these mutations affect transcriptional accuracy and whether these effects are mediated through changes in the rates of backtracking or translocation rates.

# Chapter 6

# Cell Level: The Stochastic Nature of RNA Production

In the preceding two chapters we have focused on the single molecule dynamics of the transcription elongation phase. Ultimately, however, one is interested in the process of DNA transcription as a whole and the dynamics of RNA production. The aim of this Chapter is to bridge these two levels of description by providing an integrated picture of DNA transcription and characterising how the underlying microscopic dynamics of the process affect the cellular levels of RNA. To do so we formulate a multistep, coarse grained model of DNA transcription and using stochastic simulations, we examine the statistics of RNA production in relation to transcriptional pausing. In particular, we find that long-lived elongation pauses can lead to bursts of RNA production and non-Poisson RNA statistics. Our results have direct implications for *in-vivo* transcription since they provide a microscopic mechanism for transcriptional bursts that have been observed experimentally.

## 6.1 Introduction

It has long been appreciated that life at the cellular level is noisy [122]. Indeed, all cellular processes rely on random encounters between bio-molecules and are therefore discrete and inherently stochastic in nature. This consideration along with the fact that that most

molecular species are only present in small numbers within cells constitutes stochasticity a major player at the cellular level. However, it has only been with recent advancements in experimental techniques that a more quantitative description of cellular processes has become possible [70,115,154]. In particular, the advent of fluorescence techniques, allowing to track levels of chemical species within cells, renewed the interest in the stochastic nature of cellular processes and its consequences [83].

This experimental endeavour has largely been complemented by mathematical and computational models that take the apparent stochasticity into account [107, 133]. Such models are essential not only for interpreting experimental data but also for providing fresh insight into the processes that underpin life. However, any attempt of mathematical or computational modelling is severely hindered by the inherent complexity of life processes and our limited knowledge. To be useful and instructive, therefore, models have to rely on certain assumptions regarding which are the critical aspects of the process considered and which can safely be neglected. The validity of these assumptions is ensured through experimental studies that ultimately verify or disprove the predictions made by different models.

Of particular importance is understanding the stochastic nature of gene expression and gene regulation. These processes underlie every aspect of the cell and therefore their stochastic dynamics ought to have the most direct implications regarding cell behaviour and fate [25, 72, 83]. One of the major assumptions behind standard models of gene expression and gene regulation is the Poisson character of the steps involved [107, 133]. For example, transcription is usually described as a single-step reaction occurring at a constant rate. However, as we have seen in the previous chapters, this is roughly the case. In particular, transcription as well as translation are in themselves multi-step processes involving initiation, elongation and termination. Most importantly, these processes can exhibit biochemical fluctuations at each of these stages due to their complex microscopic dynamics and cannot in general be described as simple Poisson processes. Several question therefore arise concerning such simplifications. Under what conditions are they valid? Are we missing key aspects of gene expression by ignoring the microscopic dynamics of the processes involved?

Such questions become even more relevant in the light of recent experimental observations that highlight the non-Poisson character of DNA transcription. Utilising artificial reporter genes, which give rise to mRNA chains carrying several binding sites for fluorescently labeled probes (see Fig. 6.1), experimental studies [27,55,114] suceeded in tracking mRNA levels within living cells with single molecule resolution. The key finding of these studies was that mRNA production both in bacterial and eukaryotic cells occurs in bursts.

In particular, Golding *et al.* [55] obsereved intense periods of rapid mRNA production followed by periods of transcriptional inactivity (see Fig. 6.1). This mode of mRNA production gives rise to enhanced variability in the mRNA levels and cannot be captured by simple Poisson models of transcription.

The aim of this Chapter is to provide a quantitative picture of how the microscopic dynamics of DNA transcription affect gene expression and in particular RNA production. The remainder of the Chapter is organised as follows. We first give a brief overview of a simple model that has found wide appreciation in describing stochastic gene expression. We mainly focus on assumptions underlying the model as well as the predictions the model makes. We then motivate the need for more detailed picture of DNA transcription by considering a model that incorporates the microscopic elongation dynamics discussed in Chapter 4. Finally, we present a coarse grained model of DNA transcription involving elongation pauses. Using stochastic simulations of the model we examine the effect that the microscopic dynamics of the process (*i.e.*, pausing) have on the statistics of mRNA production. Our results indicate that long-lived elongation pauses can play a significant role in the fluctuations of RNA species leading to bursts of RNA production and non-Poisson RNA statistics.

## 6.2 Standard Models of Stochastic Gene Expression

In this section we present a simple model of stochastic gene expression. The model, hereafter referred as the *standard model* (SM), captures the apparent stochasticity of gene expression by considering the random birth and death of RNA and protein molecules [107, 133]. Effectively, SM coarse grains all processes involved into elementary reactions obeying Poisson statistics. Despite its simplicity, SM (and its different variants) have been successfully used to interpret experimental data and to provide a first handle of the stochastic nature of gene expression [42, 109, 158]. However, SM relies on certain assumptions that limit its validity. We discuss some of these assumptions and motivate the need for more detailed, microscopically grounded, models, especially for the case of DNA transcription.

### 6.2.1 Mathematical Formulation

As described in chapter 2, at a coarse grained level, the expression of a protein-coding gene can be considered as two-step process involving (i) transcription and (ii) translation. During transcription mRNA molecules are produced from the DNA. At the subsequent step of translation each mRNA molecule is used as a template for the production of pro-

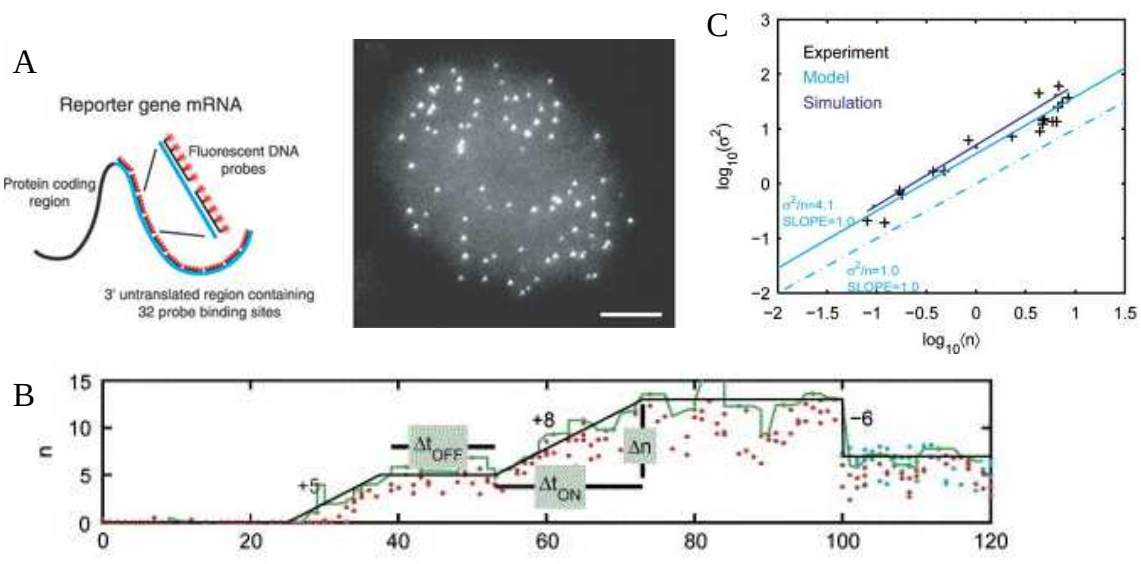



Figure 6.1: Experimental results demonstrating bursts of mRNA transcription. (A) Schematic illustration of the mRNA detection method. Multiple fluorescent labeled probes bind to each mRNA molecule, yielding a bright signal that enables detection of the mRNA. Reprinted from A. Raj *et al.*, PLOS Biol., **4** (2006). (B) Number of mRNA molecules $n$ per cell, as a function of time. Intense periods of mRNA production are followed by periods og transcriptional inactivity. Reprinted from I. Golding *et al.*, Cell, **123** (2005). (C) Variance ($\sigma^2$) versus average ($\langle n \rangle$) of mRNA numbers. The ratio $\sigma^2/\langle n \rangle = 4.1$ is significantly higher than that predicted from a simple Poisson model of transcription ($\sigma^2/\langle n \rangle = 1$). Reprinted by permission from Elsevier: I. Golding *et al.*, Cell, **123** (2005) Copyright(2005).

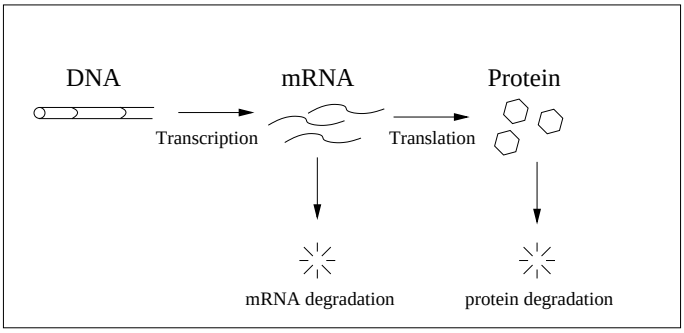



Figure 6.2: Schematic illustration of a simple model of gene expression. Transcription, translation as well as RNA and proteins degradation are captured as single-step reactions.

| Processes | Reaction representation | Probabilities |
|---|---|---|
| Transcription ($m \to m+1$) | $\varnothing \to$ mRNA | $k_m\ \Delta t$ |
| Translation ($n \to n+1$) | mRNA $\to$ mRNA + protein | $m\ k_p\ \Delta t$ |
| mRNA degradation ($m \to m-1$) | mRNA $\to \varnothing$ | $m\ d_m\ \Delta t$ |
| Protein degradation ($n \to n-1$) | protein $\to \varnothing$ | $p\ d_p\ \Delta t$ |

Table 6.1: Reactions involved in the standard model of stochastic gene expression.

teins. Of course, due to active cell processes or cell dilution mRNA and protein molecules are constantly lost. This simple picture, schematically illustrated in Fig. 6.2, sets the starting point for the formulation of SM.

Let us focus on gene expression dynamics for a single gene present on the DNA. The system consists of the mRNA and protein molecules produced from this gene, which we denote by $m$ and $p$, respectively. SM assumes that all processes have a constant probability of occurring over some time interval $\Delta t$ [108]. For example, transcription events, resulting in the production of mRNA ($m \to m+1$), occur with probability $k_m \Delta t$. Translation on the other hand, resulting in the production of proteins ($p \to p+1$), occurs with probability proportional to the number of mRNA molecules present, *i.e.*, $mk_p\Delta t$. Finally, degradation of mRNA ($m \to m-1$) and proteins ($p \to p-1$) occur with probabilities $md_m\Delta t$ and $pd_p\Delta t$, respectively. The reactions involved in the SM are summarised in Table 6.1.

Implicit in the above picture is the Markovian assumption. In particular, SM describes the evolution of the system at a coarse-grained time-scale $\Delta t$ during which transcription, translation, and degradation events have a constant probability to occur. Hence, at this level of description, the change observed in the mRNA and protein molecules between $t$ and $t+\Delta t$ has a certain probability distribution, which depends on the state of the system at time $t$ but not on previous times.

As we have seen in Chapter 3 (3.2.4) the above consideration allow us to formulate the Master equation describing the dynamics of $P(m,p,t) \equiv P(m,p,t|m_0,p_0,t_0)$, the PDF of observing $m$ mRNA molecules and $p$ proteins at time $t$ given that at $t = t_0$ one has $m_0$ and $p_0$ molecules, respectively. The Master equation given by

$$\begin{aligned} \frac{d\,P(m,p,t)}{dt} &= k_m\,P(m-1,p,t) - k_m\,P(m,p,t) && \text{(transcription)} \\ & \quad (m+1)\,d_m\,P(m+1,p,t) - m\,d_m\,P(m,p,t) && \text{(mRNA degradation)} \\ & \quad m\,k_p\,P(m,p-1,t) - m\,k_p\,P(m,p,t) && \text{(translation)} \\ & \quad (p+1)\,d_p\,P(m,p+1,t) - p\,d_p\,P(m,p,t) && \text{(protein degradation)} \end{aligned} \tag{6.1}$$

The picture conveyed by the above equation is the following. All processes that alter the state of the system obey exponential temporal statistics and, therefore, appear uncorrelated in time. For example, the time $\tau$ between successive transcription events is distributed according to

$$P(\tau) = k_m e^{-k_m\,\tau}. \tag{6.2}$$

Similar, exponential distributions describe the successive translation events of individual mRNA molecules as well as the lifetime of any mRNA or protein. The dynamics of the system is simply a combination of all these mutually independent processes.

### 6.2.2 Remarks on the Standard Model

As formulated above, SM attributes fluctuations in gene expression to the apparent randomness with which the processes considered (*i.e.*, transcription, translation, and degradation) occur over time. In this respect, SM only captures the *intrinsic* fluctuations of the system, and disregards external sources that effect the system in an apparently random fashion. Particular examples of *extrinsic* sources are bio-molecules that are actively involved in the processes of transcription (*e.g.*, RNAP), translation (*e.g.*, ribosomes), or degradation (*e.g.*, proteases). Such bio-molecules demonstrate fluctuations in their numbers that affect the expression of genes. Such effects can be introduced in the SM by allowing the rates of transcription, translation, and degradation to vary in some stochastic manner.

Here, we should also stress the fact that SM captures the intrinsic fluctuations of gene expression in a phenomenological manner, since it disregards all the microscopic dynamics of the processes involved. Processes are effectively treated as elementary chemical reactions obeying either zero or first order kinetics. As we will see in greater detail below,

the phenomenology invoked by SM relies on the assumption that all processes involve a rate limiting step that dominates their microscopic dynamics. In this respect they can be approximated by single-step processes.

For example, in the case of transcription this rate limiting step is assumed to be due to the slow time-scale at which the RNAP recognises the promoter sequences and initiates transcription. In general, however, the frequency of transcription initiation has a wide dynamical range *in-vivo* [85], and *in-vitro* studies have shown that initiation times can be as fast as a few seconds [89, 127, 160]. Clearly then, rapid initiation times can be significantly shorter than the time needed for elongation, especially for long DNA templates or bacterial genes transcribed in operons. In these cases, a Poisson representation of the process might be an inadequate approximation. Indeed, recent experimental studies focusing on the *in-vivo* transcription have demonstrated the non-Poisson character of the process [27, 55, 114], highlighting the need for more detailed microscopic models able to capture the intrinsic fluctuations of the process.

With the above in mind, in the following, we aim to qualitatively and quantitatively characterise the effect that the microscopic dynamics of DNA transcription have on the statistics of mRNA production. In particular, we use the model of elongation dynamics presented in Chapter 4 (4.4) as a starting point to demonstrate the effect of pauses due to backtracking on the statistics of the mRNA population. We then formulate a more general model of transcription incorporating elongation pauses and study the problem in greater detail.

## 6.3 Incorporating Elongation Dynamics

The elongation phase of transcription demonstrates non-trivial dynamics [82], such as RNAP pausing, that can significantly alter the statistics of the process. Here, we present an integrated model of DNA transcription and demonstrate how transcriptional pausing can qualitatively alter the statistic of mRNA production. The model is based on the model of elongation dynamics presented in Chapter 4 [see Eq. (4.31)]

As described in Chapter 4 elongation dynamics can be captured in terms of two discrete variables $(n, m)$. Variable $n$ denotes the position of the last transcribed nucleotide (or length of the RNA), whereas $m$ the position of the active site relative to $n$. From the active state $(n, m = 0)$ the TEC can proceed with polymerisation $(n + 1, m = 0)$ or depolymerisation $(n - 1, m = 0)$ of the nascent RNA at rates $p_+$ and $p_-$, respectively. Moreover, it backtracks $(n, m = -1)$ at a rate $p_b$. During backtracking the TEC hops between contiguous translocation state $(n, m = l) \rightarrow (n, m = l \pm +1)$ at rate $c$.

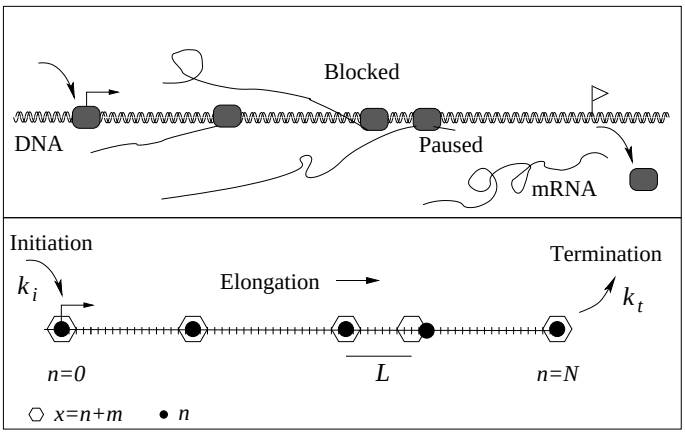



Figure 6.3: Schematic illustration of an integrated model involving initiation, elongation and mRNA degradation. Initiation occurs at a constant rate $k_i$ and multiple TECs are allowed to transcribe the same DNA template. During elongation the state of individual TECs is characterised by two variables $n$ and $m$. Variable $n$ denotes the position of the last transcribed nucleotide, whereas $m$ the position of the active site relative to $n$. The actual position of the TEC along the DNA template is given by $x = n + m$. Initiation involves the the formation of a TEC in state $(n = 0, m = 0)$ and termination of transcription occurs when state $(n = N, m = 0)$ has been reached. For RNA degradation a constant rate $k_d$ has been assumed.

Backtracking is restricted up to some boundary $(m, m = M)$ and polymerisation can proceed when the active state $(n, m = 0)$ is reattained. The elongation phase starts at state $(n = 0, m = 0)$ and terminates at state $(n = N, m = 0)$.

To provide a more complete model of transcription, we regard that that the initiation step, involving the loading of the RNAP on the DNA template and the formation of a TEC occupying state $(n = 0, m = 0)$ occurs at a constant rate $k_i$. Furthermore, we assume that termination takes place instantaneously when the transcript reaches its designated size $N$. To assess the dynamics of the RNA population we also include degradation which we model as a first order process with rate constant $k_d$. The combination of mRNA production and degradation gives a first handle on RNA levels and fluctuations in the cell.

In fact, RNA production is complicated by the fact that multiple initiation events can occur within the time it takes to produce a single RNA. This would lead to several TECs moving in tandem on the same DNA template [57], each synthesising a different RNA. To capture the physical restriction that two TECs cannot come in close proximity due to non-specific interactions between them or to the additional work required to deform the DNA helix [28, 88], we set a minimum (exclusion) distance of $L$ nucleotides ($L \ll N$) between the active sites of any two contiguous TECs. In terms of variables $n$ and $m$

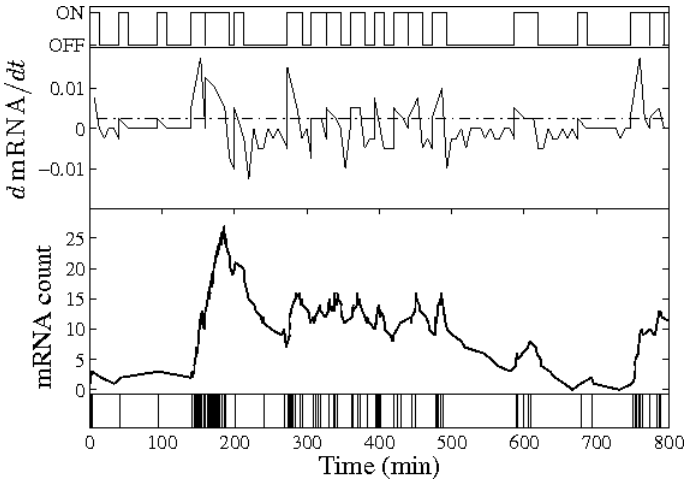



Figure 6.4: Results obtained from stochastic simulations of the integrated model of DNA transcription, illustrating the burst-like RNA production induced by backtracking pauses. The bottom panel shows the mRNA production events in time and the trace above illustrates the resulting mRNA count fluctuations. In the third panel $d$mRNA$/dt$ is plotted ($dt = 6min$), along with an arbitrary threshold (dotted line, set to $1/dt$ mRNA/sec). The threshold enables us to visualise the transcriptional process as a telegraph process with 'off' and 'on' states corresponding to low and high rates of mRNA production (top panel). The parameters used in the simulation are summarised in Section 6.5

the active site of a TEC is located at position $x = n + m$ along the DNA template. Therefore, a TEC, positioned at $x_1$, can translocate forward (backward) if the leading (trailing) TEC, positioned at $x_2$, is at distance of more than $L$ nucleotides, *i.e.*, $|x_1 - x_2| > L$. A similar argument also applies for transcription initiation, that is, no RNAP can initiate transcription if a TEC is present at position $x \leq L$. A schematic illustration of the model is given in Fig. 6.3.

Stochastic simulations (see section 6.5) of the model described above indicate that transcriptional pausing due to backtracking can give rise to burst-like production of RNA transcripts (see Fig. 6.4). Intuitively, sufficiently long pauses induced via backtracking can shut down mRNA production by blocking trailing TECs. In the intervals between pauses, multiple blocked TECs that have accumulated at the congestion site are likely to be transcribed in a burst of rapid mRNA production. In the following Section we study this phenomenon in greater detail using a coarse grained model of DNA transcription.

## 6.4 Coarse-Grained Model of DNA Transcription

In the previous Section we devised an integrated model of the transcription process and demonstrated that backtracking can result (under certain conditions) into bursts of mRNA production. However, long lived transcriptional pauses can be induced, besides backtracking, through a wide variety of mechanism such as sequence encoded signals [7], nucleosome packaging [24, 82] and DNA lesions [95].

Here we formulate a more general model of DNA transcription with the aim of quantitatively studying the effect of transcriptional pausing on the statistics of RNA production. The model is inspired by asymmetric exclusion processes (ASEP) that have been widely used in non-equilibrium statistical mechanics to model transport and traffic [36, 43].

### 6.4.1 Model Formulation

At a coarse grained level, DNA transcription can be described by a one dimensional totally asymmetric exclusion process [36, 43]. Within this picture, TECs are thought as particles moving on a chain, which represents the DNA template. Each site of the chain maps to a DNA region rather than a single nucleotide. As described in the previous section the length of this region is set by the minimum distance that two complexes can approach each other due to steric interactions between them or the additional work required to deform the DNA helix. Since at any point during transcription the footprint of a TEC is approximately $30$ nucleotides long [58], a reasonable estimate of the exclusion distance would of the order of $50 - 100$ nucleotides.

Transcription initiation occurs with rate $k_i$ and involves loading of a particle at position $n = 1$. While moving on the chain, particles can exist in two states representing active and paused TECs. Active particles hop forward with rate $k_f$ provided that the next site is not occupied. Forward movement is in kinetic competition with pausing which occurs at rate $k_p$. Once paused a particle can hop forward with a reduced rate $\bar{k}_f$ ($\bar{k}_f < k_f$) and its state is reset to active. Finally, a particle terminates transcription from site $n = N$ with rate $k_t$. The above transitions are schematically illustrated in Fig 6.5.

The four relevant time-scales associated with the model are

- $\tau_i = 1/k_i$: time-scale of initiation
- $\tau_f = 1/k_f$: time-scale of active elongation
- $\tau_p = 1/\bar{k}_f$: time-scale of a single pause
- $\tau_t = 1/k_t$: time-scale of termination

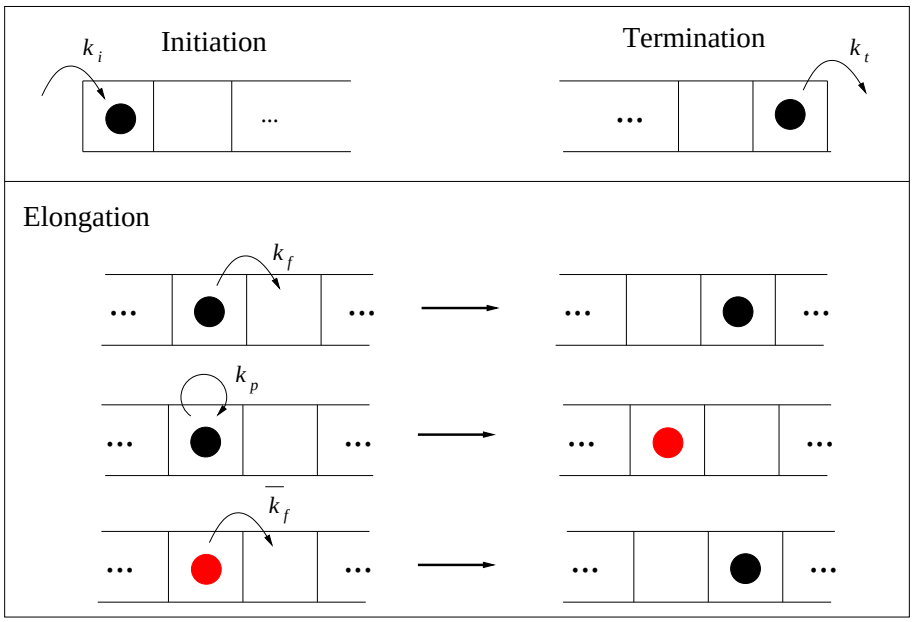



Figure 6.5: Schematic illustration of the state transitions involved in the coarse-grained ASEP type model of DNA transcription. Initiation, *i.e.*, loading of a particle at position $n = 0$, occurs at rate $k_i$. During elongation particles move forward on the chain ($n \rightarrow n+1$) at rate $k_f$. At any position particles can enter a paused state (red) at rate $k_p$. Forward movement of a paused particle occurs at rate $\bar{k}_f$. Termination occurs from position $n = N$ at rate $k_f$.

The overall dynamics of the process depend on the relationship between these time-scales. In particular, we define two dimensionless quantities $E$ and $S$ as:

$$\mathcal{S} \equiv \frac{k_f}{\bar{k}_f}, \tag{6.3}$$

$$\mathcal{E} \equiv \frac{k_p}{k_f}. \tag{6.4}$$

$\mathcal{S}$ ($\mathcal{S} \geq 1$) quantifies the time overhead introduced by transcriptional pausing, that is $S \approx 1$ indicates short pauses, while $S \gg 1$ long lived ones. On the other hand, $\mathcal{E}$ relates to the probability of entering the paused state at a specific site via

$$\text{Probability to pause} = \frac{\mathcal{E}}{1+\mathcal{E}}. \tag{6.5}$$

As $\mathcal{E} \rightarrow 0$, pauses become more and more infrequent while $\mathcal{E} \rightarrow \infty$ essentially guarantees pausing at each site.

### 6.4.2 Inter-arrival Statistics

Using stochastic simulations (see section 6.5) of the model presented above we examine the steady state statistics of the *inter-arrival times* ($T$), defined as the intervals between successive termination (RNA production) events. Our choice of studying the inter-arrival times instead of RNA populations levels enables us to disregard the process of mRNA degradation and focus solely on the microscopic dynamics of transcription. Furthermore, advancements in fluorescent techniques, allowing for single molecule resolution, make the direct measurement of inter-arrival times possible [70, 115, 154].

In particular we focus on the squared coefficient of variation $CV_T^2$ defined as

$$CV_T^2 = \frac{\sigma_T^2}{\langle T \rangle^2}. \tag{6.6}$$

$CV_T^2$ is a normalised measure for the dispersion of a probability distribution and provides a first handle on the temporal fluctuations of the process. Furthermore, it provides a useful measure for qualitative comparison with the Poisson process, which has been used in standard models of gene expression to model the transcription step. Events occurring according to a Poisson process are randomly and independently distributed in time. Therefore, the inter-arrival times follow an exponential distribution that yields $CV_T^2 = 1$. Consequently, super-Poisson (high variance) processes are indicated by $CV_T^2 > 1$, while sub-Poisson (low variance) processes by $CV_T^2 < 1$.

### 6.4.3 Statistics of RNA Production in the Absence of Pauses

We start our analysis by considering the simplest scenario, in which TECs are not allowed to enter the paused state, *i.e.*, $\mathcal{E} = 0$. As illustrated in Fig. 6.6, the relation between the three relevant time-scales $\tau_i$, $\tau_f$, and $\tau_f$ alter the statistics of the inter-arrival times from Poisson to sub-Poisson.

In particular, for $\tau_i \gg \tau_f, \tau_t$ [regime (I) in Fig. 6.6], initiation becomes the rate limiting step and fully determines the dynamics of the process. In this regime the mean inter-arrival time scales like $1/k_i$ and the squared coefficient of variation approaches unity (see Fig. 6.7). Effectively, the model becomes equivalent to a Poisson process with rate parameter $k_i$ and hence the inter-arrival times obey an exponential distribution (see Fig. 6.8). Similar results are also obtained for $\tau_f \gg \tau_i, \tau_f$ [regime (III) in Fig. 6.7].

As $\tau_f$ is increased relative to the two other time-scales, the elongation phase starts adding more and more to the total transcription time. This has as a consequence the decrease of the temporal fluctuations (see Fig. 6.7), since the dynamics of the process

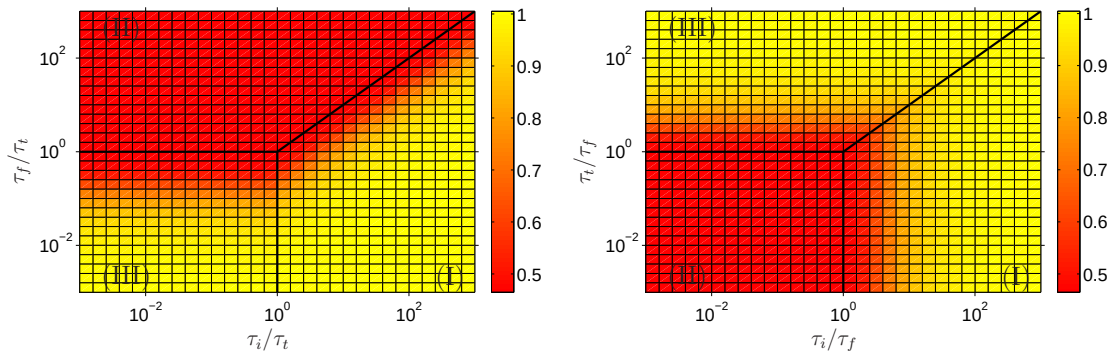



Figure 6.6: Heat maps of the squared coefficient of variation of the inter-arrival times ($CV_T^2$) in the absence of transcriptional pauses ($\mathcal{E} = 0$). Depending on the relation between the three relevant time-scales $\tau_i$, $\tau_f$, $\tau_t$, the behaviour of the model can be classified into three regimes. Regime (I) and (III) correspond to Poisson statistics ($CV_T^2 = 1$), whereas, regime (II) corresponds to sub-Poisson statistics ($CV_T^2 < 1$). Results were obtained using stochastic simulation of the model for $N = 20$.

cease to be governed by a single rate limiting step. When $\tau_f \gtrsim \tau_i, \tau_t$ the dynamics of transcription are dominated by the elongation phase, which makes the process appear sub-Poisson [regime (II) in Fig. 6.6]. In this regime the DNA template is fully occupied by TECs moving in tandem. A TEC will occasionally be blocked behind another one, but on average their motion will be regular and mRNA production will be occurring at rather fixed intervals. This is demonstrated in the distribution of the inter-arrival times, which becomes narrowly peaked around the mean and can be well fitted by a gamma distribution (see Fig.6.8).

In summary, when transcriptional pauses are negligible the dynamics of the process depend on whether a single rate limiting step is present or not. Given sufficiently low rates of initiation or termination the process demonstrates Poisson characteristics, while when the elongation phase becomes significant temporal fluctuations tend to get averaged out.

### 6.4.4 The Effect of Pause Lifetimes

We now turn to the question of how transcriptional pauses affect the statistics of the inter-arrival times. Inclusion of transcriptional pauses adds an additional time-scale $\tau_p$ and the relation between this time-scale and those of initiation ($\tau_i$) and active elongation ($\tau_f$) dictates the behaviour of the process. As illustrated in Fig. 6.9, we can distinguish three main regimes in the parameter space giving rise to qualitatively different behaviour.

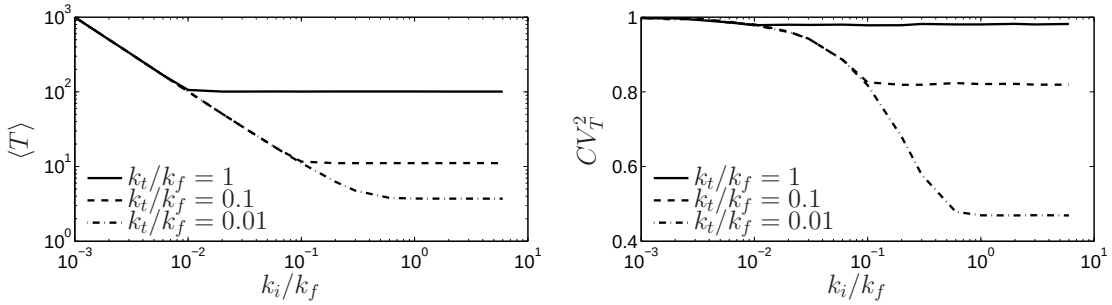



Figure 6.7: The mean inter-arrival time ($\langle T \rangle$) and the squared coefficient of variation ($CV_T^2$) as a function of the initiation rate ($k_i/k_f$) for $\mathcal{E} = 0$. When initiation is the rate limiting step $\langle T \rangle$ demonstrates a linear variation and $CV_T^2 = 1$, indicating the Poisson character of the process. For higher values of $k_i$ the time spent on active elongation becomes significant and wipes out temporal fluctuation. Results were obtained using stochastic simulation of the model for $N = 20$.

In the limit of $\tau_i \gg \tau_f, \tau_p$ [regime (I) in Fig. 6.6] initiation dynamics dominate the process. In this regime the density of TECs on the DNA template is low and therefore transcriptional pauses and interactions between TECs are expected to have only marginal effects. Therefore, as discussed above, the model becomes equivalent to a Poisson process and inter-arrival times obey an exponential distribution. For $\tau_f \gg \tau_i, \tau_p$ [regime (III) in Fig. 6.9] fast transcription initiation is blocked by the slow movement of the TECs on the DNA template, while the relatively short-lived transcriptional pauses, as in the case above, play no significant role. In particular, in this regime the density of the TECs along the DNA is maximal and their regular motion gives rise to a sub-Poisson statistics ($CV_T^2 < 1$).

However, the behaviour of the model changes significantly when pauses dominate transcription. In particular, for $\tau_p \gg \tau_i, \tau_f$ [regime (II) in Fig. 6.6] we observe $CV_T^2 > 1$ indicating the super-Poisson behaviour of the process. In particular, the distribution of inter-arrival times becomes heavy-tailed and two bumps appear in its shape, indicative of a burst-like production of RNA transcripts (see Fig. 6.10). The physical picture behind such behaviour is a simple one. Long lived transcriptional pauses can create congestion points by blocking the movement of trailing TECs, while the leading TECs continue to transcribe normally. In this way the uniform [regime (I)] or Poisson [regime (III)] distribution of TECs on the DNA template is disrupted, resulting in a burst-like production of mRNA transcripts.

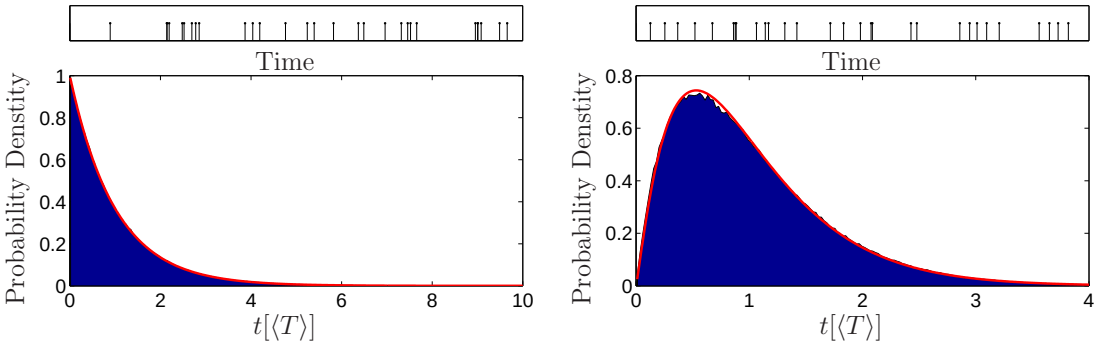



Figure 6.8: The distribution of the inter-arrival times (scaled by the mean) in the absence of pauses ($\mathcal{E} = 0$) at the two limiting regimes: $\tau_i \gg \tau_i, \tau_t$ (left panel) and $\tau_f \gg \tau_i, \tau_t$ (right panel). For low rates of initiation the inter-arrival times are in agreement with an exponential distribution with rate parameter $k_i$ (red line). For higher values the distribution become narrowly peaked around the mean value. Here the red line denotes a Gamma distribution with the same mean and variance. Results we obtained using stochastic simulation of the model for $N = 20$, $k_t/k_i = 1$ $k_i/k_f = 10^{-2}$ (left panel) and $k_i/k_f = 1$ (right panel).

## 6.5 Numerical Methods

For the model presented in Section 6.3 results were obtained using stochastic simulations (Gillespie algorithm) [52] with the following set of parameters: $N = 4$ kbp, $L = 100$ bp, $M = 10$ bp, $p_+ = 50$ s$^{-1}$, $p_- = 0.5$ s$^{-1}$, $c = 0.1$ s$^{-1}$, $k_i = 0.02$ s$^{-1}$ and $k_d = 310^{-4}$ s$^{-1}$ and $p_b = 0.05$ s$^{-1}$ (yielding approximately 1 pause/kb). The code was implemented in JAVA and a single run was performed, shown in Fig. 6.4. The system was monitored using

- a list of state variables $(n_i, m_i)$, denoting the state of the $i$th TEC along the DNA template,
- a counter $C_{mRNA}$ keeping track of RNA molecules,
- a timer $t$.

The system was initialised with an empty list of state variables (no TECs on the DNA template), $C_{mRNA} = 0$, and $t = 0$. Each time an initiation event occurred a new set of variables $(n = 0, m = 0)$ was added at the beginning of the list. In the case of a termination or degradation event $C_{mRNA}$ was updated accordingly. At each step of the algorithm, all permissible transitions for each TEC present on the DNA template were calculated based on the list of state variable and were added to an "event" list. This list

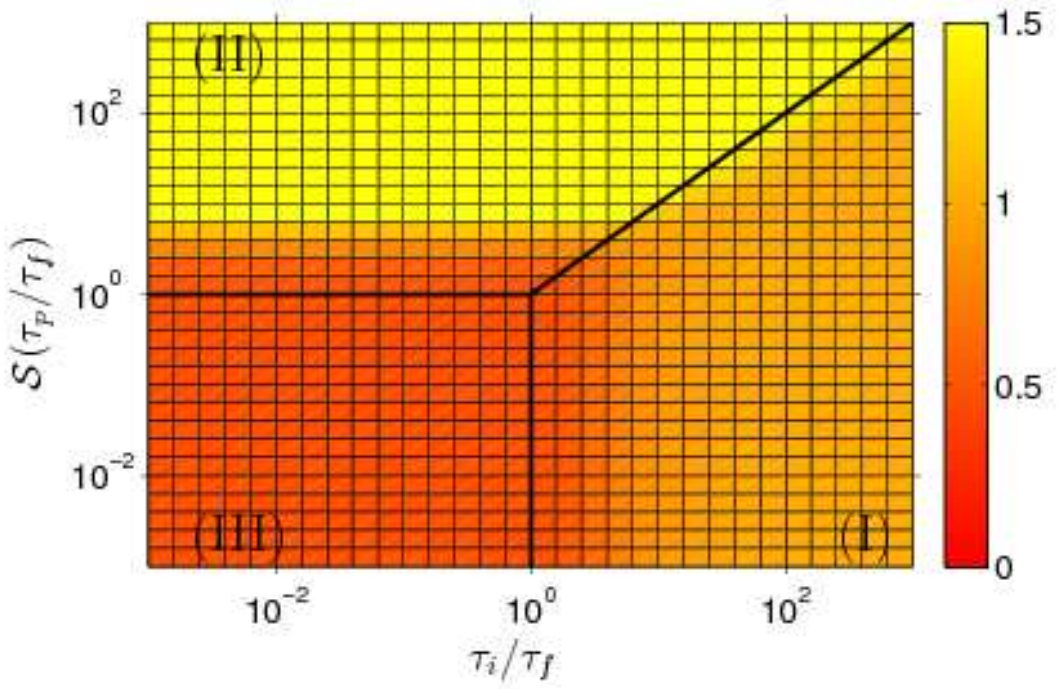



Figure 6.9: Heat maps of the squared coefficient of variation of the inter-arrival times ($CV_T^2$) in the presence of transcriptional pauses ($\mathcal{E} > 0$). Depending on the relation between the three relevant time-scales $\tau_i$, $\tau_f$, $\tau_p$, the behaviour of the model can be classified into three regimes. Regime (I) ($\tau_i > \tau_f, \tau_p$) corresponds to sub-Poisson statistics ($CV_T^2 < 1$), regime (II) ($\tau_p > \tau_f, \tau_p$) to super-Poisson statistics ($CV_T^2 < 1$), and regime (III) ($\tau_i > \tau_f, \tau_p$) to Poisson statistics ($CV_T^2 = 1$). Results were obtained using stochastic simulation of the model for $N = 20$, $\mathcal{E} = 0.05$.

was also appended with the events of initiation (when $n_1+m_1 < L$), and RNA degradation (when $C_{mRNA} > 0$). From the list of events, one was chosen with probability proportional to the corresponding rate [see Chapter 3 (3.3.4)] and the system state was updated.

All results presented in Section 6.4 were obtained using stochastic simulation of the coarse grained model of DNA transcription. As above the state of the model was monitored using

- a list of state variables $(n_i, l_i)$, denoting the position ($n_i$) of the $i$th particle along the chain and its current state ($l_i = 0, 1$, either paused or active)
- a list of termination times $\mathcal{T}_i$
- a timer $t$.

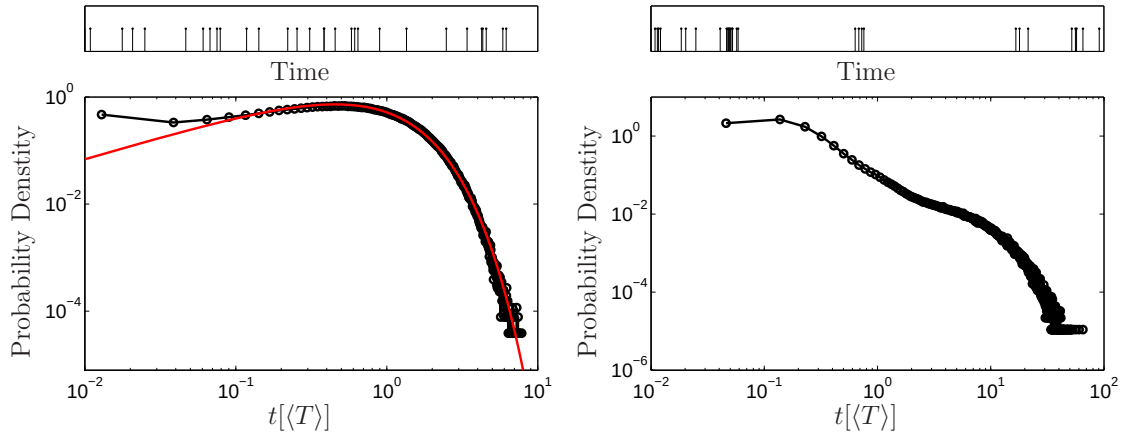



Figure 6.10: The distribution of the inter-arrival times (scaled by the mean) in the presence of pauses ($\mathcal{E} > 0$) at two limiting regimes: $\tau_i \gg \tau_i, \tau_t$ (left panel) and $\tau_f \gg \tau_i, \tau_t$ (right panel). For low rates of initiation the inter-arrival times are in good agreement with an exponential distribution with rate parameter $k_i$. For higher values the distribution becomes fat-tailed and RNA production appears to occur in bursts. Results were obtained using stochastic simulation of the model for $N = 20$, $\mathcal{E} = 0.05$

The system was initialised with an empty list of state variables and termination times and $t = 0$. Each time an initiation event occurred a new set of variables $(n = 0, l = 1)$ was added at the beginning of the list. In the case of a termination event the current time was appended in the list of termination times, *i.e.*, $\mathcal{T}_{last} = t$. At each step of the algorithm, all permissible transitions for each particle on the DNA chain were calculated based on the list of state variable and were added to an "event" list. From the list of events, one was chosen with probability proportional to the corresponding rate [see Chapter 3 (3.3.4)] and the system was updated accordingly. The code was implemented in JAVA and a single simulation run was performed for each for each set of parameters allowing the list of times to reach a size of $10^5$ elements. For the analysis, however, the first $10^3$ elements were neglected to ensure that the density of the particles on the DNA chain had reached a steady state. Inter-arrival times were calculated by subtracting consecutive elements of the list, i.e., $T_i = \mathcal{T}_{i+1} - \mathcal{T}_i$.

For the distribution presented in Fig. 6.8 and 6.10 the data obtained ($T_i$) were rescaled by their mean value,

$$\bar{T}_i = \frac{T_i}{\sum_j T_j} \tag{6.7}$$

and binned. Bin frequencies were subsequently transformed into probabilities by division with the size of the sample and finally into probability densities by division with the bin

| Regime | | Behaviour |
|---|---|---|
| $\tau_i \gg \tau_p, \tau_f$ | $\tau_p \gg \tau_f$ | Poisson |
| | $\tau_f \gg \tau_p$ | Poisson |
| $\tau_f \gg \tau_p, \tau_i$ | $\tau_i \gg \tau_p$ | sub-Poisson |
| | $\tau_3 \gg \tau_1$ | sub-Poisson |
| $\tau_p \gg \tau_i, \tau_f$ | $\tau_i \gg \tau_f$ | super-Poisson |
| | $\tau_f \gg \tau_i$ | super-Poisson |
| $\tau_i \sim \tau_f \gg \tau_p$ | | sub-Poisson |
| $\tau_i \sim \tau_p \gg \tau_f$ | | super-Poisson |
| $\tau_f \sim \tau_p \gg \tau_i$ | | super-Poisson |

Table 6.2: Table summarising the behaviour of RNA production in the different limiting regimes.

size.

## 6.6 Summary and Discussion

In this Chapter we have presented a integrated model of DNA transcription linking the microscopic dynamics of the process to fluctuations in mRNA production and gene expression. The model incorporated the initiation, elongation, and termination phases of DNA transcription and was formulated in terms of a totally asymmetric exclusion process to take into account that multiple RNAPs with repulsive interactions can simultaneously transcribe the DNA template. Our results indicate that the interplay between the different time-scales of the model in combination with the exclusive interactions between transcribing TECs can significantly alter the temporal statistics of mRNA production. A qualitative description of the different classes of behaviour obtained is presented in Table 6.2.

Following the work presented in previous chapters we particularly focused on characterising the effect of transcriptional pauses on the statistics of mRNA production. Our results suggest rare and long pauses can result in a burst-like production of mRNA transcripts and super-Poisson mRNA statistics. The effect of pauses can be linked heuristically to a switching mechanism between high and low rates of mRNA production. In particular, sufficiently long pauses shut down mRNA production by jamming TEC trafficking on the DNA template. Once the leading TEC resumes elongation multiple blocked TECs that have accumulated at the congestion site are likely to terminate transcription resulting in burst of rapid mRNA production. Similar findings illustrating the effect of transcriptional pauses on the statistics of RNA production were independently reported in Ref. [37].

Interestingly, recent experiments have provided evidence of the existence of bursts of transcription both in bacterial [55] and eykaryotic cells [27, 114]. Our model attributes this phenomenon to particularly long pauses that occur during transcription elongation. Such pauses can be attributed to a wide range of factors such as RNAP backtracking, sequence encoded signals [7], molecules that interact with the transcribing RNAP, DNA lesions ,or nucleosome packaging [24, 82]. We note, however, that burst of mRNA production can also be attributed to other phenomena. For example, changes in the state of the promoter due to chromatin remodelling [27, 114] or the the diffusive motion of regulatory molecules [142] can also provide a switching mechanism between rapid and slow mRNA production

Advancements in experimental techniques, which allow one to track levels of chemical species within cells, have renewed the interest in the stochastic nature of gene expression and its implications regarding cell behaviour and fate. So far, however, modelling attempts have focused on a coarse grained level of description ignoring the microscopic details of the processes involved in gene expression. The results presented in this Chapter can also be relevant for translation and highlight the need for a finer level of description to understand gene expression and regulation and fluctuations therein.

# Chapter 7

# Population Level: The Social Behaviour of Bacteria

---

The stochastic nature of subcellular processes plays a crucial role in determining cellular behaviour and cell fate. However, cells rarely exist in isolation, and their behaviour is also shaped to a large extent by inter-cellular communication. In this Chapter, we aim to study in a simplified context how the dynamics and behaviour of a cell population, shaped by interactions between individual cells, is affected by intra-cellular fluctuations. Inspired by real life bacterial communication, we propose and study an artificial gene regulation network. The network couples bacterial cells via two distinct communication channels and gives rise to two mutually exclusive bacterial behaviours. Beyond some critical threshold of coupling, coordination at the population level is achieved, with the majority of the cells adopting one of the two behaviours. Our results indicate that subcellular fluctuations raise the critical coupling strength at which transition to majority consensus is observed. We provide a physical explanation of the phenomenon using a coarse-grained, Ising-type model of the bacterial population. The *in-silico* paradigm of bacterial social behaviour presented in this Chapter illustrates the bidirectional relationship between cellular and population-level dynamics exemplifying possible effects that intra-cellular fluctuations can have at the population level.

## 7.1 Introduction

Cells are constantly presented with "choices" regarding their fate and behaviour. The mechanisms underlying their apparent decision making are intricate networks of regulatory interactions between genes and proteins. These networks function as genetic programs giving rise to distinct cellular behaviours in response to changes in environmental conditions or changes of the cell's internal state. However, these modules of cellular functionality are far from reliable. Instead it has long been appreciated that the inherent stochasticity of subcellular processes renders randomness a key player in dictating cellular phenotype, behaviour and fate [86, 102], one that cells must adapt to cope with or occasionally exploit to their advantage.

A simple, yet illustrative example comes from the lifestyle of $\lambda$ phage, a virus infecting bacterial (*Escherichia Coli*) cells. Upon infection, the genome of the phage ($\sim 50$ genes) is integrated into the bacterial DNA, and subsequently host machinery facilitates the expression (transcription and translation) of its genes. The $\lambda$ phage genome contains a rather simple genetic programme enabling the phage to choose between two distinct lifestyles, the *lysogenic* and the *lytic* one [112]. Under conditions that allow bacterial proliferation, the phage adopts the lysogenic lifestyle, where the protein product of a master regulator gene is responsible for repressing the the rest of the phage genes. Hence, the phage remains dormant and its genetic material is passively replicated along with the rest of the bacterial DNA. When, however, the bacterial population is stressed through exposure to UV light, the phage switches to its lytic lifestyle. Expression of phage genes is rapidly turned on and as many as $100$ phage particles are assembled causing the bacterial cell to lyse (burst) [112]. Importantly, switching from the lysogenic to the lytic lifestyle can also be triggered in the absence of environmental stimuli, solely due the stochastic nature of the processes involved in gene expression. Not surprisingly $\lambda$ phage has evolved elegant mechanisms for minimising these randomly induced lytic events [142], The $\lambda$ phage paradigm illustrates the crucial role of fluctuations in dictating the behaviour and fate of individual cells.

Of course, one should also appreciate the fact that cells rarely exist in isolation. In multicellurar organisms, for example, cells are constantly signalling to each other, synchronising their activities in this manner and coordinating their fates during development [21]. Similar cell-to-cell communication is observed in the bacterial kingdom. Bacterial communication, termed as *quorum sensing*, is mediated by small molecules called *autoinducers* that bacterial cells produce, release to their environment, and detect [150]. When the autoinducer molecules reach some critical concentration within a bacterium

they trigger a quorum response by activating certain transcription factor proteins that regulate the expression of quorum-specific genes [150]. In this manner, bacterial cells are constantly communicating with one another orchestrating their behaviour in response to environmental stimuli and changes in their density.

The dependence of population-wide dynamics on the inter-cellular communication raises the question of how noise present at the intra-cellular level affects the behaviour at the cell population level. Such a question is particularly interesting to the physics community that has extensively studied the collective behaviour of noise-driven, non-linear systems in many different contexts [48]. Specific examples of particular interest are ensembles of noise-driven bistable switches [111]. Effectively, each switch can be considered as a system possessing a double-well energy landscape with the two wells corresponding to the two discrete states that the switch can attain. In the absence of any coupling, due to intrinsic fluctuations individual switches undergo random transitions from one state to the other. The presence of a uniform all-to-all coupling, however, gives rise to a critical coupling strength at which the population undergoes a phase transition from a "disordered" state – where noise dominates and the switches are partitioned between the two states – to an "ordered" one – where the majority of the switches occupy one the two states. In this Chapter, we study how intra-cellular fluctuations affect the behaviour at the population level using a gene-regulatory network that demonstrates qualitatively similar behaviour to the toy model described above.

The remainder of this Chapter is organised as follows. We start with a brief review of bacterial communication and its importance for bacterial life. Next, inspired by real-life bacterial behaviour, we propose and analyse an *in-silico* gene regulatory network. This network enables us to dissect bacterial communication and study it in a simplified context. More importantly, it serves as a fine system to study how intrinsic fluctuations at the cellular level affect the behaviour of bacterial populations. In a nutshell, the circuit enables cells to choose between two antagonistic social behaviours. Beyond some critical threshold of cell coupling, coordination at the population level is achieved, with the majority of the cells adopting one of the two behaviours. Our results illustrate that subcellular fluctuations hinder the ability of cells to achieve majority consensus, making the population appear more disordered. Finally, to gain a deeper insight into the transition between the two regimes of behaviour we present and analyse a coarse grained, Ising-type model of the dynamics at the population level.

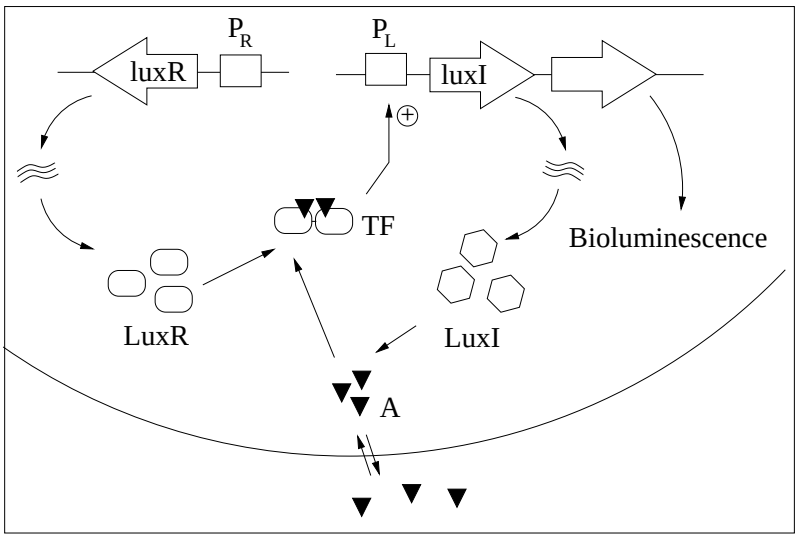



Figure 7.1: The Quorum sensing system in *V. fisheri* (adapted from Ref. [150]). Communication in *V. fisheri* is mediated by small molecules called autoinducers (AI). These molecules are produced by specific enzymes (LuxI synthetase) and in turn diffuse in and out of bacterial cells. When present in sufficiently high concentrations within cells, indicative of high cell density, autoinducers trigger a quorum response by activating specific proteins (LuxR receptor) that control the expression of genes.

## 7.2 Bacterial Communication

The gene regulatory network we propose and study here enables coupling between cells via two mutually inhibiting quorum sensing modules. Before presenting the actual network it is therefore essential to give a brief overview of quorum sensing and its importance for bacterial life as well as to present specific real-life examples of bacterial communication that have motivated the design of the network.

### 7.2.1 The *Vibrio fischeri* Paradigm

Quorum sensing was first discovered and described in the marine bacterium *Vibrio fischeri* [98]. This bacterium colonises the light organ of the Hawaiian squid, *Eupryman scolopes* [144], where necessary nutrients are provided for its proliferation. In exchange, *V. fischeri* uses quorum sensing to induce expression of bioluminesence genes once it has grown to sufficiently high cell densities. The light emitted by the bacterial colony is used by the squid to mask its shadow and avoid predation [98].

As illustrated in Fig. 7.1, the quorum sensing system in *V. fischeri* consists of two proteins, LuxI and LuxR. The former (I protein) is involved in the synthesis of autoinducer

molecules (AI molecules), the chemical signal used for bacterial communication. In the case of *V. fischeri* this signal is an Acyl-homoserine lactone (AHL). Following production, autoinducer molecules freely diffuse in and out of the cell and their concentration increases with increasing cell density. The second protein (R protein) is the autoinducer receptor. When present at sufficiently high concentrations, autoinducer molecules readily bind to LuxR and promote its dimerisation [26, 61]. In this form, LuxR can recognise and bind specific promoter sequences upregulating the expression of certain genes [126, 129]. Among these genes are ones responsible for bioluminesence as well as the gene encoding for the LuxI protein [126]. This gives rise to a positive feedback loop that locks cells into a quorum sensing mode [150].

### 7.2.2 An Overview of the Complexity in Bacterial Communication

Following the seminal discovery of quorum sensing in *V. fischeri* it was appreciated that a large number of bacteria possess communication systems obeying similar principles [87, 150]. In particular, different autoinducer molecules are produced by many bacterial species. These molecules either diffuse or are actively transported to the extracellular environment and their concentration is constantly gauged. Beyond some critical concentration (corresponding the high cell density) autoinducers trigger a quorum response by regulating the expression of specific genes. The similarities of quorum systems across different bacterial species, points to a common ancestral origin and is perhaps the strongest evidence for the importance of quorum sensing for bacteria and their survival. Nonetheless, closely related (*homologous*) quorum sensing systems of different bacteria demonstrate extreme specificity: differences in the structure of the autoinducers as well as in the structure of the receptor proteins play an important role in conveying signalling specificity [150]. That is, autoinducers can only activate their cognate receptor proteins and therefore allow only for intraspecies communication.

The social life of bacteria becomes even more intriguing when one recognises that many bacterial species possess multiple quorum sensing systems. Such systems are most often interweaved with one another, arranged in parallel [23], in series [123] and in some cases in direct competition with one another [59] resulting in rather complex behaviour. One particular example comes from the well studied bacterium *Bacillus subtilis*. When presented with stress conditions, *B. subtilis* commits to one of two mutually exclusive lifestyles: sporulation or competence. In the first state the bacterium undergoes a physiological change that enables it to survive for extended periods of time under unfavourable environmental conditions. The second state enables the bacterial cells to uptake exoge-

nous DNA to be utilised as energy source or incorporated into the genome. Interestingly *B. subtilis* relies on inter-cellular communication through two competitive (inhibiting) quorum sensing systems to decide which of the two fates to choose [29, 59, 97].

Unlike the classic *V. fischeri* example, bacterial communication is blocked in many cases by signals coming from the host or even other bacteria growing in the same niche. Such inhibition, termed as *quorum quenching*,enables hosts to prevent colonisation by pathogenic bacteria or allows certain bacterial species to proliferate faster than others [38, 150]. A particularly interesting example of the second scenario comes from the social life of *Staphylococcus aureus*. This pathogenic bacterium comes in different strains that are classified according to the autoinducer molecules they produce [39, 103]. Surprisingly, autoinducers of one strain directly inhibit the quorum sensing machinery of other strains [84]. For example such behaviour imposes direct competition between populations of different strains of *S. aureus* when they co-infect a host.

## 7.3 An *in-silico* Paradigm for Bacterial Communication

In this section we present an artificial gene regulatory network consisting of two mutually inhibiting quorum sensing modules similar to the one found in the bacterium *V. fischeri*. Our primary goal is to study the dynamics that the regulatory network conveys at the population level, and in particular the effect of subsecullar fluctuations. The construction of the network was inspired by the complex social lives of *B. subtilis* and *S. aureus* presented in the preceding section. In this respect, the proposed regulatory network can also serve as an paradigm for bacterial communication, enabling one to dissect complex bacterial social behaviour and study it in a simplified context, in the spirit of *synthetic biology*.

Synthetic biology is a young discipline that is already changing the life sciences as we know them. The main aim of synthetic biology is the bottom up construction of novel biological systems, ranging from small genetic circuits to fully functional cells and even ecosystems [113]. From an engineering perspective, such systems have potential applications in a wide range of areas, with medicine [6], drug synthesis [2] and sustainable energy production [120] being a few indicative examples. On the other hand the construction of simple synthetic systems with predefined functions enables one to dissect life processes and study them within a simplified context and under controlled conditions. In this manner, synthetic biology has a crucial role to play in understanding natural biological processes and the first principles underpinning life.

Early efforts in synthetic biology have been particularly successful in assembling small regulatory networks from basic elements, such as promoters and genes encoding

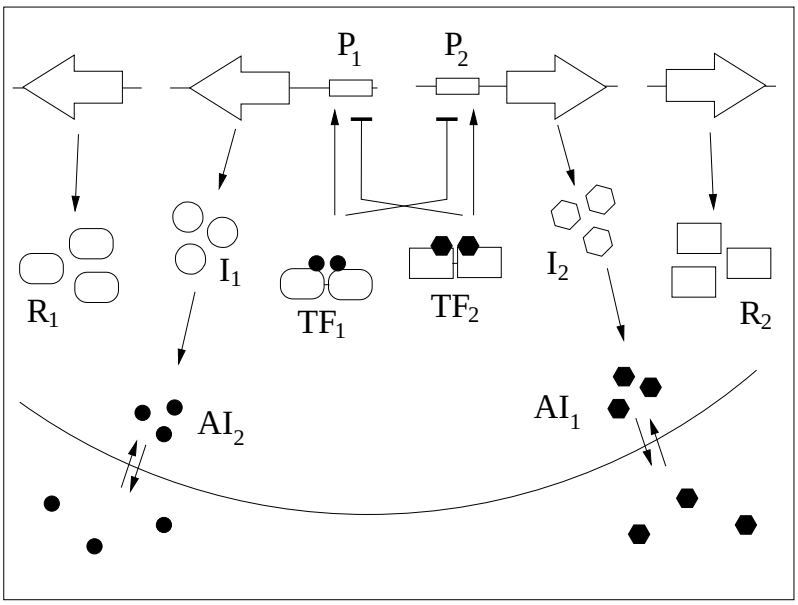



Figure 7.2: Schematic illustration of the proposed gene regulatory network. It consists of two mutually inhibiting quorum sensing modules giving rise to two mutually exclusive behaviours. Two species of autoinducers ($AI_1/AI_2$) are produced intra-cellularly by synthetase enzymes ($I_1/I_2$). In turn, autoinducer molecules diffuse in and out of the cells. When an autoinducer species is present in high concentrations it triggers a quorum response by binding to its cognate receptor protein ($R_1/R_2$) and activating it. Activated receptor proteins ($TF_1/TF_2$) upregulate the expression of the cognate synthetase and repress the expression of non-cognate one.

transcriptional factors [113]. Such circuits were used to generate different kinds of behaviour, including oscillations [8, 41, 54, 130, 135, 139] , bistability [8, 51, 77], pulse generation [13], spatial patterning [12] and logic functions [117, 153]. Also, synthetic paradigms have been extensively used to study the design principles and dynamical properties of small, naturally occurring, regulatory motifs that include linear and feed-forward regulatory cascades and autoregulation [9, 14, 40, 67]. More recently, attention has also been given to synthetic ecosystems and the design of synthetic gene networks that are capable of conveying nontrivial population wide behaviour. Examples include usage of synthetic quorum sensing modules to achieve regulation of cell density [157], predator-prey dynamics [11] and coordinated behaviour between cells [125]. The artificial gene regulatory network we propose here can, therefore, be particularly motivating with regard to these recent bio-engineering efforts.

### 7.3.1 The Synthetic Circuit

Figure 7.2 gives a schematic illustration of the proposed gene regulatory network. It consists of two mutually repressing quorum sensing modules. Similar to *V. fischeri*, each module consists of two genes encoding for the autoinducer synthetase enzyme (denoted by $I_1/I_2$) and the autoinducer receptor (denoted by $R_1/R_2$). Initially, both of the genes are expressed at a basal rate. Proteins $I_1$ and $I_2$ produce distinct autoinducer signalling molecules (denoted by $AI_1/AI_2$) that are free to diffuse in and out of the cell. This enables cells to communicate via two distinct channels. When autoinducer molecules are present at high concentrations (corresponding to a high cell density) they convey a quorum response by readily binding to their cognate receptor proteins and enabling dimerisation. In the dimeric form (denoted by TF), R proteins bind to promoter sites on the DNA (denoted by $P_1/P_2$) and regulate the expression of the synthetases. In particular, each promoter contains two binding sequences, one for each TF. These sequences enable each TF to upregulate the expression of its cognate I protein while downregulating the expression of the non-cognate one.

The positive feedback established for each quorum sensing module along with the mutual inhibition established between them allow cells to adopt one of two mutually exclusive behaviours (states): expressing one of the two autoinducer synthetase proteins and therefore communicating via one of the two channels. Such behaviour where the bacterium chooses between two distinct physiological states using two mutually inhibiting quorum sensing modules is reminiscent of *B. subtilis*. The picture is also similar to the competition observed between different *S. aureus* strains. This is readily seen if the gene regulatory network is broken into two parts and placed in distinct cell types, as Fig. 7.3 illustrates. In this case, each cell type is capable of producing only one autoinducer signal but responds to both. In particular, each cell type responds to cognate (non-cognate) autoinducer molecules by up(down)-regulating the production of the I protein. In this manner, when present in the same environment the two cell types are in direct competition with each other.

### 7.3.2 Modelling the Dynamics

At a coarse grained level the dynamics of gene regulatory networks can be described in terms of chemical reactions occurring at constant rates. Here we summarise the reactions that capture the key behaviour of the gene regulatory network and their corresponding rates. The dynamics of the network can be broken up into the following three components.

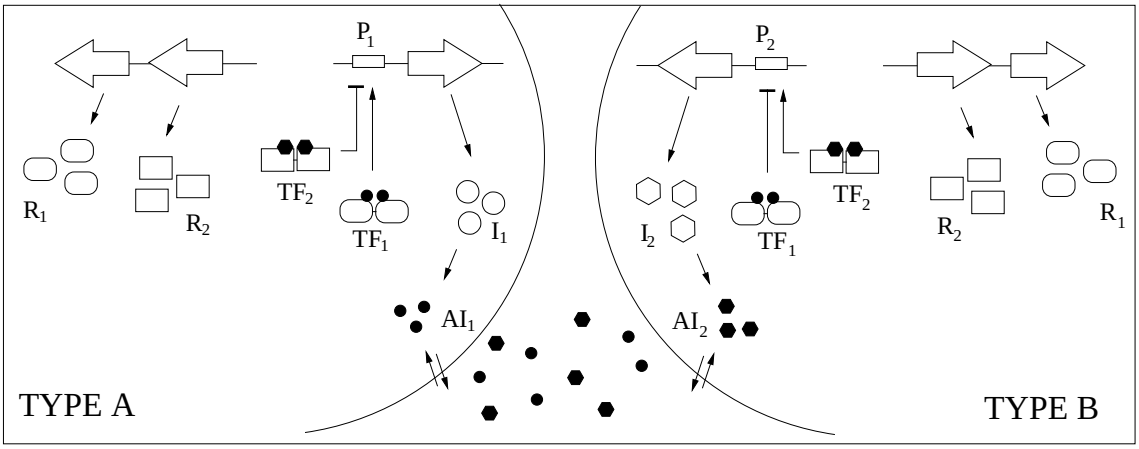



Figure 7.3: Schematic illustration of the proposed gene regulatory network giving rise inter-species competition. The two bacterial cell types (A and B) produces different autoinducer molecules ($AI_1/AI_2$). However, they are capable of detecting and responding to both signals. Cognate autoinducer molecules upregulate the production of the synthetase proteins ($I_1/I_2$) whereas non-cognate ones downregulate it. Regulation is achieved via receptor proteins ($R_1/R_2$), which – upon binding to their cognate autoinducers – are able to dimerise and form active transcription factors ($TF_1/TF_2$). In turn, transcription factors molecules bind to the promoters ($P_1/P_2$) driving the expression of the synthetase proteins.

### Autoinducer Dynamics

Autoinducer molecules, $AI_i$, are produced by their cognate synthetases, $I_i$, at rate $\alpha_A$. Following their production, $AI_i$ diffuse in and out of the cell with rates $\eta$ and $\eta_{\text{ext}}$, respectively. Following [49] we define the diffusion rates as $\eta = \sigma\mathcal{A}/V_c$ and $\eta_{\text{ext}} = \sigma\mathcal{A}/V_e$, where $\sigma$ represents the membrane permeability, $\mathcal{A}$ the surface area, and $V_c$, $V_e$ denote the intra-cellular and extra-cellular volumes, respectively. Finally, due to different conditions, autoinducer molecules degrade with rates $\delta_A$ and $\delta_{A,\text{ext}}$ depending on whether they reside inside or outside the cell.

$$\mathrm{I}_i \rightarrow \mathrm{I}_i + \mathrm{A}_i \quad : \quad \alpha_A \tag{7.1a}$$

$$\mathrm{AI}_i \rightleftharpoons \mathrm{Ai}_{i,\text{ext}} \quad : \quad \eta(\rightarrow), \eta_{\text{ext}}(\leftarrow) \tag{7.1b}$$

$$\mathrm{AI}_i \rightarrow \emptyset \quad : \quad \delta_A \tag{7.1c}$$

$$\mathrm{AI}_{i,\text{ext}} \rightarrow \emptyset \quad : \quad \delta_{A,ext} \tag{7.1d}$$

### Transcription Factor Formation

Autoinducer receptor proteins $R_i$ are constitutively expressed at all times and we therefore assume their numbers constant. As the autoinducer molecules start growing in numbers

| **Species** | **Description** |
|---|---|
| $AI_i$ | autoinducer |
| $I_i$ | autoinducer synthetase |
| $R_i$ | autoinducer receptor |
| $R_i$-$AI_i$ | receptor-autoinducer complex |
| $TF_i$ | transcription factor |
| $P_i$ | promoter driving the expression of the synthetase |

Table 7.1: Summary of species involved in the gene regulatory network.

they bind to $R_i$ forming $R_i$-$AI_i$ complexes. These complexes then dimerise to form the active transcription factors, $TF_i$.

$$R_i + AI_i \rightleftharpoons R_i\text{-}AI_i \quad : \quad k_1(\rightarrow), k_2(\leftarrow) \tag{7.2a}$$

$$2R_i\text{-}AI_i \rightleftharpoons TF_i \quad : \quad k_3(\rightarrow), k_4(\leftarrow) \tag{7.2b}$$

**Autoinducer Synthetase Dynamics**

In the absence of any signal, $I_i$ is expressed at some basal rate $\alpha_I$. As transcription factors, $TF_i$, start to form they bind to promoters $P_i$ either activating or repressing the expression of $I_i$. In particular, binding of the cognate $TF_i$ to the promoter $P_i$ increases the rate of expression to $\alpha_I'$, while binding of the antagonist $TF_{\bar{i}}$ blocks any transcription reducing the rate of expression to zero. Finally the rate of $I_i$ degradation is $\delta_I$.

$$P_i \rightarrow P_i + I_i \quad : \quad \alpha_I \tag{7.3a}$$

$$P_i^{\text{active}} \rightarrow P_i^{\text{active}} + I_i \quad : \quad \alpha_I' \tag{7.3b}$$

$$P_i^{\text{repressed}} \rightarrow P_i^{\text{repressed}} + I_i \quad : \quad 0 \tag{7.3c}$$

$$P_i + TF_i \rightleftharpoons P_i^{\text{active}} \quad : \quad k_{\text{ON}}(\rightarrow), k_{\text{OFF}}(\leftarrow) \tag{7.3d}$$

$$P_i + TF_{\bar{i}} \rightleftharpoons P_i^{\text{repressed}} \quad : \quad k_{\text{ON}}(\rightarrow), k_{\text{OFF}}(\leftarrow) \tag{7.3e}$$

$$I_i \rightarrow \emptyset \quad : \quad \delta_I \tag{7.3f}$$

### 7.3.3 Formulating a Rate Equation Model

We can use the above chemical reaction picture to formulate a rate-equation model, *i.e.*, a system of ordinary differential equations describing the time evolution of the concentration of the different species (denoted by square brackets). Such a model will be valid as long as all participating species are: (i) present in large numbers (so that their concen-

trations can be represented as continuous variables) and (ii) homogeneously distributed within and outside the cell (so that reactions occur homogeneously in time and space). The rate-equation model therefore captures the deterministic behaviour of the gene regulatory network embedded within a homogeneous population. The stochastic nature of the system will be phenomenologically captured by subsequently adding white noise terms to our ordinary differential equations.

**Autoinducer Synthetase Dynamics**

Reactions (7.3) give rise to the following set of rate equations:

$$\frac{d\left[\mathrm{P}_i^{\text{active}}\right]}{dt} = k_{\text{ON}}[\mathrm{TF}_i]\left[\mathrm{P}_i\right] - k_{\text{OFF}}\left[\mathrm{P}_i^{\text{active}}\right], \tag{7.4}$$

$$\frac{d\left[\mathrm{P}_i^{\text{repressed}}\right]}{dt} = k_{\text{ON}}[\mathrm{TF}_{\bar{i}}]\left[\mathrm{P}_i\right] - k_{\text{OFF}}\left[\mathrm{P}_i^{\text{repressed}}\right], \tag{7.5}$$

$$\frac{d\left[\mathrm{P}_i\right]}{dt} = k_{\text{OFF}}\left(\left[\mathrm{P}_i^{\text{active}}\right] + \left[\mathrm{P}_i^{\text{repressed}}\right]\right) - k_{\text{ON}}\left([\mathrm{TF}_{\bar{i}}]\left[\mathrm{P}_i\right] + [\mathrm{TF}_i]\left[\mathrm{P}_i\right]\right), \tag{7.6}$$

$$\frac{d[\mathrm{I}_i]}{dt} = \alpha_I\left[\mathrm{P}_i\right] + \alpha_I'\left[\mathrm{P}_i^{\text{active}}\right] - \delta_I[\mathrm{I}_i]. \tag{7.7}$$

Of course the total number of promoters is conserved so we additionally have

$$\mathrm{P}_i^{\text{total}} = \left[\mathrm{P}_i\right] + \left[\mathrm{P}_i^{\text{active}}\right] + \left[\mathrm{P}_i^{\text{repressed}}\right]. \tag{7.8}$$

In what follows we will assume that binding and unbinding of transcription factors occurs on a short time-scale. This will enable us to eliminate the fast varying variables $\left[\mathrm{P}_i^{\text{active}}\right]$, $\left[\mathrm{P}_i^{\text{repressed}}\right]$ and $[P_i]$ and end up with a single equation describing the slow dynamics of $[\mathrm{I}_i]$.

By setting $d\left[\mathrm{P}_i^{\text{active}}\right]/dt = 0$ and $d\left[\mathrm{P}_i^{\text{repressed}}\right]/dt = 0$ we obtain:

$$\left[\mathrm{P}_i^{\text{active}}\right] = \frac{k_{\text{ON}}}{k_{\text{OFF}}}\left[\mathrm{P}_i\right]\left[\mathrm{TF}_i\right], \tag{7.9}$$

$$\left[\mathrm{P}_i^{\text{repressed}}\right] = \frac{k_{\text{ON}}}{k_{\text{OFF}}}\left[\mathrm{P}_i\right]\left[\mathrm{TF}_{\bar{i}}\right]. \tag{7.10}$$

Substituting the above relations into Eq. (7.8) yields:

$$\left[\mathrm{P}_i\right] = \frac{K_I \mathrm{P}_i^{\text{total}}}{K_I + [\mathrm{TF}_i] + [\mathrm{TF}_{\bar{i}}]}, \tag{7.11}$$

where $K_I \equiv \frac{k_{\text{ON}}}{k_{\text{OFF}}}$. Finally, upon substitution, Eq. (7.7) becomes

$$\begin{aligned} \frac{d[\text{I}_i]}{dt} &= [\text{P}_i^{\text{total}}]\frac{\alpha_I K_I + \alpha_I'[\text{TF}_i]}{K_I + [\text{TF}_i] + [\text{TF}_{\bar{i}}]} - \delta_I[I_i] \\ &= \frac{\bar{\alpha}_I K_I + \bar{\alpha}_I'[\text{TF}_i]}{K_I + [\text{TF}_i] + [\text{TF}_{\bar{i}}]} - \delta_I[I_i], \end{aligned} \tag{7.12}$$

where we have absorbed the quantity $\left[\text{P}_i^{\text{total}}\right]$ into the rates of production, *i.e.*, $\bar{\alpha}_I = \alpha_I \left[\text{P}_i^{\text{total}}\right]$ and $\bar{\alpha}_I' = \alpha_I' \left[\text{P}_i^{\text{total}}\right]$.

**Transcription Factor Formation**

Turning on the dynamics of the transcription factor formation, reactions (7.2) give rise to

$$\frac{d\left[\text{R}_i\text{-AI}_i\right]}{dt} = k_1[\text{R}_i][\text{AI}_i] - k_2\left[\text{R}_i\text{-AI}_i\right]^2, \tag{7.13}$$

$$\frac{d[\text{TF}_i]}{dt} = k_3\left[\text{R}_i\text{-AI}_i\right]^2 - k_4[\text{TF}_i]. \tag{7.14}$$

We will also assume that the characteristic time-scale of the binding events leading to the formation of the transcription factor are fast. Therefore assuming that variables $[\text{R}_i\text{-AI}_i]$ and $[\text{TF}_i]$ are at quasi-steady state and setting the above equations to zero we obtain

$$[\text{TF}_i] = \frac{\left(R_0[\text{AI}_i]\right)^2}{K_D K_C^2}, \tag{7.15}$$

where $K = \frac{k_4}{k_3}$ and $K_C = \frac{k_2}{k_1}$. Furthermore, since the concentration of $R$ does not change due to the internal dynamics of the system we have regarded it as a constant parameter, *viz.* $[R_i] = R_0$.

**Autoinducer Dynamics**

The intra-cellular and extracellular concentrations of $\text{AI}_i$ denoted by $[\text{AI}_i]$ and $[\text{AI}_{i,\text{ext}}]$, respectively are described by the following rate equations

$$\frac{d[\text{AI}_i]_n}{dt} = \alpha_A[\text{I}_i]_n - \delta_A[\text{AI}_i]_n + \frac{\eta}{V_c}([\text{AI}_i]_n - [\text{AI}_{i,\text{ext}}]) \tag{7.16}$$

$$\frac{d[\text{AI}_{i,\text{ext}}]}{dt} = -\delta_{A,\text{ext}}[\text{AI}_{1,\text{ext}}] + \frac{\eta}{V_e}\sum_{n=1}^{N}([\text{AI}_{i,\text{ext}}] - [\text{AI}_i]_n) \tag{7.17}$$

where the subscript $n$ was used to denote individual cells in the population and the sum in the second equation is taken over a population of $N$ cells. Note that the above set of equations states that cells are coupled via the extracellular concentrations of the autoinducer molecules. Our assumption of spatial homogeneity therefore gives rise to an all-to-all coupling; that is autoinducer molecules, irrespective of where they are produced, rapidly diffuse throughout the whole extra-cellular volume and are therefore detected equally by all cells.

Once again invoking the quasi steady-state approximation for variables $[\mathrm{AI}_{1,\mathrm{ext}}]$ and $[\mathrm{AI}_{2,\mathrm{ext}}]$ we can write

$$[\mathrm{AI}_{i,\mathrm{ext}}] = \frac{\eta N/V_e}{\delta_{A,\mathrm{ext}} + \eta N/V_e}\langle[\mathrm{AI}_i]\rangle \equiv Q\langle[\mathrm{AI}_i]\rangle \tag{7.18}$$

where

$$\langle[\mathrm{AI}_i]\rangle = \frac{1}{N}\sum_{n=1}^{N}[\mathrm{AI}_i]_n.$$

The equation describing the dynamics of the intra-cellular concentration of the autoinducer molecules then takes the form

$$\frac{d[\mathrm{AI}_i]_n}{dt} = \alpha_A[\mathrm{I}_i]_n - \delta_A[\mathrm{AI}_i]_n + \frac{\eta}{V_c}([\mathrm{AI}_i]_n - \langle[\mathrm{AI}_i]]\rangle). \tag{7.19}$$

Note that the above equation was obtained assuming that diffusion of autoinducer molecules in and out of the cells is a rather fast process so that quasi-steady state is established for the extracellular concentrations of $\mathrm{AI}_i$. This assumption allows us to regard *mean-field coupling* between cells, i.e, cells are coupled to each other via the mean-field quantities $\langle[\mathrm{AI}_i]\rangle$. Furthermore, parameter $Q$ quantifies the strength of this coupling. In what follows we use $Q$ as a control parameter to study the effect that the population has on internal dynamics of each cell; how this effect feeds back to the population, causing consensus behaviour, and how intrinsic fluctuation affect the population dynamics.

### 7.3.4 Reduced Mean-Field Model

Equations (7.12) and (7.19) along with the relation given by Eq. (7.15) constitute a reduced model describing the deterministic dynamics of the system under the assumptions of mean-field coupling and spatial homogeneity. To take into account the intrinsically stochastic nature of the processes involved in our system we add to this set of equations Gaussian white noise terms, effectively turning them into stochastic differential equations (or *Langevin* equations). This is, indeed, a phenomenological way to proceed since the

intrinsic fluctuations of the system are captured by *ad hoc* terms and are not derived by considering the randomness of the processes involved [141]. However, this phenomenological approach is justified by the scope of our model, which is to study the effect of fluctuations on the population level dynamics and not to accurately capture the stochastic dynamics of the system from first principles.

The final rate-equation model with the addition of Gaussian white noise takes the form:

$$\frac{d[\mathrm{I}_1]_n}{dt} = \frac{\bar{\alpha}_I K + \alpha'_I [\mathrm{AI}_1]_n}{K^2 + [\mathrm{AI}_1]_n^2 + [\mathrm{AI}_2]_n^2} - \delta_I [\mathrm{I}_1] + \sqrt{D_1^I}\xi_n(t) \quad (7.20)$$

$$\frac{d[\mathrm{I}_2]_n}{dt} = \frac{\bar{\alpha}_I K + \alpha'_I [\mathrm{AI}_2]_n}{K^2 + [A_2]_n^2 + [A_1]_n^2} - \delta_I [I_2] + \sqrt{D_2^I}\zeta_n(t) \quad (7.21)$$

$$\frac{d[\mathrm{AI}_1]_n}{dt} = \alpha_A [\mathrm{I}_1]_n - \delta_A [\mathrm{AI}_1]_n + \frac{\eta}{V_c}([\mathrm{AI}_1]_n - Q\langle[\mathrm{AI}_1]\rangle) + \sqrt{D_1^A}\lambda_n(t) \quad (7.22)$$

$$\frac{d[\mathrm{AI}_2]_n}{dt} = \alpha_A [\mathrm{I}_2]_n - \delta_A [\mathrm{AI}_2]_n + \frac{\eta}{V_c}([\mathrm{AI}_2]_n - Q\langle[\mathrm{AI}_2]\rangle) + \sqrt{D_2^A}\kappa_n(t) \quad (7.23)$$

where $K = \dfrac{(K_I K_D)^{1/2} K_C}{R_0}$, and index $n = 1 \ldots N$ denotes the cell. Terms $\xi_n(t)$, $\zeta_n(t)$, $\lambda_n(t)$, $\kappa_n(t)$ are Gaussian white noise with zero mean and delta-peaked auto-correlation functions, *i.e.*,

$$\begin{aligned} \langle \xi_i(t) \rangle &= 0, \quad \langle \xi_i(t)\xi_j(t') \rangle = \delta_{ij}\delta(t-t') \\ \langle \zeta_i(t) \rangle &= 0, \quad \langle \zeta_i(t)\zeta_j(t') \rangle = \delta_{ij}\delta(t-t') \\ \langle \lambda_i(t) \rangle &= 0, \quad \langle \lambda_i(t)\lambda_j(t') \rangle = \delta_{ij}\delta(t-t') \\ \langle \kappa_i(t) \rangle &= 0, \quad \langle \kappa_i(t)\kappa_j(t') \rangle = \delta_{ij}\delta(t-t'). \end{aligned} \quad (7.24)$$

Finally, $D_1^I$, $D_2^I$, $D_1^A$, $D_2^A$ quantify the magnitude of the fluctuations for each chemical species.

### 7.3.5 Numerical Results

To study the effect of coupling on the population dynamics we focus our attention on the mean-field quantity

$$\langle [\mathrm{I}_1] \rangle = \frac{1}{N} \sum_{n=1}^{N} [\mathrm{I}_1]_n. \quad (7.25)$$

We first set to study the deterministic behaviour of the system by setting

$$D_1^A = D_2^A = D_1^I = D_2^I = 0. \quad (7.26)$$

Figure 7.4 (solid and dotted lines) illustrates the long-time steady state behaviour of the population dynamics as a function of the coupling strength $Q$. In the absence of intrinsic noise a transition is observed at some critical value of coupling ($Q = Q_c \approx 0.13$). Below $Q_c$, cells are effectively behaving independently. In particular, $I_1$ and $I_2$ are expressed at basal rates and autoinducers are not present in high enough levels to trigger the quorum response. As $Q$ is increased above the critical value, two stable steady-states appear at the population level. Each branch corresponds to a state in which one of the two quorum sensing modules is activated in bacterial cells and the other one is repressed. Of course, in the absence of noise, the choice between the two stable points depends solely on the initial conditions. As $Q$ reaches unity all cells become synchronised communicating via the same channel.

Intrinsic fluctuations qualitatively change the above picture by shifting transition to consensus behaviour to higher $Q$ values. The above is illustrated in Fig. 7.4 (circles). As before for $Q > Q_c$, each individual triggers a quorum response, activating one of the quorum sensing systems. However, in this case the choice is not fixed. Rather, due to intrinsic fluctuation each cell randomly switches between the two states. Hence at the population level, this is perceived as disorder, with approximately half of the cells occupying each of the two states (see left and centre panels in Fig. 7.5). The behaviour changes once again when the coupling strength exceeds some other critical value, $Q > \bar{Q}_c$. Above $\bar{Q}_c$ coupling is strong enough to make random transitions between the two states less frequent. The population, therefore, relaxes to one of the two consensus states with the majority of the cells communicating through the same quorum sensing system. Random fluctuations can still induce random transitions between the two states, though on a much slower time-scale (see right panel in Fig. 7.5).

### 7.3.6 Numerical Methods

The results presented in Fig. 7.4 and 7.5 were obtained by numerical integration of the reduced mean-field model given by Eqs. (7.20)-(7.23). The parameter values that were used are summarised in Table 7.2.

In the absence of noise, the system was integrated using MATLAB built-in ODE solver (function *ode45* with default settings). The function implements an explicit Runge-Kutta variant with adaptive timestep. For every value of the coupling strength $Q$ ($0$ to $1$ with step size $0.01$), $10$ random sets of initial conditions were prepared in the range $[\mathrm{I}_1]_n = [0, 10]$, $[\mathrm{I}_2]_n = [0, 10]$, $[\mathrm{AI}_1]_n = [0, 10]$, $[\mathrm{AI}_2]_n = [0, 10]$ along with the set $[\mathrm{I}_1]_n = 0$, $[\mathrm{I}_2]_n = 0$, $[\mathrm{AI}_1]_n = 0$, $[\mathrm{AI}_2]_n = 0$. The system was then integrated with

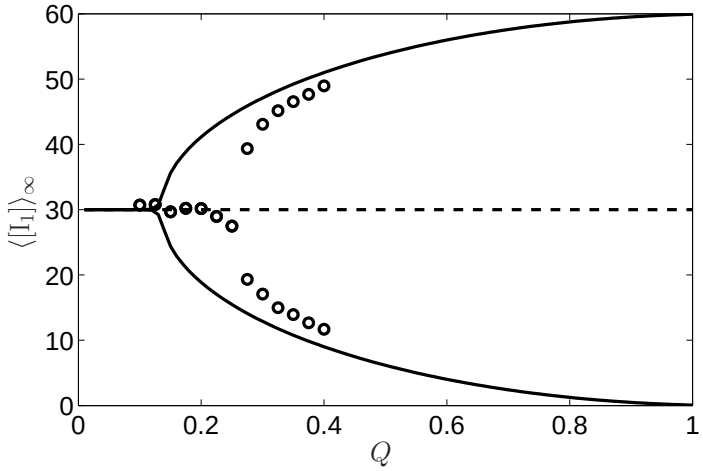



Figure 7.4: Bifurcation diagram for the long time (steady state) levels of $\langle [\mathrm{I}_1] \rangle$. Solid lines indicate the behaviour in the absence of intrinsic fluctuations. At some critical coupling strength $Q_c$ the population undergoes a supercritical pitchfork bifurcation. Above $Q_c$ a quorum response is triggered leading to two steady states, each one corresponding to cells sharing the same communication channel. In the presence of noise (circles) the critical coupling strength for a majority consensus to be reached is shifted. A higher coupling strength is necessary for the population to reach the consensus regime.

these initial condition and in each case the steady-state, mean-field quantity $\langle [I_1] \rangle_\infty$ (see Eq. 7.25) was recorded. We ensure steady state has been reached by checking that all system variables do not change more than $10^{-3}$ between successive time-steps. In all cases integration up to $\tau = 500$ fulfilled this criterion.

For numerical integration in the presence of noise, the Euler method was used as implemented in the XPPAUT software (version $5.98$) with time step $\delta t = 10^{-3}$. For every value of $Q$ ($0.1$ to $0.4$ with step size $0.025$), $10^3$ independent runs were performed using as initial conditions $[\mathrm{I}_1]_n = 0$, $[\mathrm{I}_2]_n = 0$, $[\mathrm{AI}_1]_n = 0$, $[\mathrm{AI}_2]_n = 0$. In each run integration was performed as before up $\tau = 500$, allowing the system to reach steady state and $\langle [I_1] \rangle^\infty$ was calculated. The data obtained for each value of $Q$ were non-parametrically fitted to a probability distribution using MATLAB built-in function *ksdensity* (default settings). The function essentially computes a smooth estimate of the probability density function from the histogram using a Gaussian kernel. The circles shown in Fig. 7.4 denote the position and number of peaks in the estimated probability distribution for each value of $Q$.

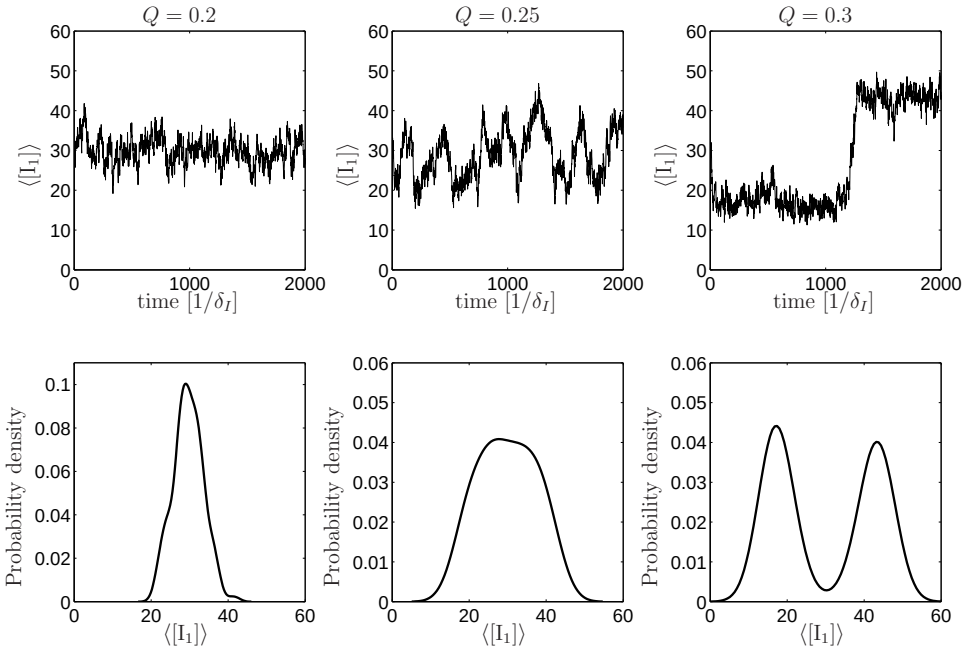



Figure 7.5: Time traces and the stationary distribution of the mean-field quantity $\langle[\mathrm{I}_1]\rangle$ for different values of $Q$. For coupling strengths high enough to trigger a quorum response, the dynamics at the population level initially appear disordered due to random transitions of individual cells between the two states (left panel). Within a critical regime of coupling the distribution flattens (centre panel) and above this regime the two meta-stable states of the population are clearly discerned (right panel). The stationary distributions presented in the second row are non-parametric fits to simulation data (see section 7.3.6).

| **Parameter** | **Value** | **Dimensionless value** | Reference |
|---|---|---|---|
| time ($t$) | sec | $\tau = t \cdot \delta_I = t/t^*$ | |
| $\alpha_I$ | $0.02$ nM/sec | $60$ nM | [56] |
| $\alpha'_I$ | $0.002$ nM/sec | $30$ nM | Tunable parameter |
| $\alpha_A$ | $0.01$ sec$^{-1}$ | $33$ | [56] |
| $\delta_I$ | $0.0003$ sec$^{-1}$ | $1$ | [56, 91] |
| $\delta_A$ | $0.0003$ sec$^{-1}$ | $1$ | [56, 91] |
| $\delta_{Ae}$ | $0.0003$ sec$^{-1}$ | $1$ | [56, 91] |
| $K_I$ | $10^2$ nM | | [56, 91] |
| $K_C$ | $10^2$ nM | | [56, 91] |
| $K_D$ | $10^2$ nM | | [56, 91] |
| $R_0$ | $10^2$ nM | | Tunable parameter |
| $K$ | $100$ nM | | Tunable parameter |
| $V_c$ | $10^{-15}$ L | | [3] |
| $N$ | $40$ | | |
| $\eta$ | $6 \cdot 10^{-21}$ m$^3$/sec | $15V_c$ | [91] |
| $\sqrt{D_1^A}$ - $\sqrt{D_2^A}$ | $10$ | | |
| $\sqrt{D_1^I}$ - $\sqrt{D_2^I}$ | $5$ | | |

Table 7.2: Parameters used in the reduced rate-equation model.

## 7.4 An Ising Model of the Population Dynamics

The results presented in the preceding section illustrate that sub-cellular fluctuations hinder the ability of cells to coordinate their behaviour and achieve consensus. Under weak coupling (yet strong enough to trigger the quorum response) intrinsic fluctuations induced random switching between the two lifestyles in individual cells. In this regime, the behaviour at the population level appears disordered with roughly one half of the population occupying each state. As the coupling strength is increased the population undergoes a transition into a ordered state where the majority of cells occupy one of the two states. In the presence of noise, therefore, higher values of coupling strength are necessary for the population to reach the consensus regime.

In this section, we present and study a coarse grained model that demonstrates quantitatively similar behaviour to the bacterial quorum. In a nutshell, the model considers cells in the quorum as a population of interacting bistable switches. To the physicist this coarse grained picture will bear close resemblance to an Ising-type model capturing the collective behaviour of mean-field coupled spins. We use this Ising-type model to study the transition between the two regimes of behaviour in greater detail. For a finite systems this transition is blurred in the region around the critical coupling strength. We find a condition that marks the clear transition to the ordered state, linking the coupling strength

to the magnitude of fluctuations and size of the population.

### 7.4.1 Master Equation Formulation

Consider a population of $N$ cells capable of occupying two distinct states, $A$ and $B$. The number of cells occupying each state is denoted by $n_A$ and $n_B = N - n_A$, respectively. Furthermore, we allow cells to interact with each other. For simplicity, we restrict ourselves to a mean-field coupling and ignore any spatial effects. In particular, we regard that cells occupying state $A(B)$ exert a force $+F(-F)$ on every other cell. Furthermore, we limit the forcing capability of each cell by imposing a functional relationship of the force magnitude on the size of the population:

$$|F| = \frac{F_0}{N + K} \tag{7.27}$$

where $K$ is some arbitrary, non-negative constant. For $K \gg N$ interactions between cells are negligible, while for $K \ll N$ each cell exerts a maximum force $F_0/N$ on every other. The total force, $F_T$, exerted on each cell is therefore given by

$$F_T = |F|(n_A - n_B) = \frac{NF_0}{N + K}\frac{2n_A - N}{N} \equiv Q\frac{2n_A - N}{N} \tag{7.28}$$

Here, parameter $Q$ quantifies the coupling strength between the cells and will be used to study the effect of the interactions on the dynamics of the population.

Since we consider the size of the population fixed we can study its dynamics by considering the time evolution of a single variable, for example $n_A$, $n_B$ or $m \equiv n_A - n_B$. In terms of $n_A$, the Master equation describing the stochastic dynamics of the system is

$$\begin{aligned} \frac{dP(n)}{dt} &= (N - n + 1)W_+(n-1)P(n-1) \\ &\quad +(n+1)W_-(n+1)P(n+1) \\ &\quad -\left[(N-n)W_+(n) + nW_-(n)\right]P(n) \end{aligned} \tag{7.29}$$

where $P(n) = P(n_A = n, t|n_A = n_o, t_0)$ is the probability of observing $n$ cells occupying state $A$ at time $t$ given that at time $t_0$ there were $n_0$ such cells. Moreover, the transition rates $W_\pm$ are given by

$$W_\pm(n) = w_\pm \exp\left[\pm\beta Q\frac{2n - N}{N}\right]. \tag{7.30}$$

Prefactors $w_+$ and $w_-$ represent the basal switching rates from state $B$ to state $A$ and vice

versa when the net force acting on each cell is $F_T = 0$. One can regard them as being $\sim \exp[-\beta E_b]$, where $E_b$ is the energy barrier the cell must overcome to get from one state to the other and $\beta$ a temperature parameter quantifying the magnitude of the intrinsic fluctuations driving the transitions. The symmetric construction of the synthetic circuit (see Fig. 7.2) enables us to consider a symmetric energy potential so that $w_+ = w_- = w$. When $F_T \neq 0$ the exponential factor in Eq. (7.30) tilts the energy landscape, hence biasing the transitions to one of the two states.

### 7.4.2 The Macroscopic Behaviour

As described in Chapter 3 (3.3.3), a rate equation describing the macroscopic dynamics ($N \to \infty$) of the system, $\phi(t)$, can be obtained as the lowest order terms in the system-size expansion of the Master equation. In our case, Eq. (7.29) [141] yields

$$\frac{d\phi(t)}{dt} = -V'(\phi(t)), \tag{7.31}$$

where

$$V'(\phi(t)) = \phi \exp\left[-\beta Q(2\phi - 1)\right] - (1-\phi)\exp\left[\beta Q(2\phi - 1)\right]. \tag{7.32}$$

The function $V(\phi)$ can be considered as a potential landscape driving the time evolution of variable $\phi$, *i.e.*, $\phi$ will move towards values minimising $V(\phi)$.

One can study the long time (steady state) behaviour of the system by setting $d\phi/dt = 0$, and looking for the steady state points $\phi_s$ as solutions of the equation $V'(\phi_s) = 0$. Inspection yields the trivial root $\phi_s = 1/2$, however, a closed formula for any other root is not possible. Alternatively, one can Taylor expand $V'(\phi)$ around $\phi_s = 1/2$ and look for roots in this neighbourhood. For clarity we use the transformation $\bar{\phi}_s = 2\phi_s - 1$ and obtain

$$(\beta Q - 1)\bar{\phi}_s - \frac{(\beta Q)^2}{2}\left(1 - \frac{\beta Q}{3}\right)\bar{\phi}_s^3 + \mathcal{O}(\bar{\phi}_s^5) = 0 \tag{7.33}$$

The first two terms of the Taylor expansion suffice to describe the behaviour of system. We note that for $\beta Q$ close to unity the coefficient of the cubic term is strictly negative and therefore the number of roots depends solely on the sign of the linear term. If $\beta Q \leq 1$ a single root exists corresponding to a stable fixed point ($\phi_s = 1/2$) (see left and centre panel in Fig. 7.7). In the case $\beta Q > 1$ the picture is altered: two stable steady states exist separated by an unstable one at $\phi_s = 1/2$ (see right panel in Fig. 7.7). This describes a

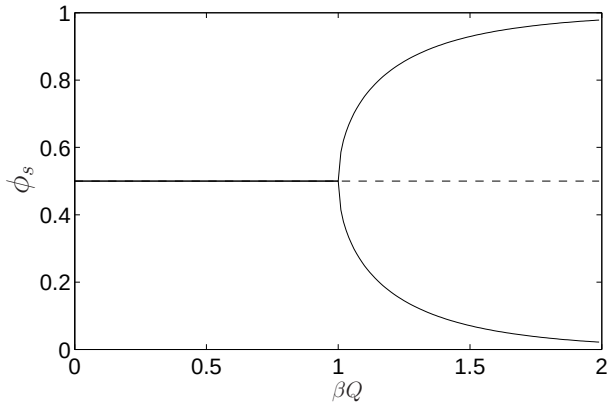



Figure 7.6: Bifurcation diagram for the deterministic behaviour ($N \to \infty$) of the Ising-type model. The system undergoes a supercritical pitchfork bifurcation at $Q = 1/\beta$. For $Q \leq 1/\beta$ the population exists in a disordered state where half of the cells occupy each state ($\phi_s = 1/2$). For $Q > 1/\beta$ the majority of the population occupies one of the two states. The stationary points $\phi_s$ were obtained by numerically solving $V'(\phi_s) = 0$ [see Eq. (7.32)], using Gauss-Newton algorithm (MATLAB built-in function *fsolve*).

supercritical pitchfork bifurcation at $\beta Q = 1$ [132], with the population crossing from the disordered (low coupling) state to the ordered (high-coupling) one. The bifurcation diagram obtained by numerically solving $V'(\phi_s) = 0$ is illustrated in Fig. 7.6.

### 7.4.3 Stationary Distribution

So far we have presented the deterministic behaviour of the model in the limit $N \to \infty$. Now we turn and study the behaviour of the model for finite $N$. Once again we focus on the long time limit ($t \to \infty$) and present analytic results for the stationary distribution $P_s(n) = P(n_A = n, t = \infty | n_A = n_0, t = t_0)$. These results will used in subsequent sections to obtain the transition times between the two stable states.

As seen in Chapter 3 (3.3.2) for finite systems the stationary distribution, $P_s$ obeys the recursion relation [50, 141]

$$P_s(n) = \frac{(N-n+1)W_+(n-1)}{nW_-(n)} P_s(n-1), \tag{7.34}$$

from which one obtains

$$P_s(n) = \frac{(N-n+1)(N-n+2)\cdots N}{n!} \cdot \frac{W_+(n-1)\cdots W_+(0)}{W_-(n)\cdots W_1} \cdot P_s(0) \tag{7.35}$$

Finally, using the definition of the transition probabilities [Eq (7.30)] the stationary dis-

tribution can be written as

$$P_s = \frac{1}{\mathcal{N}}\binom{N}{n}\exp\left[-\frac{2\beta Q}{N}n(N-n)\right], \tag{7.36}$$

where $\mathcal{N}$ is the normalisation constant such that

$$\sum_{n=0}^{N} P_s(n) = 1. \tag{7.37}$$

In general, $\mathcal{N}$ depends on parameters of the model, namely $\beta$, $Q$ and $N$. For example, in the trivial case of $Q = 0$, one readily obtains $\mathcal{N} = 1/2^N$. As expected, in this scenario the stationary distribution is just the binomial distribution for equally likely events. Below, we provide approximations for $\mathcal{N}$ in the three regimes of behaviour, $\beta Q < 1$, $\beta Q = 1$ and $\beta Q > 1$, using standard perturbation techniques. Our approximation are valid for sufficiently large populations for which the discrete quantity $x = n/N$ can be treated as a continuous variable.

We first express $P_s(n)$ as

$$P_s(n) = \frac{\exp\left[-E(n)\right]}{\mathcal{N}}, \tag{7.38}$$

where $E(n)$ can be though as the energy landscape. From Eq. (7.36) one readily sees that

$$E(n) = -\ln\left[\frac{N!}{n!(N-n)!}\right] + \frac{2\beta Q}{N}n(n-N). \tag{7.39}$$

Writing the above expression in terms of the intensive variable $x = n/N$ and Taylor expanding yields

$$\begin{aligned} E(n) &\approx \left[\frac{n}{N}\ln\left(\frac{n/N}{1-n/N}\right) + \log\left(1-\frac{n}{N}\right) - 2Q\tfrac{n}{N}\left(\frac{n}{N}-1\right)\right]N \\ &\quad \left[x\ln\left(\frac{x}{1-x}\right) + \log\left(1-x\right) - 2Qx\left(x-1\right)\right]N \\ &= U(x)N \end{aligned} \tag{7.40}$$

Hence, in the continuum limit the stationary distribution becomes

$$P_s(x) = \frac{\exp\left[-U(x)N\right]}{\mathcal{C}}, \tag{7.41}$$

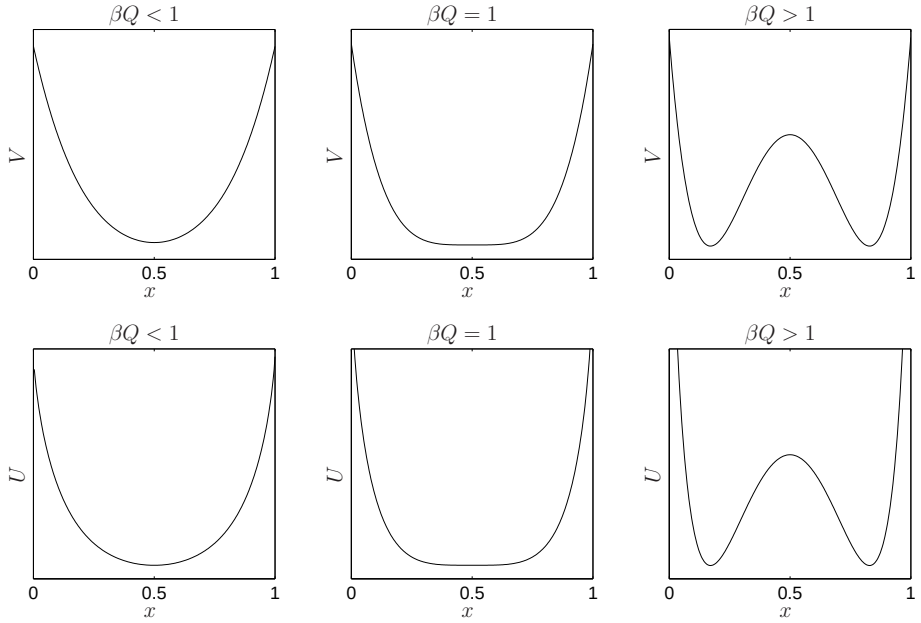



Figure 7.7: The shape of $V$ and $U$ for $\beta Q < 1$, $\beta Q = 1$ and $\beta Q > 1$. $V$ corresponds to the potential landscape driving the macroscopic ($N \to \infty$) behaviour of the population. Minima of $V$ correspond to stable steady state solutions of Eq. (7.32), while maxima to unstable ones. The energy landscape $U$ dictates the stationary distribution for finite yet large population sizes, $N \gg 1$ [see Eq. (7.41)]. For $Q \leq 1$, both $V$ and $U$ have a single minimum at (or around) $x = 1/2$ (left and centre column). For $Q > 1$ two minima exist at $x = x_a > 1/2$ and $x = x_b < 1/2$ separated by a maximum at $x = 1/2$ (right panels).

where $\mathcal{C}$ and can be evaluated from the normalisation constrain

$$\int_0^1 P_s(x)dx \;=\; 1 \;\Rightarrow\; \mathcal{C} \;=\; \int_0^1 \exp\left[-U(x)N\right]. \tag{7.42}$$

Figure 7.7 illustrates the general shape of $U(x)$. $U(x)$ is closely related to the potential $V$ [Eq. (7.32)] driving the macroscopic ($N \to \infty$) behaviour of the population. In particular, they both undergo the same change of shape as $\beta Q = 1$ is crossed and both possess the same minima and maxima. Therefore, one should expect that for finite system sizes the stationary distribution of the system is peaked around the stable points of the macroscopic behaviour.

**Case I ($\beta Q < 1$)**

For $Q < 1$ the integral in Eq. (7.42) can be evaluated using the Laplace method [34]. The method relies on approximating the integral of a function that possesses a sharp peak at

some point with the integral of its parabolic approximation around that point. In our case, for large $N$ the mass of the probability density function is located around the deterministic stable point $x = 1/2$. This allows us to write $P_s(x)$ as

$$P_s(x) \approx \begin{cases} \frac{1}{\mathcal{C}} \exp\{-[U(1/2) - (x-1/2)^2 U''(1/2)]N\} & |x - 1/2| \ll 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{7.43}$$

Using the above equation we can evaluate the integral in Eq. (7.42) as

$$\begin{aligned} \mathcal{C} &\approx \int_0^1 \exp\left\{-[U(1/2) - (x-1/2)^2 U''(1/2)]N\right\} dx \\ &\approx \int_{-\infty}^{\infty} \exp\left\{-[U(1/2) - (x-1/2)^2 U''(1/2)]N\right\} dx \\ &= 2^N \exp\left[\frac{-\beta Q N}{2}\right] \sqrt{\frac{\pi}{N(1-\beta Q)}}. \end{aligned} \tag{7.44}$$

We note that in the second step of the above calculation, the limits of the integral were replaced by $\pm\infty$; errors introduced at this point are negligible since the contribution from any region outside the neighbourhood of $x = 1/2$ are expected to be exponentially small.

Substituting back to Eq. (7.41) one obtains a Gaussian stationary distribution

$$P_s(x) \approx \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{7.45}$$

with $\mu = 1/2$ and $\sigma^2 = \dfrac{1}{4N(1-\beta Q)}$. Alternatively, the same result can be obtained by performing the system size expansion [141] on Eq. (7.29).

**Case II ($\beta Q = 1$)**

Using similar arguments one can obtain an approximation for the stationary distribution in the case $Q = Q_c = 1/\beta$. Since $Q_c$ is the critical point where the bifurcation occurs the second and third derivatives of $U(x)$ vanish at $x = 1/2$. Hence the following approximation for $P_s(x)$ should be used

$$P_s(n) \approx \begin{cases} \frac{1}{\mathcal{C}} \exp\left\{-[U(1/2) - (x-1/2)^4 U^{(4)}(1/2)/24]N\right\} & |x - 1/2| \ll 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{7.46}$$

The integration yields

$$\begin{aligned} \mathcal{C} &\approx \int_0^1 P_s(x)dx \\ &\approx \int_{-\infty}^{\infty} P_s(x)dx \\ &= 2^N \exp\left[\frac{-N}{2}\right] \frac{\Gamma\left(\frac{5}{4}\right)\sqrt{2\sqrt{3}}}{N^{1/4}}. \end{aligned} \tag{7.47}$$

From the above one can easily calculate the variance to be

$$\begin{aligned} \sigma^2 &= \int_0^1 \left(x-\frac{1}{2}\right)^2 P_s(x) \\ &\approx \int_{-\infty}^{+\infty} \left(x-\frac{1}{2}\right)^2 P_s(x) \\ &= \frac{2\Gamma\left(\frac{7}{4}\right)}{\sqrt{3N}\Gamma\left(\frac{1}{4}\right)}, \end{aligned} \tag{7.48}$$

where $\Gamma(x)$ is the Gamma function. Therefore at the critical point $Q_c$, stationary fluctuations around the mean are amplified, *i.e.*, $\sigma \sim N^{-1/4}$ rather than $N^{-1/2}$ [as in Eq. (7.45)].

**Case III ($\beta Q > 1$)**

For $\beta Q > 1$, minor complications are introduced due to the existence of two minima in the shape of $U(x)$ at $x_a < 1/2$ and $x_b > 1/2$. From Eq. (7.41) it is readily seen that the two minima in $U(x)$ correspond to maxima of the stationary distribution $P_s$. That is the probability mass is concentrated around the points $x_a$ and $x_b$. Hence, to evaluate the integral in Eq. (7.42) one must make use of the following parabolic approximation

$$P_s(n) = \begin{cases} \dfrac{1}{\mathcal{C}} \exp\left\{-[U(x_a)-(x-x_a)^2 U''(x_a)/2]N\right\} & |x-x_a| \ll 1 \\ \dfrac{1}{\mathcal{C}} \exp\left\{-[U(x_b)-(x-x_b)^2 U''(x_b)/2]N\right\} & |x-x_b| \ll 1 \\ 0 & \text{elsewhere}, \end{cases} \tag{7.49}$$

and the normalisation constant is given by

$$\begin{aligned} \mathcal{C} &\approx \int_0^{1/2} e^{-[U(x_b)-(x-x_b)^2U''(x_b)/2]N}dx + \int_{1/2}^{1} e^{-[U(x_a)-(x-x_a)^2U''(x_a)/2]N}dx \\ &\approx \int_{-\infty}^{1/2} e^{-[U(x_b)-(x-x_b)^2U''(x_b)/2]N}dx + \int_{1/2}^{\infty} e^{-[U(x_a)-(x-x_a)^2U''(x_a)/2]N}dx. \end{aligned} \tag{7.50}$$

Points $x_a$ and $x_b$ coincide with the stable stationary points corresponding to the macroscopic behaviour of the model. Although there is no close formula giving them as a function of $Q$, they can approximated numerically, *e.g.*, using Newton's method [34]. The above formulae can consequently be used to obtain $\mathcal{C}$. Similar to $Q < 1/\beta$, the approximation yields a stationary distribution that will be the mixture of two Gaussian peaks centred around $x_a$ and $x_b$ and with widths that scales like $N^{-1/2}$.

**Some Final Remarks**

At this point we should note that a more careful examination of the validity of our approximations for $Q \lesseqgtr 1/\beta$ is needed. To illustrate our point we note that near $x = 1/2$, $P_s(x)$ can be approximated (keeping up to $4^{\text{th}}$ order terms) by

$$P_s(x) \approx \frac{2^N e^{-\frac{\beta Q N}{2}}}{\mathcal{C}} \exp\left[-2N(1-\beta Q)(x-1/2)^2 - \frac{4N}{3}(x-1/2)^4\right]. \tag{7.51}$$

In our treatment so far, the use of the parabolic approximation, implicitly assumed that for all $Q < 1/\beta$ the quadratic term in the exponent is the dominant one. Let us now be a bit more precise. As $Q \to 1/\beta$ from below, the peak of $P_s$ around $x = 1/2$ becomes wider and wider, and as we have seen the width of the distribution at the critical point scales like $N^{-1/4}$. Therefore, to accurately capture the shape of the distribution at all times the approximation should be valid for $|x - 1/2| \sim N^{-1/4}$. Now, by comparing the two terms in the exponent of Eq. (7.51) it is evident that if

$$1 - \beta Q \gg \frac{1}{\sqrt{N}}, \tag{7.52}$$

the quadratic term is the leading term whereas for

$$1 - \beta Q \ll \frac{1}{\sqrt{N}}, \tag{7.53}$$

the fourth order term dominates. Therefore, when $1-\beta Q \lesssim N^{-1/2}$ more accurate approximations can be obtained by Taylor expanding the exponential factor in Eq. (7.51)

$$P_s(x) \approx \frac{2^N}{\mathcal{C}} e^{-\frac{\beta Q N}{2}} e^{-\frac{4N}{3}(x-1/2)^4} \left[1 + \frac{N(1-Q)}{2}(1 + 4(x-1/2)^2 + \cdots)\right]. \tag{7.54}$$

It should be noted that the above approximation scheme is equally valid as $Q$ approaches unity from above. To illustrate this point further, from Eq. (7.33) we find that as $\beta Q$ approaches unity $x_a$ and $x_b$ are given by

$$x_a \approx \frac{1-\sqrt{3(\beta Q-1)}}{2}, \tag{7.55a}$$

$$x_b \approx \frac{1+\sqrt{3(\beta Q-1)}}{2}. \tag{7.55b}$$

Hence, the distance separating the two points scales like

$$\delta = x_a - x_b \sim (\beta Q - 1)^{1/2}. \tag{7.56}$$

Now, the condition $1-\beta Q \lesssim N^{-1/2}$ can be translated into $\delta \lesssim N^{-1/4}$, *i.e.*, the two maxima being sufficiently close together can be approximated as a single peak.

Summarising, the validity of approximations presented above for the $\beta Q \lessgtr 1$ cases, does not depend solely on the $N \gg 1$ condition but also on how $Q$ approaches the critical point. In particular, they accurately capture the stationary distribution provided that

$$|1-\beta Q| \gg \frac{1}{\sqrt{N}}. \tag{7.57}$$

Otherwise alternative approximation schemes [see Eq. (7.54)] are more suitable.

### 7.4.4 Transition Times in the $\beta Q > 1$ Regime

As we have seen for $\beta Q > 1$, the energy landscape $U(x)$ possesses two wells which correspond to the two *meta-stable states* of the system. Therefore, starting from some initial configuration the system will end up jittering in one of the two wells of $U(x)$. Of course giant fluctuations can still induce random transitions between the two stable states (hence the term meta-stable). In this section we examine the statistics of the transition times between the two wells, a well studied problem tackled in Kramer's rate theory [50, 141]. We use this result to provide the physical picture behind the transition to the ordered (or majority consensus) regime observed in our results for the gene regulatory network

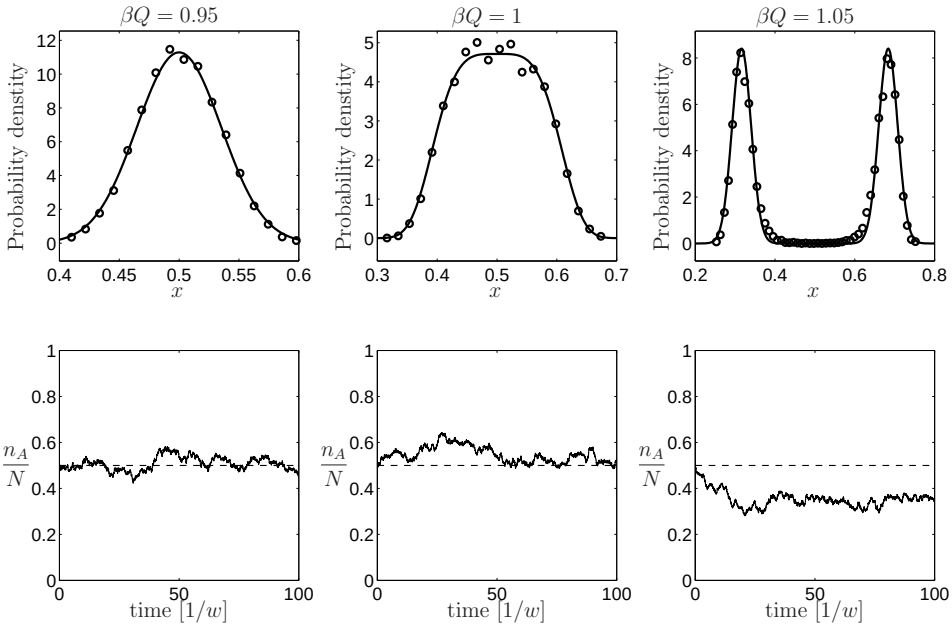



Figure 7.8: The stationary distribution $P_s(x)$ and time traces of the system for different values of $\beta Q$ and $N \gg 1$. Solid lines corresponds to analytic approximations while markers denote results obtained from stochastic simulations of the model ($10^4$ independent runs). For the simulations $N = 4000$ was used.

(see section 7.3.5).

**The First Passage Problem**

As we have discussed in Chapter 3 ( 3.4), first passage theory [116] states that the mean time needed for the system originally occupying state $n = n_1$ to reach state $n = n_2$ is given by

$$\mathcal{T}(n_1 \to n_2) = \sum_{n=n_1}^{n_2} \frac{1}{W_+(n)P_s(n)} \sum_{n'=0}^{n} P_s(n'). \tag{7.58}$$

This result allows us to calculate the mean transition time between the two stable points as

$$\tau = \mathcal{T}(n_b \to n_a). \tag{7.59}$$

where $n_a$ and $n_b$ correspond to the two minima of the energetic landscape $U(n)$.

The above yield an exact result of the problem as defined by the formulation the Eq. (7.29). However, to evaluate the sums and obtain and closed formula for $\tau$ we once again make use of the large $N$ approximation to enable us to treat our model as a continuous one. Hence. in terms of the intensive variable $x = n/N$, the transition time is given by

$$\tau \approx \int_{x_b}^{x_a} \frac{N dx}{W_+(x)P_s(x)} \int_0^x N P_s(x')dx', \tag{7.60}$$

where we have replaced sums by integrals and $x_a$ and $x_b$ are the minima of $U(x)$. Once again the integrals that appear in Eq. (7.60) can be evaluated asymptotically using the parabolic approximation [34, 141]. In particular, the main contribution for the outer integral comes from the neighbourhood around $x = 1/2$ where $1/P_s(x)$ demonstrates a maximum. The contributions from $W_+(x)$ are negligible since it varies slowly compared to $P_s(x)$,; hence, it can safely be replaced by $W_+(1/2) = N/2$ (higher order approximations can however be used). Subsequently, the inner integral is large around $x = x_b$ and otherwise exponentially smaller. Therefore, $\tau$ can be written as

$$\begin{aligned} \tau &\approx \int_{-\infty}^{+\infty} 2N e^{N\left[U(1/2) - \frac{|U''(1/2)|}{2}(x-1/2)^2\right]} dx \int_{-\infty}^{+\infty} e^{N\left[-U(x_b) - \frac{|U''(x_b)|}{2}(x-x_b)^2\right]} dx \\ &= \frac{4N\pi}{\sqrt{U''(x_b)|U''(1/2)|}} \exp\left[\frac{U(1/2) - U(x_b)}{N}\right]. \end{aligned} \tag{7.61}$$

This is a well celebrated result of Kramer's rate theory. It gives the rate of transition $r = 1/\tau$ in terms of some general characteristics of the energy landscape $U(x)$. In particular, the rate includes a prefactor that depends on the curvature near the maximum and at the

bottom of the well. The flatter these areas are the harder the transitions become. The exponential term is the Arrhenius factor and depends solely on the height of the transition barrier.

In the above treatment we have tacitly assumed that points $x_a$, $x_b$ and $x = 1/2$ are well separated. This assumption allowed us to evaluate the integrals in Eq. (7.61) by using the parabolic approximation. This assumption, however, imposes some additional conditions on the validity of our result. In particular, the width of each peak of $P_s(x)$ obtained by the parabolic approximation is $\delta_0 \sim 1/\sqrt{N}$. One would therefore additionally require

$$\delta \equiv x_a - x_b \gg \delta_0 \tag{7.62}$$

so that the two points $x_a$ and $x_b$ are sufficiently far to allow a clear distinction of the two peaks. As we have seen when $\beta Q \to 1^+$ one obtains

$$x_b \approx \frac{1 - \sqrt{3(\beta Q - 1)}}{2}, \tag{7.63a}$$

$$x_a \approx \frac{1 + \sqrt{3(\beta Q - 1)}}{2}. \tag{7.63b}$$

Therefore, the distance between the two points is given by

$$\delta \sim \sqrt{3(\beta Q - 1)}, \tag{7.64}$$

and the condition for two peaks to be well separates takes the form

$$\beta Q \gg 1 + 1/\sqrt{N}. \tag{7.65}$$

When the condition given in Eq. (7.65) breaks down a more appropriate scheme for the calculation of $\tau$ is needed. It involves inclusion of up to fourth order terms in the evaluation of the integral in Eq. (7.60) [118].

### The Physical Picture Behind the Transition to the Consensus Regime

The physical picture behind the transition to the ordered regime is indeed a simple one, involving the separation between two time-scales [141]. The first time-scale is the one set by $\tau$ at which transitions between the two meta-stable states are observed. We calculate

for $Q \to 1^+$

$$U''(x_a) \approx 8(\beta Q - 1), \tag{7.66a}$$

$$U''(1/2) = -4(\beta Q - 1), \tag{7.66b}$$

$$U(1/2) - U(x_a) = \frac{3}{4}(\beta Q - 1)^2. \tag{7.66c}$$

and hence from Eq. (7.60) one obtains

$$\tau \approx \frac{\pi}{\sqrt{2}N(Q-1)} \exp\left[\frac{3}{4}(Q-1)^2 N\right]. \tag{7.67}$$

The second time-scale, $\tau_{eq}$, is determined by the rate at which equilibrium is established around each stable point. This is essentially the autocorrelation time of the process and depends on the curvature at of $U(x)$ at $x = x_a$ and $x = x_b$, respectively [141], that is

$$\tau_{eq} = \frac{1}{NU''(x_a)} \sim \frac{1}{N(\beta Q - 1)}. \tag{7.68}$$

It is easily verified that the condition given by Eq. (7.65) ensures a clear separation between these two time-scales,

$$\tau \gg \tau_{eq}. \tag{7.69}$$

In other words the system rapidly equilibrates around one of the two stable points before giant fluctuations induce a transition to the other one. Equation (7.65), therefore, gives a relation between the coupling strength $Q$, the noise intensity $\beta$ and the population size $N$ ensuring a clear transition into the ordered regime. When this condition breaks the distinction between the two meta-stable states of the population is not clear.

### 7.4.5 A Two Population Model

So far we have considered an Ising-type model demonstrating qualitatively similar behaviour to a population of cells bearing the gene regulatory network presented in Fig. 7.2. We now turn briefly to the alternative design presented in Fig. 7.3. As discussed this design gives rise to two competing bacterial populations bearing resemblance to the case of competing *S. aurues* strains. Using a similar coarse-grained, Ising-type model we demonstrate the relationship between the two designs.

At a coarse grained level, we can characterise the behaviour of the two-population model by considering a mixed population of two distinct cell types $A$ and $B$. The number of type-A and type-B cells is denoted by $N_A$ and $N_B$, respectively. For reasons that will

become apparent below we set the total size of the population to $2N$, that is $N_A + N_B = 2N$.

Both cell types are capable of switching between two states: a social (quorum-aware) one and a solitary (quorum-unaware) one. We use $n_A$ and $n_B$ to denote the number of social cells of type A and B respectively, whereas the number of solitary one is denoted by $\bar{n}_A$, $\bar{n}_B$. As in the model presented above, we allow all-to-all interactions but in this case, we allow only social cells to exert forces. In particular, social cells of each type exert forces on their own kind pulling solitary ones into social behaviour and keeping social ones in their current state. Furthermore, they interact with social cell on the other kind pushing them towards isolation. The total force exerted on each cell is therefore

$$F_T = Q\frac{n_A - n_B}{2N}, \tag{7.70}$$

where once again Q quantifies the strength of coupling between individual cells.

The dynamics of $P(n_1, n_2) = P(n_A = n_1, n_B = n_2, t | n_A = n_A^0, n_B = n_B^0, t)$, the probability of observing $n_1$ type-A and $n_2$ type-B social cell at time $t$ having initially ($t = t_0$) $n_A^0$ and $n_B^0$, respectively are described by the Master equation

$$\begin{aligned}\frac{dP(n_1,n_2)}{dt} = &(\mathbf{E}_A^{-1} - 1)(N_A - n_1)W_+^A(n_1,n_2)P(n_1,n_2)\\ &+(\mathbf{E}_A^{+1} - 1)n_1 W_-^A(n_1,n_2)P(n_1,n_2)\\ &+(\mathbf{E}_B^{-1} - 1)(N_B - n_2)W_+^B(n_1,n_2)P(n_1,n_2)\\ &+(\mathbf{E}_B^{+1} - 1)n_2 W_-^B(n_1,n_2)P(n_1,n_2)\end{aligned} \tag{7.71}$$

where for compactness we introduced the step operators [141]

$$\mathbf{E}_A^a f(n_1, n_2) = f(n_1 + a, n_2), \tag{7.72a}$$

$$\mathbf{E}_B^a f(n_1, n_2) = f(n_1, n_2 + a). \tag{7.72b}$$

Furthermore the transition rates are given by

$$W_\pm^A(n_1, n_2) = w_\pm^A \exp\left[\pm\beta_A Q\frac{n_1 - n_2}{N}\right], \tag{7.73a}$$

$$W_\pm^B(n_1, n_2) = w_\pm^B \exp\left[\mp\beta_B Q\frac{n_1 - n_2}{N}\right]. \tag{7.73b}$$

As before, $w_\pm^A$ and $w_\pm^B$ represent the basal switching rates between two states for the two cell types, when the net force acting on each individual is $F_T = 0$. The exponential factor, captures the change of the basal rates due to interaction forces.

Once again one can readily obtain the deterministic law describing the dynamics of the system as $N_A \to \infty$ and $N_B \to \infty$. In terms of the extensive variable $y_1 = n_1/2N$ and $y_2 = n_2/2N$, one has

$$\frac{dy_1}{dt} = \left(\frac{N_A}{2N} - y_1\right) W_+^A(y_1, y_2) - y_1 W_-^A(n_1, n_2) \tag{7.74a}$$

$$\frac{dy_2}{dt} = \left(\frac{N_B}{2N} - y_2\right) W_+^B(y_1, y_2) - y_2 W_-^B(n_1, n_2) \tag{7.74b}$$

where $W_+^A(y_1, y_2)$ and $W_+^A(y_1, y_2)$ are the rates given by Eq. (7.73) rewritten in terms of $y_1$ and $y_2$.

One readily sees by imposing the symmetry conditions

$$w_\pm^A = w_\pm^B \equiv w, \tag{7.75a}$$

$$\beta_A = \beta_B \equiv \beta, \tag{7.75b}$$

the above system of equations reduces to

$$\frac{x}{dt} = \left(\frac{N_A}{2N} - x\right) W_+(x) - \left(\frac{N_B}{2N} + x\right) W_-(x) \tag{7.76}$$

where

$$\begin{aligned} x &= y_1 - y_2, \\ W_\pm(x) &= w \exp\left[\pm \beta Q x\right]. \end{aligned}$$

In fact, by applying the transformation $x = 2x' - 1$ one retrieves the deterministic law obtained for the preceding model [see Eq. (7.32)], provided that $N_A = N_B$. The symmetries render the two models equivalent, hence the stochastic dynamics of the current model can be captured by a single variable $m = n_A - n_B$. In particular, Eq. (7.71) reduces to

$$\frac{dP(m)}{dt} = (\mathbf{E}^{-1} - 1)(N_A - m) W_+(m) P(m) + (\mathbf{E} - 1)(N_B + m) W_-(m) P(m) \tag{7.77}$$

which is equivalent to Eq. (7.29) when $N_A = N_B$. Therefore results presented in the preceding sections also apply for this model provided that $N_A = N_B$.

## 7.5 Summary and Future directions

In this Chapter, motivated by the complex social behaviour of bacteria we proposed and analysed an artificial gene regulatory network. The main aim of our work was to study,

in a simplified context, how population dynamics – shaped by interactions between individual cells – is affected by fluctuations dominating at the intra-cellular level. The gene network, which we proposed, consists of two mutually repressing quorum sensing modules similar to the one found in the bacterium *V. fischeri*. The reciprocal repression gives rise to two distinct states that a cell can occupy when present in a quorum.

We studied the dynamics that the gene network conveys at the population level using a stochastic differential equation model of the genetic circuit. Our results indicate that the bacterial population can exhibit two different behaviours depending on the strength of the coupling between cells. In the low coupling regime the population appears mixed (disordered) with approximately one half of the population occupying each state. In the high coupling regime cells coordinate their behaviour, with the majority occupying one of the two states; hence the population appears ordered. The crossover between the two regimes depends on the intra-cellular fluctuations. We also used a coarse grained, Ising-type model to study in greater detail the transition between the two regimes of behaviour. In particular, we find a condition that marks the clear transition to the ordered state, linking the coupling strength to the magnitude of fluctuations and size of the population.

The work presented here sets the starting point for a more thorough analysis of our *in-silico* paradigm of bacterial communication that is left for the future. In particular, in our analysis so far we have assumed a mean field, all-to-all coupling between cells. In practise, spatial aspects ought to play an important role. For example, diffusion of signalling molecules, spatial inhomogeneities of the population, and cellular motility, can give rise to pattern and clique formation, phenomena particularly interesting to the physics community [16].

Furthermore, our current design of the gene regulatory circuit is perfectly symmetric with regard to the two quorum sensing modules. Investigation of how different asymmetries introduced in the system affect the population dynamics are also left for the future. This will be particularly relevant since it can shed light on different adaptations bacterial species can exploit to outperform competing species.

Similar artificial gene regulatory networks, enabling interaction between cells and, hence, conveying population wide behaviour (*e.g.,* oscillations, bistability), have been proposed in the literature [49, 91, 140, 148]. The novel ingredient of our gene network is the mutual inhibition between the two distinct quorum sensing channels, giving rise to competitive behaviour, similar to the one observed for *S. aureus* and *B. subtilis*. In this respect our *in-silico* paradigm of bacterial communication can be particularly motivating for synthetic biology efforts on understanding complex bacterial behaviour. Recently, several gene regulatory systems, giving rise to non-trivial population dynamics have been

engineered in living cells [11, 22, 125, 157]. By constructing and analysing such synthetic ecosystems we ought to improve our understanding of naturally occurring systems as well as uncover design principles underpinning how cells interact and coordinate their fate and behaviour.

# Chapter 8

# Discussion

In this Thesis we presented a theoretical study of gene expression at different organisational levels of life. At the microscopic (single-molecule) level the stochastic dynamics of RNA polymerase were considered. In particular, a stochastic model of the transcription elongation phase was proposed and used to study the phenomenon of transcriptional pausing induced via backtracking. Following that, the model was extended with the aim to study transcriptional error correction and the role of backtracking in achieving reduced error rates. Then we aimed to understand how the microscopic dynamics of the process affect RNA levels observed at the cellular level. By constructing an integrated stochastic model model of DNA transcription we studied the effect of transcriptional pausing on the fluctuations of RNA production. Finally, we aimed in understanding how cellular fluctuations of molecular species could affect the dynamics and behaviour of cell populations. To this end, we proposed a simplified system for bacterial communication and studied the effect of intrinsic fluctuations on the ability of cells to coordinate their behaviour.

Special emphasis was placed on the quantitative characterisation of transcriptional pauses caused by backtracking of the RNAP. These pauses dominate *in-vitro* transcription [63] and the existence of specific DNA signals inducing them as well as the presence of accessory proteins assisting their recovery indicate their important role in the regulation of the elongation phase [7]. To understand the phenomenon and its implication in greater detail we presented a stochastic model of the transcription elongation phase, which incorporates polymerisation and backtracking dynamics. Unlike previous modelling attempts [10, 60, 137], our main goal was to to provide a quantitative picture of temporal dynamics of the process. Our results show that owning to the diffusional character of backtracking this class of pauses should obeys a broad temporal distribution, with a power law decay ($t^{-3/2}$). Such finding is consistent with the non-exponential, heavy-tailed distribution of pause lifetimes observed in bacterial and eukaryotic transcription [47, 99, 124].

The phenomenon of RNAP backtracking is also thought to convey transcriptional proofreading [58], however the microscopic details of how error correction is accomplished remain elusive. Motivated by recent experiments [124, 159], we extended our stochastic model of the elongation dynamics to incorporate polymerisation of correct and incorrect nucleotides, and RNA cleavage. Our aim here was to provide a quantitative picture of transcriptional proofreading based on the underlying microscopic dynamics of backtracking. In analogy with kinetic proofreading, in our model backtracking provides a multiple-checking reaction, which probes the fidelity of the last few nucleotides several times before the next polymerization occurs. In fact, the greater the delay introduced by this mechanism, the greater the accuracy of the process [68, 101]. Our model makes specific prediction regarding the observed error rate in terms of the microscopic rates involved in the process. and can be used to assess the overall role of backtracking in enhancing transcriptional fidelity.

At a higher organisation level, one is particularly interested in the role of fluctuations in gene expression and its implications regarding cell behaviour and fate [72, 83]. To this end we aimed to bridge the gap between the microscopic dynamics of DNA transcription and apparent randomness in the production of RNA species by studying a integrated model of DNA transcription. The model involved the initiation, elongation, and termination phases of the DNA transcription and was formulated in terms of totally asymmetric exclusion process to take into account that multiple RNAPs with repulsive interactions can simultaneously transcribe the DNA template. Our results indicate that the interplay between the different time-scales of the model in combination with the exclusive interactions between transcribing TECs can significantly alter the temporal statistics of RNA production. In particular, we found is that rare and long pauses can result in a burst-like production of RNA transcripts and hence super-Poisson RNA statistics. The effect of pauses can be linked heuristically to a switching mechanism between high and low rates of mRNA production. More specifically, sufficiently long pauses shut down RNA production by jamming TEC trafficking on the DNA template. Once the leading TEC resumes elongation multiple blocked TECs that have accumulated at the congestion site are likely terminate transcription resulting in burst of rapid RNA production. Our findings are particularly relevant for *in-vivo* systems demonstrating burst-like RNA production [27, 55, 114].

At an even higher level, that of cell populations, we aimed to understand how cellular fluctuations of gene expression affect population dynamics. Motivated by the complex social behaviour of bacteria we proposed and analysed an artificial gene regulatory network. The gene network consisted of two mutually repressing quorum sensing modules

similar to the one found in the bacterium *V. fischeri*. The reciprocal repression gives rise to two distinct states that a cell can occupy when present in a quorum. Our results indicated that owning to intra-cellular fluctuations the bacterial population can exist in two different states depending on the strength of the coupling between cells. In the low coupling regime the population appears mixed (disordered) with approximately one half of the population occupying each state. In the high coupling regime cells coordinate their behaviour, with the majority occupying one of the two states, hence the population appears ordered. The crossover between the two regimes depends on the intra-cellular fluctuations. We also used a coarse grained, Ising-type model to study in greater detail the transition between the two regimes of behaviour. In particular, we found a condition that marks the clear transition to the ordered state, linking the coupling strength to the magnitude of fluctuations and size of the population. The work presented here sets the starting point for a more thorough analysis of our *in-silico* paradigm of bacterial communication that is left for the future.

Similar artificial gene regulatory networks, enabling interaction between cells and, hence, conveying population wide behaviour (*e.g.*, oscillations, bistability), have been proposed and studied both *in-silico* [49, 91, 140, 148] and *in-vivo* [11, 22, 125, 157]. The novel ingredient of our gene network is the mutual inhibition between the two distinct quorum sensing channels, giving rise to competitive behaviour, similar to the one observed for *S. aureus* and *B. subtilis*. In this respect our *in-silico* paradigm of bacterial communication can be particularly motivating for synthetic biology efforts seeking to understand complex bacterial behaviour. By constructing and analysing such synthetic ecosystems we ought to improve our understanding of naturally occurring systems as well as uncover design principles behind how cells interact and coordinate their fate and behaviour.

# Appendix A

# Transcriptional error correction: $M > 1$ case

Here, we present a detailed treatment of the transcriptional error correction model for the case of $M > 1$. We will restrict our analysis in the limit $\epsilon \ll 1/M$, which allows to safely assume that at most one error can occur every $M$ nucleotides.

## Dynamics at the single nucleotide level

In the general case of $M > 1$, the transition matrix $\mathbf{W}^{(s)}$ will depends on the last $M$ entries of the index $s$. We use the notation $s^*$ to denote all transcripts that have no erroneous nucleotides at the $M$ last places of their sequence, *i.e.*,

$$s^0 = \ldots, \underbrace{0, \ 0, \ \ldots, \ 0}_{M \text{ elements.}}$$

Similarly, we use $s^l$ ($0 \leq l \leq M-1$) to denote all transcripts that have one error in position $n-l$. For example

$$s^1 = \ldots, \underbrace{0, \ 0, \ \ldots, \ 1}_{M \text{ elements.}}$$

Using the transition matrix $\mathbf{W}$ corresponding to each of the sequences $s^*$ $s^l$ ($0 \leq l \leq M-1$) one can obtain from Eq. (5.6) all the splitting probabilities: $p_i(s^l) \equiv \bar{p}_i(l) =$ (the probabilities of hitting boundary $i$ given an error in position $n-l$ of the transcript) and $p_i(s^*) \equiv p_i$ (the probabilities of hitting boundary $i$ given no errors in the last $M$

nucleotides) . In particular, the splitting probabilities corresponding to boundary $i = 0$ (polymerisation) in the limit $K \ll alpha_1 \ll \epsilon$ become

$$p_0 \approx \frac{1}{2M}\frac{K}{\alpha_1}, \tag{A.1a}$$

$$\bar{p}_0(l) \approx \frac{1}{2(M-l)}\frac{K}{\alpha_2}, \quad 0 \le l \le M-1 \tag{A.1b}$$

## Effective model

As in the case for $M = 1$ to calculate the probabilities, $\mathcal{P}_n$, $\bar{\mathcal{P}}_n$, of reaching the the terminal position $n = N$ having transcribed a correct or wrong nucleotide at position $n = n'$ we make use of the effective model of an elongation dynamics. In particular, the splitting probabilities divided by some coarse-grained time-scale $\tau$ yield the effective rates, $r_i$ and $\bar{r}_i$ ($i = 0, 1$), and Eq. (5.11) can be used to describe the dynamics of the system. Similar to the case of $M = 1$ presented in the main text we proceed our analysis by breaking the domain of the process into 3 regions:

- Region $R_-$: $n = 0, \ldots, n' - 1$,
- Region $R_0$: $n = n', \ldots, n' + M - 1$,
- Region $R_+$: $n = n' + M, \ldots, N - 1$.

Let us consider the probability fluxes between these regions. The probability flux from $R_-$ to $R_0$ is due to polymerisation occurring from the boundary position $n = n' - 1$:

$$J(R_-|R_0) = \sum_{s \in S^{n-1}} r_0(s)\Pi(n-1, s, t). \tag{A.2}$$

Polymerisation will result in either a correct or an incorrect nucleotide at position $n'$, This gives rise to two independent branches in the process. The probability flux form $R_0$ to $R_-$ will be through both of this branches

$$J(R_0|R_-) = J^c(R_0|R_-) + J^w(R_0|R_-) \tag{A.3}$$

In particular each term can be decomposed into into $M$ terms, each one corresponding to cleavage from a different position in $R_0$:

$$\begin{aligned} J^c(R_0|R_-) &\equiv \sum_{i=0}^{M-1} J_i^c(R_0|R_-) \\ &\approx \sum_{i=0}^{M-1}\sum_{l=i+1}^{M} r_l \Pi(n'+i, s^*, t), \qquad \text{(A.4a)} \\ J^w(R_0|R_-) &\equiv \sum_{i=0}^{M-1} J_c^i(R_0|R_-) \\ &\approx \sum_{i=0}^{M-1}\sum_{l=i+1}^{M} \bar{r}_l(i) \Pi(n'+i, s^i, t). \qquad \text{(A.4b)} \end{aligned}$$

In the second line of the above equations the fluxes where approximated using the assumption $\epsilon \gg 1/M$. This effectively allows us to neglect misincorporations and consequently any further branching of the process within the region $R_0$. Therefore probability flows between states belonging in $R_0$ as

$$\begin{aligned} J^c(n'+l|n'+m) &\approx r_{m-l}\Pi(n'+l, s^*, t) \text{ for } 0 \le l < m \le M-1, \qquad \text{(A.5a)} \\ J^c(n'+l|n'+l+1) &\approx r_0 \Pi(n'+l, s^*, t) \text{ for } 0 \le l \le M-1, \qquad \text{(A.5b)} \\ J^w(n'+l|n'+m) &\approx \bar{r}_{m-l}(l)\Pi(n'+l, s^l, t) \text{ for } 0 \le l < m \le M-1, \qquad \text{(A.5c)} \\ J^w(n'+l|n'+l+1) &\approx \bar{r}_0(l)\Pi(n'+l, s^l, t) \text{ for } 0 \le l \le M-1. \qquad \text{(A.5d)} \end{aligned}$$

The two branches will evolve independently of one another and will lead to probability flowing into region $R_+$. In particular, probability will flow through polymerisation event occurring at the boundary of the two regions:

$$\begin{aligned} J^c(R_0|R_+) &\approx r_0 \Pi(n'+M-1, s^*, t), \qquad \text{(A.6a)} \\ J^w(R_0|R_+) &\approx r_0(M-1)\Pi(n'+M-1, s^{M-1}, t). \qquad \text{(A.6b)} \end{aligned}$$

Once in region $R_+$ we allow the process to branch once again. However, the total probability entering $R_+$ should be conserved, either flowing back to $R_0$ or to the absorbing

boundary $n = N$. This allows us to write

$$J^c(R_0|R_+) = J^c(R_+|N) + \sum_{i=0}^{M-1} J_i^c(R_+|R_0), \tag{A.7a}$$

$$J^w(R_0|R_+) = J^w(R_+|N) + \sum_{i=0}^{M-1} J_i^w(R_+|R_0), \tag{A.7b}$$

where once again we have decomposed the probability flux into $R_0$ into $M$ independent terms, $J_i^c(R_+|R_0)$, corresponding to the probability fluxes into each position of $n = n' + i$ of $R_0$ respectively.

In the long time limit $t \to \infty$ the fluxes in and out of the different regions will balance and a steady probability flow towards the terminal position $n = N$ will be achieved. In this limit one obtains a set of equations relating the Laplace transform of the aforemen-

tioned probability fluxes

$$
\begin{aligned}
&\sum_{i=0}^{M-1}\Big[\tilde{J}^c(R_0|R_-)+\tilde{J}^w(R_0|R_-)\Big]+\tilde{J}(R_-|R_0)+1 &&= 0\\
&\frac{\epsilon}{\epsilon+1}\tilde{J}(R_-|R_0)-\tilde{J}^c_0(R_0|R_-)+\tilde{J}^c_0(R_+|R_0)+\\
&\sum_{i=1}^{M}\tilde{J}^c(n'+i|n')-\tilde{J}^c(n'|n'+1) &&= 0\\
&\frac{1}{\epsilon+1}\tilde{J}(R_-|R_0)-\tilde{J}^w_0(R_0|R_-)+\tilde{J}^w_0(R_+|R_0)+\\
&\sum_{i=1}^{M}\tilde{J}^w(n'+i|n')-\tilde{J}^w(n'|n'+1) &&= 0\\
&\vdots\\
&-\tilde{J}^c_l(R_0|R_-)+\tilde{J}^c_l(R_+|R_0)+\sum_{i=l+1}^{M}\tilde{J}^c(n'+i|n'+l)-\\
&\sum_{i=1}^{l}\tilde{J}^c(n'+l|n'+l-i)+\tilde{J}^c(n'+l-1|n'+l)-\tilde{J}^c(n'+l|n'+l+1) &&= 0\\
&-\tilde{J}^w_l(R_0|R_-)+\tilde{J}^w_l(R_+|R_0)+\sum_{i=l+1}^{M}\tilde{J}^w(n'+i|n'+l)-\\
&\sum_{i=1}^{l}\tilde{J}^c(n'+l|n'+l-i)+\tilde{J}^w(n'+l-1|n'+l)-\tilde{J}^w(n'+l|n'+l+1) &&= 0\\
&\vdots\\
&-\tilde{J}^c_{M-1}(R_0|R_-)+\tilde{J}^c_{M-1}(R_+|R_0)\\
&-\sum_{i=1}^{M}\tilde{J}^c(n'+M-1|n'+M-1-i)+\tilde{J}^c(n'+M-2|n'+M-1) &&= 0\\
&-\tilde{J}^w_{M-1}(R_0|R_-)+\tilde{J}^w_{M-1}(R_+|R_0)\\
&-\sum_{i=1}^{M}\tilde{J}^w(n'+M-1|n'+M-1-i)+\tilde{J}^w(n'+M-2|n'+M-1) &&= 0\\
&\tilde{J}^c(R_0|R_+)-\tilde{J}^c(R_+|N)-\sum_{i=0}^{M-1}\tilde{J}^c_i(R_+|R_0) &&= 0\\
&\tilde{J}^w(R_0|R_+)-\tilde{J}^w(R_+|N)-\sum_{i=0}^{M-1}\tilde{J}^w_i(R_+|R_0) &&= 0
\end{aligned}
\tag{A.8}
$$

All terms in the above set of equation have the status of probability. Note, for example that in the last line terms $\tilde{J}^w(R_+|N)$ and $\tilde{J}^w_i(R_+|R_0)$ up to division by $\tilde{J}^w(R_0|R_+)$ can be interpreted as splitting probabilities, that is, some probability $\tilde{J}^w(R_0|R_+)$ is injected into $R_+$ and subsequently divided among $M+1$ absorbing boundaries. More importantly, the

division does not depend through which of the two branches the probability ends up in region $R_+$. This consideration allows us to make the following Anstantz

$$\begin{aligned}
\tilde{J}^c(R_+|N) &= A_T^c \tilde{J}^c(R_0|R_+) \\
&\approx A_T r_0 \Pi(n' + M - 1, s^*, t), && \text{(A.9a)} \\
\tilde{J}_i^c(R_+|R_0) &= A_{n'+i}^c \tilde{J}^c(R_0|R_+) \\
&\approx A_{n'+i} r_0 \Pi(n' + M - 1, s^*, t), && \text{(A.9b)} \\
\tilde{J}^w(R_+|N) &= A_T^w \tilde{J}^w(R_0|R_+) \\
&\approx A_T \bar{r}_0 (M-1) \Pi(n' + M - 1, s^{M-1}, t), && \text{(A.9c)} \\
\tilde{J}_i^w(R_+|R_0) &= A_{n'+i}^w \tilde{J}^w(R_0|R_+) \\
&\approx A_{n'+i} \bar{r}_0 (M-1) \Pi(n' + M - 1, s^{M-1}, t), && \text{(A.9d)}
\end{aligned}$$

subject to the condition

$$A_T + \sum_{i=0}^{M} A_{n'+i} = 1, \qquad \text{(A.10)}$$

Substituting in the system of Equations (A.8) the approximations given by Eqs. (A.4)-(A.6) and (A.9) one can solve for all $\Pi(n'+l, s^0, t)$ and $\Pi(n'+l, s^{l+1}, t)$ and subsequently obtain an approximate expression for the probabilities of interest:

$$\mathcal{P}_{n'} = \tilde{J}^c(R_+|N) \approx A_T^c r_0 \Pi(n' + M - 1, s^*, t), \qquad \text{(A.11a)}$$

$$\bar{\mathcal{P}}_{n'} = \tilde{J}^w(R_+|N) \approx A_T^w \bar{r}_0 (M-1) \Pi(n' + M - 1, s^{M-1}, t). \qquad \text{(A.11b)}$$

## Error fraction

In particular, one finds that the the error fraction at position $n'$ is given by

$$\mathcal{E}_{n'} \equiv \frac{\bar{\mathcal{P}}_{n'}}{\mathcal{P}_{n'}} = \epsilon \prod_{i=0}^{M-1} \left( \frac{\bar{p}_0(i)}{p_0} \frac{w_{M-1-i}}{\bar{w}_{M-1-i}} \right) \qquad \text{(A.12)}$$

where $w_n^k$ and $\bar{w}_n^k$ are defined as follows

$$\begin{aligned}\bar{w}_k &= 1 - \bar{p}_0(M-k)\sum_{i=1}^{k-1}\frac{\bar{p}_i(M-i)}{\bar{w}_i}\prod_{j=1}^{i-1}\frac{\bar{p}_0(M-i)}{\bar{w}_i} \\ &\quad -\bar{p}_0(M-k)A_{n'+M-k}\prod_{i=1}^{k-1}\frac{\bar{p}_0(M-i)}{\bar{w}_i} \qquad \text{(A.13a)}\\ w_k &= 1 - p_0\sum_{i=1}^{k-1}\frac{p_i}{w_i}\prod_{j=1}^{i-1}\frac{p_0}{w_i} - p_0 A_{n'+M-k}\prod_{i=1}^{k-1}\frac{p_0}{w_i} \qquad \text{(A.13b)}\end{aligned}$$

with $w_0 = \bar{w}_0 = 0$. Of course $A_{n'+k}$ terms are still unknown, however, they can be calculated by treating the process in the Region $R_+$, with $R_0$ and $n = N$ being absorbing boundaries. One can readily see that for $M = 1$, Eq. (A.12) reduces to the result obtained in Chapter 5 [see Eq. (5.26)] In particular, one has $w_1 = 1 - A_{n'}p_0$ and $\bar{w}_n^1 = 1 - A_{n'}\bar{p}_0(0)$, where $A_{n'}$ corresponds to the spitting probability of exiting region $R_+$ through the boundary at $n = n'$.

Using induction once can show that in the limit $K \ll \alpha_1 \ll \epsilon \ll 1/M$ both $w_i$ and $\bar{w}_i$ approach unity. Therefore, in this limit the error fraction becomes [using Eq. (A.1)]

$$\begin{aligned}\mathcal{E}_n &\approx \epsilon\prod_{i=0}^{M-1}\left(\frac{\bar{p}_0(i)}{p_0}\right) \\ &= \epsilon^{M+1}\frac{M^M}{M!}. \qquad \text{(A.14)}\end{aligned}$$

# Appendix B

# Published Work

334 Biophysical Journal Volume 94 January 2008 334–348

# Fluctuations, Pauses, and Backtracking in DNA Transcription

Margaritis Voliotis,*† Netta Cohen,* Carmen Molina-París,† and Tanniemola B. Liverpool†‡
*School of Computing, †Department of Applied Mathematics, University of Leeds, Leeds, United Kingdom; and ‡Department of Mathematics, University of Bristol, Bristol, United Kingdom

ABSTRACT Transcription is a vital stage in the process of gene expression and a major contributor to fluctuations in gene expression levels for which it is typically modeled as a single-step process with Poisson statistics. However, recent single molecule experiments raise questions about the validity of such a simple single-step picture. We present a molecular multistep model of transcription elongation that demonstrates that transcription times are in general non-Poisson-distributed. In particular, we model transcriptional pauses due to backtracking of the RNA polymerase as a first passage process. By including such pauses, we obtain a broad, heavy-tailed distribution of transcription elongation times, which can be significantly longer than would be otherwise. When transcriptional pauses result in long transcription times, we demonstrate that this naturally leads to bursts of mRNA production and non-Poisson statistics of mRNA levels. These results suggest that transcriptional pauses may be a significant contributor to the variability in transcription rates with direct implications for noise in cellular processes as well as variability between cells.

## INTRODUCTION

It has long been appreciated that noise and fluctuations play an important role in the cellular environment (1). Small numbers of molecules as well as the intrinsically stochastic nature of biochemical reactions mean that fluctuations must be taken into account to understand cellular function. More recently there has been renewed interest in genetic noise (see, e.g., (2–4)) and fluctuations at the molecular level, driven by new observational techniques which allow one to track levels of chemical species in bacterial and yeast cells (5–7). These experiments have allowed the identification of a number of different sources of fluctuations in the expression levels of a particular gene. Low numbers of macromolecules that participate in gene regulation and expression, as well as macroscopic fluctuations in the environment, are likely to affect the statistics of gene expression. In addition, the stochastic nature of the production and degradation of RNA transcription products introduces an important source of intrinsic genetic noise.

Within the central dogma of molecular biology, gene expression can be split into two distinct phases, transcription of DNA to mRNA and translation of mRNA into protein. However, the production (and degradation) of proteins and mRNA transcripts are themselves multistage processes. Transcription, in particular, can be crudely broken up into three main stages: initiation, elongation, and termination. During initiation, RNA polymerase (RNAP) binds to a promoter sequence on the DNA and opens the double helix, uncovering the template strand to be transcribed. The subsequent transcription of the first few (8–12) nucleotides leads to the formation of the transcription elongation complex (TEC) which consists of the RNAP, the DNA, and the nascent mRNA (8). The formation of the TEC signals the entrance into the elongation phase where, under normal conditions, the TEC slides along the DNA, extending the transcript one nucleotide at a time. Destabilization of the TEC (at specific sites or by certain factors) leads to the termination of the process and the release of the nascent mRNA (9).

In fact, the transcription process can exhibit biochemical fluctuations at each stage and cannot, in general, be described by the simple exponential (Poisson) birth and death Markov processes that are currently used to analyze experiments ((4,10) and references therein). This naturally leads one to ask under what conditions is the Poisson approximation valid (11). To answer this question, a more detailed analysis of the dynamics of transcription is required. Recent single molecule experiments (12,13) also provide a new window into the dynamics of transcription, offering a motivation as well as a solid basis for constructing more detailed mathematical models.

As demonstrated below, implicit in the Poisson approximation for the stochastic description of transcription is the assumption that the rate-limiting step is initiation, i.e., that the time taken for the polymerase to find the promoter sequence by random diffusion is longer than the total time for elongation. If so, fluctuations in the initiation step would be the major contributor to genetic noise due to transcription.

In general, the frequency of transcription initiation has a wide dynamical range in vivo (14), and in vitro studies have shown that initiation times can be as fast as a few seconds (15–17). Clearly then, rapid initiation times can be significantly shorter than the time needed for elongation, especially for long DNA templates or bacterial genes transcribed in operons. In such cases, modeling transcription as a Markovian process, obeying Poisson statistics, may be an inadequate approximation. In fact, transcription elongation

*Submitted February 2, 2007, and accepted for publication August 7, 2007.*

Address reprint requests to Tanniemola B. Liverpool, E-mail: t.liverpool@bristol.ac.uk.

Editor: Michael Edidin.



0006-3495/08/01/334/15 $2.00

doi: 10.1529/biophysj.107.105767

demonstrates features that suggest that it could play as significant a role in the overall rate of transcription and hence the regulation of gene expression (18).

Of particular interest are transcriptional pauses that disrupt the processive mRNA synthesis. Single-molecule techniques have made a more quantitative characterization of elongation pauses possible. Recent in vitro experimental studies with *Escherichia coli* RNAP have classified elongation pauses into long (>20 s) and short (1–6 s) pauses (19,20). It has also been suggested that elongation pauses can occur either in a sequence-dependent manner (21) or irrespective of the underlying sequence (19) and some pauses were linked with the reverse translocation of the RNAP (backtracking) (19,22). Backtracking may be caused by nucleotide misincorporation or a weak RNA-DNA hybrid (8,23) and can also be regulated by specific proteins (24). In general, backtracking can significantly increase the total elongation time, and in many cases is the precursor to transcriptional arrest (25).

In this article we point out that a single step Poissonian picture of transcription implies that the rate-limiting step (in transcription) is transcription initiation, i.e., the elongation process that follows is fast and straightforward. We present a molecular model of transcription elongation (26–29) with very different, heavy-tailed distributions of transcription times. Furthermore, we show that elongation can be sufficiently slow to be rate-limiting, providing the cell with ample targets for regulation. In particular, we highlight the very important role transcriptional pauses play in determining the distribution of total transcription times and therefore the statistics of the mRNA levels. Our results should have direct implications for the fluctuations observed in the levels of gene expression, which lead to noise in cellular processes and may play a role in generating variability between cells.

We study two classes of models both analytically, within a mean field approximation, and numerically, using stochastic simulations. First in a model of transcription without transcriptional pauses (Model A), we find that the transcription-elongation adds a typical delay that scales linearly with the transcript size. In this model, the contribution from fluctuations is small (especially for large transcript lengths) and leads to elongation times that are described by a Gaussian distribution. Second, we construct a model that incorporates backtracking pauses during the elongation phase (Model B). We develop a detailed model of backtracking pauses as a first-passage process and study the distribution of their duration considering two different scenarios: 1), pauses that end with the TEC sliding back into position (case 1); and 2), backtracking pauses that can also lead to transcriptional arrest (case 2). In addition, using stochastic simulations, we investigate the effect of backtracking pauses on the distribution of elongation times, as well as on the statistics of the mRNA production. We show that pauses can dominate the elongation process and lead to a heavy-tailed distribution of elongation, and hence transcription completion times. Finally, we use Model B to perform simulations of mRNA production, allowing multiple RNAP molecules to transcribe the same gene. We demonstrate that rare and long-lived pauses result in bursts of mRNA production, in agreement with experimentally observed transcriptional bursting (11,30,31).

## TRANSCRIPTION ELONGATION COMPLEX

At a typical template position the RNAP covers a region of ~25 DNA basepairs (bp), of which the central part (12 bp) is melted, forming the transcription bubble (32). Within the bubble, a hybrid (8–9 bp) is formed between the nascent mRNA and the complementary DNA strand that contributes to the stability of the TEC (33). Elongation (polymerization) describes the addition of a nucleotide to the 3′ end of the transcript, which is catalyzed by the active site of the RNAP and hence conditional on the active site being locked in the appropriate position. In the simplest scenarios, polymerization of the nascent mRNA can be interrupted by the reverse process of pyrophosphorolysis (depolymerization), which leads to shortening of the mRNA transcript (8), or by pauses, due to translocation of the TEC (see below).

After a polymerization step has taken place the TEC is thought to occupy the pretranslocated state. From this position the TEC must translocate forward on the DNA template, to the posttranslocated state, so that the active site is in position to catalyze the next nucleotide addition. In general, the TEC is also capable of translocating backward on the template (backtracking) or even ahead of the target DNA nucleotide (hypertranslocation). During backtracking the TEC is moved upstream along the DNA template. This translocation causes the 3′ end of the nascent mRNA to dissociate from the DNA and exit the TEC through the secondary channel of the polymerase (34). Effectively, this rearward motion dissociates the active site from the 3′ end of the transcript, temporarily halting the elongation, until the TEC is in position once again. The posttranslocated, pretranslocated, and backtracked states are illustrated schematically in Fig. 1, *a–c*.

A simple mathematical model that captures the essence of polymerization, depolymerization, and backtracking can be described in terms of two discrete variables $n$ and $m$. Variable $n$ denotes the position of the last transcribed nucleotide, or equivalently, the size of the nascent mRNA, and ranges from 0 to $N$. In our model, $n$ counts nucleotides relative to the position at which the elongation phase is entered by the formation of the stable TEC. Thus, position $n = 0$ does not correspond to the actual transcriptional starting point, but usually a few (8–10) nucleotides downstream. Finally, transcription will terminate at position $n = N$. Note that $n$ is only affected by polymerization (lengthening) and depolymerization (shortening) of the nascent mRNA. The second variable $m$ denotes the position of the polymerase's active site relative to $n$ and ranges from $-n$ to 1. States $m = 0$ and $m = 1$ are defined as the pre- and posttranslocated states

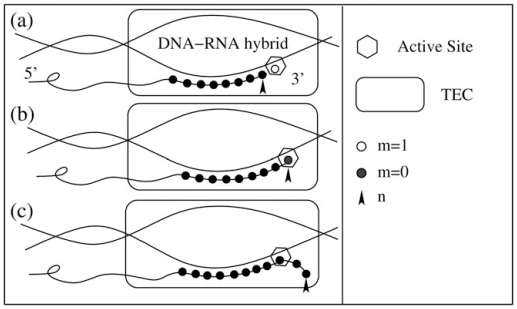
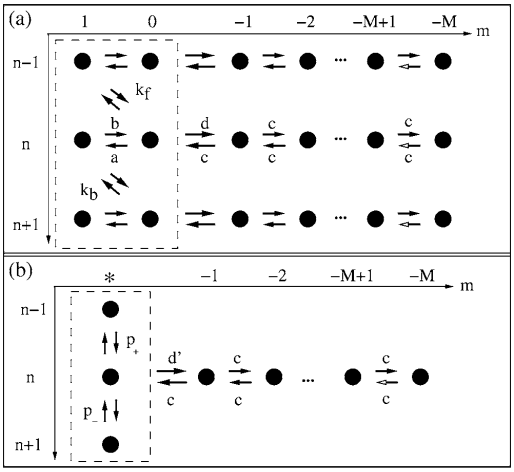



FIGURE 1 Schematic representation of the transcription elongation complex (TEC) at different translocation states: (*a*) Posttranslocated state at ($n$, $m = 1$), (*b*) pretranslocated state ($n$, $m = 0$), and (*c*) backtracked state ($n$, $m = -2$). The position of the TEC on the DNA template is characterized by the position of the active site, which in terms of variables $n$ and $m$ is $x = n + m$.




| Parameter | Value | References |
|---|---|---|
| $b/a$ | 0.8 | (61,62) |
| $k_b/k_f$ | 0.01 | (26) |
| $k_f$ | $36s^{-1}$ | (63) |

FIGURE 2 Schematic illustration of discrete models of transcription elongation. (*a*) Model A (*dotted rectangular*) includes polymerization, depolymerization, and transitions between the post- and pretranslocated states. Model B also allows for backward translocation of the TEC as far as $m = -M$, with $M \ll N$ . If $n < M$, backward translocation is permitted up to state $m = -n$ (not shown). In the case of uninterrupted backtracking (case 1), the TEC can return from state $m = -M$ (*white arrow*), whereas in the case of transcript arrest (case 2), the TEC is halted at $m = -M$ until it is rescued by accessory factors, which move it to state ($n - M$, 0). The table includes typical values for parameters of Model A. (*b*) Schematic illustration of a simplified version of Models A and B when transition between pre- and posttranslocated states is the fastest process. The active states ($m = 0$, 1) have been collapsed into one state, denoted by the asterisk (*). At each template position the TEC can either proceed with polymerization, depolymerization, or enter a backtracked state, with effective rates $p_+$, $p_-$, or $d'$, respectively.

of the TEC, respectively, while $m < 0$ corresponds to a backtracked (or reverse translocated) state. Hypertranslocation (which would lead to $m > 1$) is ignored.

The elongation phase starts with the TEC in state ($n = 0$, $m = 0$). The only transition possible from this state is to the posttranslocated state ($n = 0$, $m = 1$), from which the TEC can revert to ($n = 0$, $m = 0$) or proceed with polymerization. Polymerization, or the addition of a single nucleotide to the nascent mRNA strand, can only proceed from the posttranslocated state. Thus, with the TEC occupying the pretranslocated state ($n$, $m = 0$), polymerization by a single nucleotide requires two steps: 1), the TEC sliding forward to the posttranslocated state ($n$, $m = 1$); and 2), the extension of mRNA by one nucleotide ($n + 1$, $m = 0$), which leaves the TEC in the next pretranslocated state. Conversely, the reverse process of depolymerization can only proceed from the pretranslocated state and leaves the TEC in the previous posttranslocated state ($n - 1$, $m = 1$). Thus, at any given template position $n$, the TEC can freely move back and forth between the pretranslocated ($n$, $m = 0$) and the posttranslocated ($n$, $m = 1$) states, allowing depolymerization and polymerization, respectively, (except from the two boundary points $n = 0$ and $n = N$). A schematic diagram of state transitions for a simplified model excluding backtracking (Model A) is given in Fig. 2 *a*.

Inclusion of backtracking in the model provides an additional pathway, as the TEC can now hop from the pretranslocated state ($n$, $m = 0$) into the first backtracked state ($n$, $m = -1$). Subsequent backward translocation events can randomly shift the TEC's active site back and forth, possibly backtracking as far back as ($n$, $m = -n$) (8). In practice, backtracking is often restricted to $m = -M > -n$. In some cases, backtracking will consist of random reverse and forward translocations that eventually end as the TEC returns to the nucleotide target position (allowing polymerization to resume). In other instances, backtracking is interrupted (in a so-called transcript arrest (8)) and the pause eventually ends when the TEC is rescued by accessory factors, such as the Gre/TFIIS cleavage proteins (35,36). Note that backtracking affects only variable $m$, since it disrupts the positioning of the active site, leaving the length of the nascent mRNA (variable $n$) unaffected. In other words, both polymerization and depolymerization are blocked during backtracking until the corresponding target positions are recovered, i.e., ($n$, 1) and ($n$, 0), respectively. A schematic diagram of state transitions for a model of elongation with restricted backtracking (Model B) is given in Fig. 2 *a*.

For both Models A and B, we seek the statistics of the elongation time, i.e., the time needed for the TEC to reach position ($n = N$, $m = 0$) with the elongation phase starting with the TEC in state ($n = 0$, $m = 0$).

### Model A: translocation-limited polymerization

In this model, backtracked states are ignored, and at each template position $n$ only two translocation states are possible: $m = 1$ and $m = 0$, which allow transcript polymerization and depolymerization, respectively. The rates of polymerization and depolymerization are given by $k_f$ and $k_b$, while $a$ is the translocation rate from $m = 0$ to $m = 1$ and $b$ the reverse rate from $m = 1$ to $m = 0$. (See typical values in the table of Fig. 2.)

The dynamics of $P_{n,m}(t)$, the probability of finding the polymerase in state $(n, m)$ at time $t$, are described by the Master equation (37,38),

$$\frac{\partial P_{n,0}}{\partial t} = k_f P_{n-1,1} + bP_{n,1} - (k_b + a)P_{n,0}, \tag{1a}$$

$$\frac{\partial P_{n,1}}{\partial t} = k_b P_{n+1,0} + aP_{n,0} - (k_f + b)P_{n,1}, \tag{1b}$$

where $n$ varies from 0 to $N - 1$. We assume that depolymerization is impossible at position $n = 0$ and that the process is terminated when position $n = N$ is reached. Consequently, the boundary conditions (BC) imposed on Eq. 1 should be reflecting at $n = 0$ and absorbing at $n = N$. The reflecting BC is obtained by defining a fictitious state $n = -1$ and setting $k_b P_{0,0} = k_f P_{-1,1}$. To obtain the absorbing BC, it is convenient to introduce a fictitious position at $N$ and set $P_{N,0} = 0$ (38), which is equivalent to setting the transition rate from $(N - 1, 1)$ to $(N, 0)$ equal to zero.

A mean-field (quasi-steady-state) approximation yielding a biased random walk is obtained in the limit that the rates of polymerization are much slower than the rates of translocation (i.e., $k_f$, $k_b \ll a,b$) (26,28). The effective polymerization and depolymerization rates are $p_+ \approx k_f a/(a + b)$ and $p_- \approx k_b b/(a + b)$. We calculate $\mu$, the mean elongation time (i.e., the time it takes for the TEC to arrive at $n = N, m = 0$ from a starting position at $n = 0$, $m = 0$) and the variance $\sigma^2$ as a function of the template length $N$ (see Appendix A for a complete derivation). Under normal conditions, elongation is overwhelmingly favored over chain shortening (8) $K = p_-/p_+ \ll 1$. Therefore, we have

$$\mu = \frac{N}{p_+} + K\frac{(N-1)}{p_+} + \mathcal{O}(K^2), \tag{2a}$$

$$\sigma^2 = \frac{N}{p_+^2} + K\frac{(4N-4)}{p_+^2} + \mathcal{O}(K^2). \tag{2b}$$

Fig. 3 shows results obtained from stochastic simulations of Model A (Eq. 1), along with the analytic results obtained

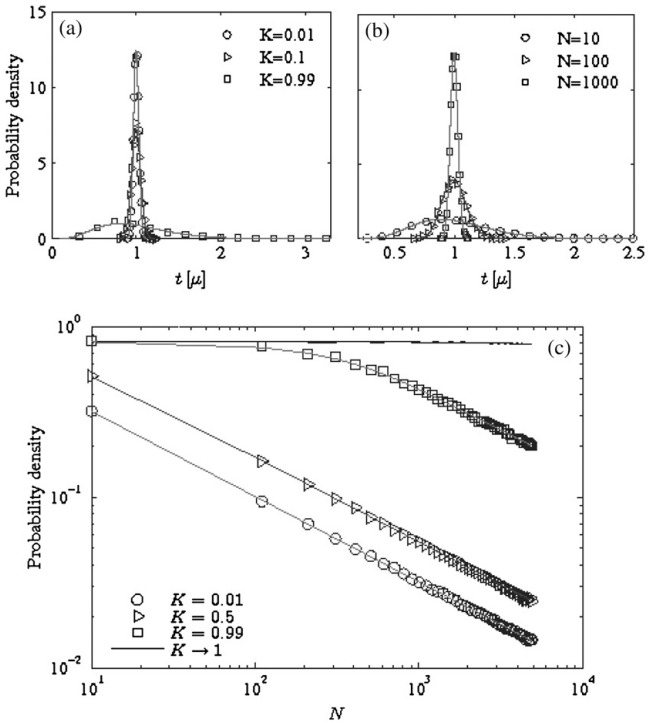



FIGURE 3 ($a$, $b$) Distribution of dimensionless elongation times (scaled by the mean elongation time) for Model A (Eq. 1). Mean-field analytic results are plotted in solid curves, and superimposed with stochastic simulations results. ($a$) Results for $N = 1000$ bp, $p_+ = 20\ s^{-1}$ and different polymerization biases $K = 0.01$, 0.5, 0.99. ($b$) Results for $K = 0.01$, $p_+ = 20\ s^{-1}$ and different template lengths $N = 10$, 100, 1000 bp. ($c$) Standard deviation over mean ($\sigma/\mu$) plotted against the template length $N$ for different values of $K$. As expected, the width of the distribution scales as $1/\sqrt{N}$.

in the mean-field approximation, for different values of $N$ and $K$. In the small $K$ regime and for small values of $N$, the elongation times are approximately $\gamma$-distributed, with shape parameter $\alpha = \mu^2/\sigma^2$ and scale parameter $\beta = \sigma^2/\mu^2$. As $N$ is increased, the distribution approaches a Gaussian, in agreement with the Central Limit Theorem, with mean and variance given by Eqs. 2a and 2b, respectively. Since both $\mu$ and $\sigma^2$ scale linearly with the template length $N$, fluctuations around the mean are of the order $1/\sqrt{N}$. As a result, the distribution becomes narrowly peaked around the mean as $N$ is increased, and in the limit $N \to \infty$, where fluctuations tend to zero, the process becomes essentially deterministic. Conversely, in the $K \to 1$ limit, polymerization and depolymerization tend to play equal roles, leading to fluctuations in the transcription time that do not vanish as $N$ is increased.

### Model B: transcription with backtracking pauses

We now extend Model A to include elongation pauses that arise when the TEC occupies backtracked states ($m < 0$). In particular, a pause is signaled when the TEC enters the backtracked state $m = -1$ from state $m = 0$. We denote the corresponding transition rate by $d$ and assume a slow rate relative to polymerization $d \ll p_+$. From $m = -1$ the TEC hops across contiguous backtracked states with rate $c$. In principle, at each template position $n$, backtracking can proceed up to $m = -n$ (8). However, in practice, different mechanisms, such as RNA hairpins, RNA-DNA interactions, and cleavage enzymes preclude extensive backtracking (33). A more reasonable assumption is that backtracking is restricted in length; we assume backtracking to be restricted to a fixed number of steps $m = -M \gg -n$, which we take to be independent of position $n$. Also, for values of template position $n$ that are smaller than $M$, backtracking is permitted to extend as far as $m = -n$. In fact, hairpins are dynamic (breaking and reforming), implying that the choice of fixed $M$ is only a first approximation. If the hairpin relaxation time is sufficiently fast (as compared with the backtracking rate), such dynamics could lead to fluctuations in the value of $M$.

### Dynamics of backtracking pauses

To gain insight into the statistics of transcriptional pauses, we describe and examine the dynamics of backtracking as a separate process. Without loss of generality, we describe backtracking by a symmetric hopping process, or unbiased random walk with rate $c$. The asymmetric case, equivalent to a biased random walk, is quite a straightforward generalization (39). For simplicity, we characterize backtracked states by a new variable $l = -m$, where $1 \le l \le M$. The probability $P(l, t)$, of finding the polymerase in state $l$ at time $t$ given that it was in state $l = 1$ at $t = 0$, follows the Master equation:

$$\frac{\partial P(l,t)}{\partial t} = cP(l-1,t) + cP(l+1,t) - 2cP(l,t). \quad (3)$$

We use the Laplace transform, $\tilde{p}(l,s) = \int_0^\infty P(l,t)e^{-st}dt$, to obtain exact expressions for the probability distribution of the duration of backtracking pauses for two different scenarios:

1. Uninterrupted backtracking: $l = M$ is a reflecting boundary, and termination of the pause occurs when the TEC eventually slides back to state $l = 0$ and
2. Transcript arrest: The TEC is irreversibly halted at $l = M$. Elongation can be resumed either from state $l = 0$ or from position $l = M$ with the aid of accessory factors. Detailed derivations are given in Appendix B.

### Case 1: uninterrupted backtracking

In this case no backward translocation is possible beyond state $l = M$, and the pause is ended when state $l = 0$ is reached. The corresponding boundary conditions for Eq. 3 are: $P(0, t) = 0$ (absorbing) and $cP(M, t) = cP(M + 1, t)$ (reflecting). The mean pause duration is $\langle t \rangle = M/c$ and an analytic expression for the probability distribution $\mathcal{P}(t)$ of pause duration is given in Appendix B. Simple expressions for $\mathcal{P}(t)$ are obtained in the following limits:

$$\mathcal{P}(t) \approx \begin{cases} \dfrac{1}{2\sqrt{\pi}\sqrt{ct^3}^{3/2}}, & \dfrac{1}{c} \ll t \ll \dfrac{M^2}{c}, \\ \dfrac{\pi c \sin\left(\dfrac{\pi}{2(M+1)}\right)}{(1+M)^2} \exp\left[e^{-\dfrac{c\pi^2}{4(1+M)^2}t}\right], & t \gg \dfrac{M^2}{c}. \end{cases} \quad (4)$$

For times short compared to the timescale of diffusion to the reflecting state $l = M$ ($t \ll M^2/c$), but still longer than the time for the TEC to diffuse by one nucleotide ($t \gg 1/c$), $\mathcal{P}(t)$ scales as $t^{-3/2}$. Interestingly, the power law behavior characteristic of this regime is consistent with the heavily skewed and heavy-tailed distribution observed by Shaevitz et al. (19). Conversely, for times much longer than $M^2/c$, which ensure reflection, the asymptotics are altered and $\mathcal{P}(t)$ exhibits a rapid exponential decay. The two different asymptotic behaviors are illustrated in Fig. 4 *a*, where the analytic results have been plotted together with the data obtained from stochastic simulations of the model.

### Case 2: backtracking with transcript arrest

As before, pauses begin with a transition into state $l = 1$ and terminate when state $l = 0$ is reached. However, in this scenario, backtracking will also be terminated by the arrest of transcription if the TEC arrives at $l = M$. Transcription can only resume from the arrested state with the aid of a rescue mechanism (35,36). The boundary conditions imposed to Eq. 3 are therefore absorbing at both ends: $P(0, t) = P(M, t) = 0$.

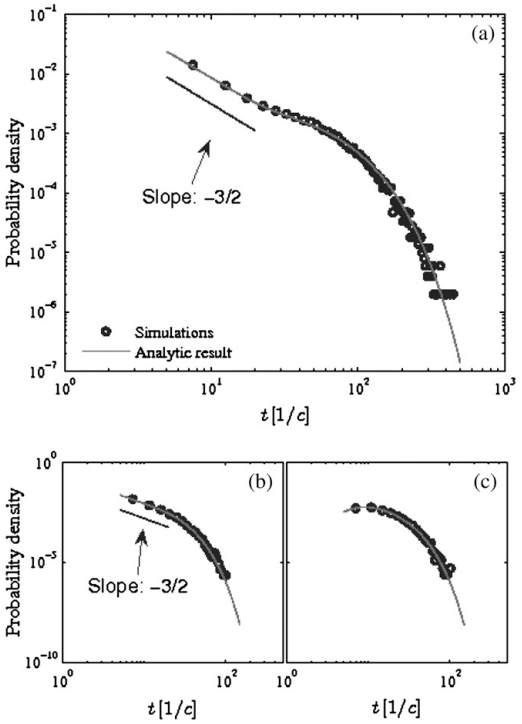



FIGURE 4 Results for case 1 (uninterrupted backtracking) and case 2 (transcript arrest) pauses with $M = 10$. Distributions of (*a*) pause duration $\mathcal{P}(t)$ for case-1; (*b*) self-recovered pause duration $\mathcal{P}_0(t)$ for case-2; and (*c*) time to arrest $\mathcal{P}_M(t)$ for case-2. Plotted are the analytic results (Eq. 39, and Eqs. 45a and 45b, respectively) as solid lines and the results of stochastic simulations as circles. $\mathcal{P}(t)$ and $\mathcal{P}_0(t)$ exhibit a power law decay for $1/c \ll t \ll M^2/c$, followed by an exponential cutoff in long time limit ($t \gg M^2/c$).

It can be shown (see Appendix B) that the probability of eventual arrest of the TEC is $p_M = 1/M$; the probability of TEC recovery from the pause is $p_0 = 1-p_M$; and the corresponding mean time for each case is $\langle t\rangle_M = (M^2-1)/6c$ and $\langle t\rangle_0 = (2M-1)/6c$. Compact expressions for $\mathcal{P}_0(t)$, the probability distribution of recovering from the pause at time $t$, are obtained in the two limits discussed above:

$$\mathcal{P}_0(t) \approx \begin{cases} \dfrac{1}{2\sqrt{\pi}\sqrt{c}t^{3/2}}, & \dfrac{1}{c} \ll t \ll \dfrac{M^2}{c}, \\ \dfrac{2\pi c \sin\left(\dfrac{\pi}{M}\right)}{M^2} \exp\left[e^{-\dfrac{\pi^2 c}{M^2}t}\right], & t \gg \dfrac{M^2}{c}. \end{cases} \quad (5)$$

Once again, the distribution demonstrates a power law decay for $1/c \ll t \ll M^2/c$, followed by an exponential cutoff. For sufficiently long times $t \gg M^2/c$ that allow diffusion to the boundary $l = M$, the probability distribution of the TEC arrest decays exponentially with $\mathcal{P}_M(t) \approx \mathcal{P}_0(t)$. The above analytic results, along with stochastic simulations, are summarized in Fig. 4.

### Stochastic simulations of Model B

Having characterized backtracking statistics, we are now in a position to examine the effects of backtracking on the total elongation time. The macroscopic (observable) properties that we must consider are: 1), the number of pauses $\delta$ over a DNA template of length $N$, and 2), the aggregate lifetime of all the pauses relative to the time spent on active polymerization. These properties are linked to the microscopic parameters $d$ and $c$, respectively. In particular, when translocation between pre- and posttranslocated states is the fastest process, the number of pauses $\delta$ is given by:

$$\frac{\delta}{N} = \frac{d\dfrac{a}{a+b}}{d\dfrac{a}{a+b} + p_+ + p_-} = \frac{d'}{d' + p_+ + p_-}, \quad (6)$$

where $d' = d(a/(a+b))$ is the effective rate of entering into a backtracked state (see Fig. 2 *b*). Moreover, the distribution of pause durations (for the case of uninterrupted backtracking) is determined by the symmetric diffusion rate $c$, with $M/c$ being the mean pause duration.

As expected, in the limit of short-lived pauses, even the aggregate pause duration will be negligible relative to the time spent on processive polymerization, $N/p_+ \gg \delta(M/c)$, and so the distribution of elongation times will approach that of Model A. Conversely, when $N/p_+ \gg \delta(M/c)$, pauses dominate the total elongation time and the distribution of elongation times is significantly affected by the large fluctuations in the duration of the pauses. In the limit $p_+ \gg d'$ and $p_+ \gg p_-$, Eq. 6 becomes $\delta/N \approx d'/p_+$ and the above limits can be written as $d'(M/c) \ll 1$ and $d'(M/c) \gg 1$. We therefore introduce $R = d'(M/c)$ as a dimensionless measure of pauses which quantifies their relative contribution to the elongation time. This measure of pause durations is particularly useful as it is directly linked to the macroscopic parameters of the system (i.e., mRNA production rate) but is derived from the microscopic rate parameters.

Figs. 5 and 6 illustrate the results of the stochastic simulations of Model B, i.e., transcription with restricted, uninterrupted backtracking, for different values of $R$ (keeping the frequency of pauses $\delta/N$ constant). As expected, for $R \to 0$ the polymerization-only model (Model A) is recovered and $\sigma/\mu = 1/\sqrt{N}$ (Fig. 6). This is also evident from the distribution of elongation times, where for small $R$ the high peak close to the mean elongation time predicted by Model A indicates that either no pauses or only brief ones occur. The effect of backtracking events is most evident in the heavier tail of the distribution since rare prolonged pauses can give rise to significantly longer elongation times. This effect is magnified as the fraction of time spent in pauses is increased (i.e., for higher values of $R$) (Fig. 5 *a*). For increasing pause frequency

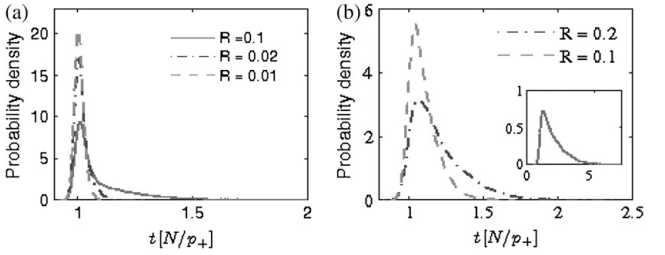



FIGURE 5 Distributions of dimensionless elongation times (scaled by $N/p_+$) for Model B for different values of $R = d'M/c$. The distributions were obtained from stochastic simulations. (*a*) $N = 4$ kb, $M = 10$ bp, $p_+ = 10\ s^{-1}$, $K = 0.01$ and $d'$ chosen to yield $\delta/N \approx d'/p_+ = 1$ pauses/kb (19,22). (*b*) $N = 1$ kb, $M = 10$ bp, $p_+ = 10\ s^{-1}$, $K = 0.01$, and $d'$ chosen to yield $\delta/N \approx d'/p_+ = 10$ pauses/kb. (*Inset*) $R = 1$. The effect of the pauses is evident in the heavy tails that broaden with decreasing $R$ or increasing $\delta/N$.

(higher $\delta/N$) the effect on the total elongation time is clearly more profound; the distribution becomes broader and exhibits a general shift toward longer elongation times (Fig. 5 *b*).

### mRNA transcript levels: production and degradation

Models A and B capture the statistics of the elongation phase. Ultimately, however, one is interested in the mRNA levels, which are the combined action of mRNA production (transcription) and degradation. In general, the transcription process involves an initiation phase (which includes promoter binding, open complex formation, and promoter clearance), an elongation phase, and termination. As a more complete model of transcription, we assume fast termination and combine the model of elongation presented above (Model B) with a simplified, first-order initiation step. Degradation is also represented as a Poisson, single-step process. Using

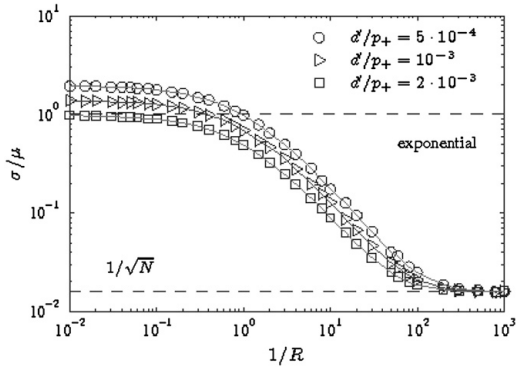



FIGURE 6 Standard deviation over mean ($\sigma/\mu$) of elongation times (Model B) plotted against $1/R$ for different values of the ratio $d'/p_+$ (pause frequency). As $1/R \to 0$, pauses become more significant and the distribution of elongation times becomes broader. In the case of frequent pausing ($d'/p_+ = 2 \times 10^{-3}$), the distribution exhibits characteristics of an exponential distribution, i.e., $\sigma/\mu = 1$ (indicated by the *upper dashed line*). As $1/R \to \infty$, the effect of pauses vanishes and Model B approaches Model A, where $\sigma/\mu \approx 1/\sqrt{N}$ (indicated by the *lower dashed line*). Parameters used: $N = 4$ kb, $M = 10$ bp, $d' = 0.01\ s^{-1}$, $K = 0.01$, and $p_+ = 2$, 10, and 20 $s^{-1}$.

stochastic simulations of this combined transcription-degradation model, we examine how the elongation and possible pauses therein affect steady-state mRNA levels.

We denote the initiation rate as $k_i$. The elongation phase proceeds as described by Model B and instantaneous termination takes place when the transcript reaches its designated size, leading to mRNA production. Finally, mRNA degradation is modeled as a first-order process with rate constant $k_d$. The combination of mRNA production and degradation gives a first handle on mRNA levels and fluctuations in the cell.

In fact, mRNA production is complicated by the fact that multiple initiation events can occur within the time it takes to produce a single mRNA. This would lead to several TECs moving in tandem on the same DNA template (40), each synthesizing a nascent mRNA. To capture the fact that two TECs cannot come in close proximity due to nonspecific interactions between them or to the additional work required to deform the DNA helix (41,42), we set a minimum (exclusion) distance of $L$ nucleotides ($L \ll N$) between the active sites of any two contiguous TECs. In terms of variables $n$ and $m$ of Model B, the active site of a TEC is located at position $x = n + m$ along the DNA template. Therefore, a TEC, positioned at $x_1$, can translocate forward (backward) if the leading (trailing) TEC, positioned at $x_2$, is at distance of at least $L$ nucleotides, i.e., $|x_1 - x_2| < L$. A similar argument applies for transcription initiation, that is, no RNA polymerase can initiate transcription if a TEC is at position $x \leq L$. A schematic illustration of the model is given in Fig. 7.

The relevant timescales associated with the above model are: 1), the time needed for transcription initiation $\tau_1 = 1/k_i$; 2), the time needed by the TEC to transcribe $L$ nucleotides $\tau_2 \approx L/p_+$; and 3), the mean time of a pause due to backtracking $\tau_3 = M/c$. When initiation is the rate-limiting step ($\tau_1 \gg \tau_2, \tau_3$), the density of TECs on the DNA template is low and therefore transcriptional pauses and interactions between TECs are expected to have marginal effects. Consequently, the rate of mRNA production is set mainly by the rate of initiation $k_i$ and the statistics of the mRNA levels are expected to be approximately Poisson with the mean equal to the variance ($\mu_{mRNA} = \sigma^2_{mRNA}$; see Fig. 8 *III*). If the rate of polymerization is the rate-limiting step

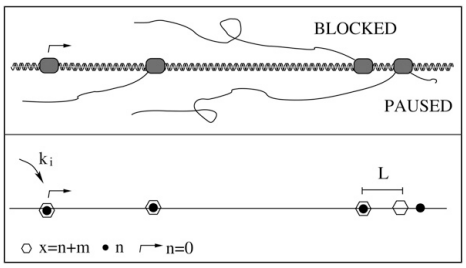



FIGURE 7 Schematic illustration of multiple RNAP molecules simultaneously transcribing a DNA template. Transcription initiation proceeds with an effective rate of $k_i$. The position of each TEC on the DNA is characterized by the position of its active site, which is given by $x = n + m$. We also set a minimum (exclusion) distance of $L$ nucleotides between any two TECs. If transcriptional pauses are sufficiently long they can block the progress of trailing RNAP molecules and subsequently lead to a burst in mRNA production. Such a scenario suggests a significant link between transcriptional pauses and mRNA production statistics.

($\tau_2 \gg \tau_1,\tau_3$), fast transcription initiation is blocked by the slow movement of the TECs on the DNA template, while the relatively short-lived backtracking events, as in the case above, play no significant role. In particular, the density of TECs along the DNA is expected to be maximal ($N/L$), with the TECs kept evenly spaced ($L$ nucleotides apart) by exclusive interactions. In this regime the statistics of the mRNA levels are sub-Poisson with more evenly distributed TECs along the DNA template ($\mu_{\mathrm{mRNA}} > \sigma^2_{\mathrm{mRNA}}$; see Fig. 8 *II*). Finally, $\tau_3 \gg \tau_1,\tau_2$ corresponds to a regime where long pauses dominate transcription. Such pauses can create congestion points by blocking the movement of trailing TECs, while the leading TECs continue to transcribe normally. In this way the uniform ($\tau_2 \gg \tau_1$) or Poisson ($\tau_1 \gg \tau_2$) distribution of TECs on the DNA template is disrupted, resulting in a burstlike production of mRNA transcripts (Fig. 9) and super-Poisson mRNA statistics (i.e., $\mu_{\mathrm{mRNA}} < \sigma^2_{\mathrm{mRNA}}$; see Fig. 8 *I*).

In the bursting regime, the effect of elongation pauses can be linked heuristically to a switching mechanism between high and low rates of mRNA production. In particular, sufficiently long pauses shut down mRNA production by blocking trailing TECs. In the intervals between pauses, multiple blocked TECs that have accumulated at a congestion site are likely to be transcribed in a burst of rapid mRNA production. A qualitative description of the different classes of behavior obtained for the integrated initiation, elongation, degradation model is presented in Table 1. Stochastic simulations of the model confirm that rare and long-lived pauses give rise to jamming of TEC trafficking during transcription and therefore bursts of mRNA production. We note that such abrupt switching between two states is reminiscent of dynamic phenomena observed in studies of the asymmetric exclusion process (43,44).

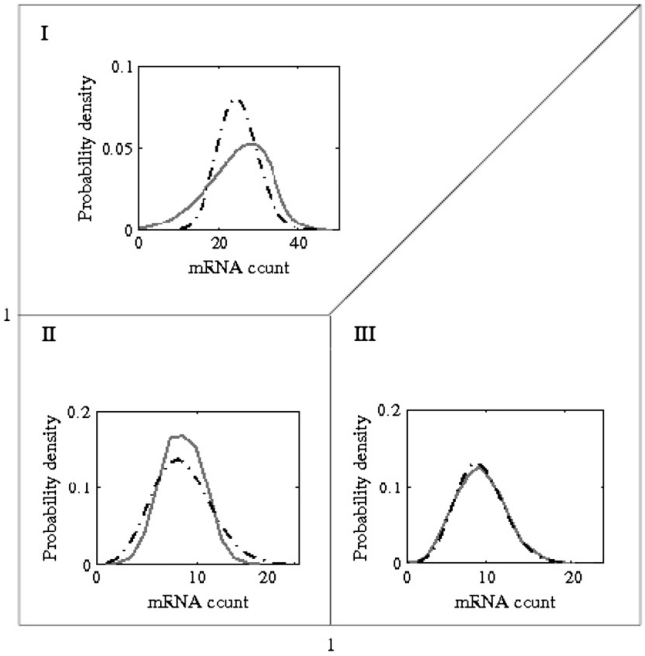



FIGURE 8 Distribution of steady-state number of mRNA molecules (*solid line*). Simulations included transcription initiation, elongation, and mRNA degradation and allowed multiple RNAP molecules to transcribe the DNA template at the same time. A Poisson distribution with the same mean value is given for reference (*dash-dotted line*). (*I*) When elongation pauses are longer than the time needed for transcription initiation and the time needed by the TEC to transcribe $L$ nucleotides ($\tau_3 \gg \tau_1,\tau_2$), the mRNA distribution is expected to be broader than Poisson. (*II*) When the movement of RNAP molecules on the DNA template is the rate-limiting step ($\tau_2 \gg \tau_1,\tau_2$), the mRNA distribution predicted by the model is sub-Poisson. (*III*) When transcription initiation is the rate-limiting step ($\tau_1 \gg \tau_2,\tau_3$), the mRNA distribution predicted by the model is Poisson.

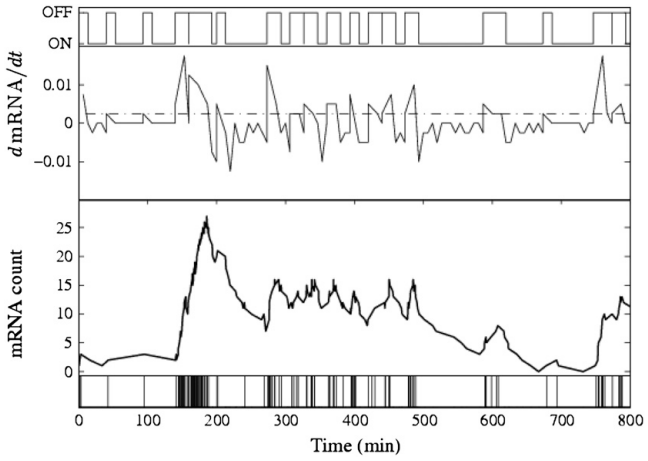



FIGURE 9 Simulation of mRNA population levels in an integrated model of transcription initiation, elongation, and mRNA degradation (parameters given in Appendix C; $10^3$ runs). The inclusion of transcriptional pausing (when multiple initiations are permitted) results in bursts of mRNA production and super-Poisson mRNA statistics ($\sigma^2_{\mathrm{mRNA}}/\mu_{\mathrm{mRNA}} = 4.25$). The bottom panel shows the mRNA production events in time and the trace above illustrates the resulting mRNA count fluctuations. In the third panel, $d$mRNA/$dt$ is plotted ($dt = 6$ min), along with an arbitrary threshold (*dotted line*, set to 1/$dt$ mRNA/s). The threshold enables us to visualize the transcriptional process as a telegraph process with off- and on-states corresponding to low and high rates of mRNA production (*top panel*).

## DISCUSSION

We have presented a stochastic model of transcription, including initiation, elongation, and mRNA degradation. Our main focus has been on the elongation phase for which we obtained analytic results both for the polymerization dynamics (ignoring backtracking) and for the dynamics of backtracking pauses. Our model of backtracking pauses as a first passage process is consistent with recent single molecule experiments (19). By means of stochastic simulations we have also examined how pauses affect the total elongation times. Finally, we have developed a model of mRNA production and degradation that combines transcription initiation, transcription elongation, and mRNA degradation. In this model, multiple RNAPs with repulsive interactions can move in tandem on the same DNA template. We used stochastic simulations of this model to examine how the dynamics of the elongation phase and backtracking pauses therein affect the statistics of the mRNA population levels.

Our key results are particularly instructive in two limits: first, when pauses cause a weak perturbation to elongation dynamics and secondly, when they significantly affect it. The third regime, in which initiation is the rate-limiting step (with relatively rapid elongation), recovers previously predicted Poisson statistics. As expected, if the elongation phase dominates transcription, but the time spent in backtracking pauses is brief relative to that spent on active polymerization, similar results to the polymerization-only model are recovered. That is, for sufficiently long sequences ($N \gg 1$) the elongation times follow a narrow Gaussian distribution with fluctuations around the mean scaling like $1/\sqrt{N}$, where $N$ is the length of the gene. This leads to a characteristic delay in the total time of transcription. Coupling fast transcriptional initiation with such a model of transcription elongation predicts a more homogeneous transcription process and hence steadier mRNA population levels than would be produced by a model of initiation alone.

**TABLE 1 Table summarizing the behavior of mRNA production in the different limiting regimes (with time-limiting initiation, polymerization, or pausing kinetics)**

| | Regime | Behavior |
|---|---|---|
| $\tau_1 \gg \tau_2,\tau_3$ | $\tau_1 \gg \tau_2,\tau_3$ | Poisson |
| | $\tau_1 \gg \tau_2,\tau_3$ | Poisson |
| $\tau_2 \gg \tau_1,\tau_3$ | $\tau_1 \gg \tau_3$ | sub-Poisson |
| | $\tau_3 \gg \tau_1$ | sub-Poisson |
| $\tau_3 \gg \tau_1,\tau_2$ | $\tau_1 \gg \tau_2$ | super-Poisson |
| | $\tau_1 \gg \tau_2$ | super-Poisson |
| $\tau_1 - \tau_2 \gg \tau_3$ | | sub-Poisson |
| $\tau_1 - \tau_3 \gg \tau_2$ | | super-Poisson |
| $\tau_2 - \tau_3 \gg \tau_1$ | | super-Poisson |

In the opposite regime, when there is a significant number of backtracking pauses whose duration is comparable to the active polymerization time, there is a dramatic change in the distribution of transcriptional times. We considered two types of backtracking pauses; pauses that end with the TEC sliding back into position and backtracking pauses that can lead to transcriptional arrest. For both classes of pauses we found a broad distribution of pause durations with a power law decay cutoff by an exponential one. Consequently, the statistics of the elongation phase can be dramatically altered, with increased mean and a significantly broader distribution of elongation times, which mirrors the distribution of pause durations.

Recent experiments have provided evidence for the existence of bursts of transcription both in bacterial (11) and eukaryotic cells (30,31). We have found that our model of the dynamics of elongation with pauses leads naturally to switching between high and low mRNA production rates, resulting in transcriptional bursts. Our findings suggest that rare and long elongation pauses (from the tails of the distribution) act as congestion points turning off mRNA

production for long periods, while allowing rapid mRNA production for short intervals. Such long pauses, therefore, give rise to more strongly fluctuating mRNA levels. Thus, in this regime, elongation pauses act as a rate-limiting step.

In fact, experimental reports of transcriptional bursting measure mRNA population levels (rather than production rates). We obtain consistent fluctuations in mRNA population levels, in a model that combines transcription with mRNA degradation kinetics. Other possible elongation pauses (which are not linked to backtracking) could result in similar bursting effects (45). Indeed, pauses can, in general, result from sequence-encoded signals (46), elongation factors, or nucleosome packaging (47,48). We note, however, that the rate-limiting step can also be provided by a number of different mechanisms associated with the transcription process, such as changes in the state of the promoter (30,31) (e.g., by chromatin remodeling) or the diffusive motion of regulatory molecules (49).

While single molecule studies have provided evidence that RNAP backtracking dominates in vitro transcription and results in pauses of significant (>20 s) duration (19), it is interesting to consider how frequent they are and what role they may play in vivo. For example, backtracking pauses have been previously implicated in mRNA editing and error correction (8,23) and could therefore partially account for discrepancies between theoretically expected and observed error rates in mRNA transcripts. Differences in free energies between correct and incorrect nucleotides yield an expected error rate of $10^{-3}$ errors/bp. This high rate contrasts with experimentally measured values of $10^{-5}$ errors/bp (50). This discrepancy in error rates could presumably be accounted for by error correction mechanisms, which may include backtracking pauses (M. Voliotis, N. Cohen, C. Molina-París, and T. B. Liverpool, unpublished). Of course, the situation in vivo is further complicated by the effects of transcription factors and other regulatory proteins. Nevertheless, if backtracking pauses are significant in the elongation process they could provide the cell with ample opportunity for a range of regulation mechanisms.

The models presented here relied on a number of simplifying assumptions. In particular, both polymerization and elongation pauses were taken to be sequence-independent. The assumption that polymerization takes place on a homogeneous DNA template is likely to be a simplification, since the local rates of translocation have been suggested to depend on the underlying local DNA sequence. Moreover, our models have neglected any sequence dependence that has been attributed to short-lived pauses (20,21). We leave the development of more detailed sequence-dependent kinetic models of elongation dynamics for future research.

While in this article we restrict our calculations to models of transcription, similar arguments regarding pauses and bursting should also be relevant for translation. Applications of these results will ultimately contribute to a more complete understanding of gene expression and regulation, and fluctuations therein. A better understanding of these processes will also shed light on the differences between the effects of gene regulatory mechanisms, which act during transcription and translation (18,52–56) as compared to those which act by controlling the initiation of these processes. Ultimately, models of noise generation in the cellular environment may lead to new insights on the ways in which cells survive and adapt, with consequences for cell development, function, and fate.

## APPENDIX A: TRANSLOCATION-LIMITED POLYMERIZATION

For Model A, the Master equation describing the dynamics of $P_{n,m}(t)$, the probability of finding the TEC in state $(n, m)$ at time $t$, starting from an initial state (0, 0) at $t = 0$, is given by Eq. 1. Since we take $N$ to be the termination site, we implement an absorbing boundary at position $(n = N, m = 0)$. Such a boundary can in general be obtained by setting the depolymerization rate at $n = N$ equal to 0. By doing so, Eq. 1b is affected only for $(n = N - 1, m = 1)$:

$$\frac{\partial P_{N-1,1}}{\partial t} = aP_{N-1,0} - (k_f + b)P_{N-1,1}. \quad (7)$$

The same result can be obtained by setting $P_{N,0} = 0$ and regarding Eq. 1b valid for every $n$ in $\{0, 1, \ldots, N - 1\}$. Also, since we assume $(n = 0, m = 0)$ to be a reflecting boundary, we set the depolymerization rate at $n = 0$ to 0 and $P_{-1,1} = 0$, i.e., there is no probability flow from or to state $(n = -1, m = 1)$. In this way, Eq. 1a is affected only for $(n = 0, m = 0)$:

$$\frac{\partial P_{0,0}}{\partial t} = bP_{0,1} - aP_{0,0}. \quad (8)$$

The same result can be obtained by setting $k_b P_{0,0} = k_f P_{-1,1}$ such that Eq. 1a is valid for every $n$ in $\{0, 1, \ldots, N - 1\}$.

We can define a mean occupancy for each translocation state $(m = 0, 1)$ by summing over all possible template positions, $\Pi_m(t) = \sum_{n=0}^{N-1} P_{n,m}(t)$. From Eq. 1a, we obtain

$$\frac{\partial \Pi_0}{\partial t} = (k_f + b)\Pi_1 - (k_b + a)\Pi_0, \quad \text{and} \quad \Pi_1 = 1 - \Pi_0. \quad (9)$$

The solution to Eq. 9 that satisfies initial conditions $\Pi_0(0) = 1$ relaxes on a timescale $\tau = (a+b+k_f+k_b)^{-1} \ll k_f^{-1}$. On timescales longer than $\tau$, this solution attains steady-state values such that $\Pi_0^s = (k_f + b)\tau$ and $\Pi_1^s = (k_b + a)\tau$. For such long times the fluctuations in $n$ and $m$ become independent and we can write $P_{m,n} = \Pi_m^s P_n$. Substituting back into Eq. 1 and summing over $m$, we obtain

$$\frac{\partial \mathcal{P}_n}{\partial t} = p_- \mathcal{P}_{n+1} + p_+ \mathcal{P}_{n-1} - (p_- + p_+)\mathcal{P}_n, \quad (10)$$

which is equivalent to a biased random walk with effective polymerization and depolymerization rates

$$p_+ = k_f(k_b + a)\tau \approx \frac{k_f a}{a + b}, \quad (11a)$$

$$p_- = k_b(k_f + b)\tau \approx \frac{k_b b}{a + b}, \quad (11b)$$

where we have used $k_f, k_b \ll a,b$. Note that the boundary conditions for Eq. 10 are $\mathcal{P}_N = 0$ (absorbing) and $p_-\mathcal{P}_0 = p_+\mathcal{P}_{-1}$ (reflecting).

The elongation time is defined as the time needed for the TEC to reach position $(n = N, m = 0)$ starting from $(n = 0, m = 0)$. In the mean-field model the mean and variance of the elongation time can be calculated using the backward Master equation (38). We denote the initial template position

of the TEC at time $t_0 = 0$ by $n_0$ and rewrite Eq. 10 in terms of conditional probabilities:

$$\frac{\partial \mathcal{P}(n,t|n_0,t_0)}{\partial t} = p_+\mathcal{P}(n-1,t|n_0,t_0) + p_-\mathcal{P}(n+1,t|n_0,t_0) - (p_+ + p_-)\mathcal{P}(n,t|n_0,t_0). \tag{12}$$

The backward Master equation is (38)

$$\frac{\partial \mathcal{P}(n,t|n_0,t_0)}{\partial t_0} = p_+[\mathcal{P}(n,t|n_0,t_0) - \mathcal{P}(n,t|n_0+1,t_0)] + p_-[\mathcal{P}(n,t|n_0,t_0) - \mathcal{P}(n,t|n_0-1,t_0)]. \tag{13}$$

Since the system is homogeneous, we can write

$$\mathcal{P}(n,t|n_0,t_0=0) = \mathcal{P}(n,0|n_0,-t), \tag{14}$$

so that the backward Master equation takes the form

$$\frac{\partial \mathcal{P}(n,t|n_0,0)}{\partial t} = p_+[\mathcal{P}(n,t|n_0+1,0) - \mathcal{P}(n,t|n_0,0)] + p_-[\mathcal{P}(n,t|n_0-1,0) - \mathcal{P}(n,t|n_0,0)]. \tag{15}$$

The boundary conditions for the backward Master equation are $\mathcal{P}(n,t|n_0 = 0,0) = \mathcal{P}(n,t|n_0 = -1,0)$ (reflecting) and $\mathrm{P}(n,t|n_0 = N,0) = 0$ (absorbing).

The probability that at time $t$ the TEC has not yet reached the absorbing boundary is given by

$$\sum_{n=0}^{N-1}\mathcal{P}(n,t|n_0,0) = G(n_0,t). \tag{16}$$

If $T$ is the elongation time (time needed to complete elongation by reaching the absorbing boundary at position $n = N$), $G(n_0, t)$ is the probability that $T \geq t$. In other words, the cumulative distribution function of the elongation times is $1 - G(n_0, t)$. We sum Eq. 15 over $n$ from $n = 0$ to $n = N - 1$ to obtain

$$\frac{\partial G(n_0,t)}{\partial t} = p_+[G(n_0+1,t) - G(n_0,t)] + p_-[G(n_0-1,t) + G(n_0,t)], \tag{17}$$

subject to the initial condition $G(n_0,0) = 1$ and boundary conditions $G(N,t) = 0$ and $G(0,t) = G(-1,t)$.

Equation 17 can be expressed and solved in terms of the first and second moments of the elongation time $T$, which can be written as

$$T(n_0) = \langle T\rangle = -\int_0^{+\infty} t\partial_t G(n_0,t)dt = \int_0^{+\infty} G(n_0,t)dt, \tag{18}$$

$$T_2(n_0) = \langle T^2\rangle = -\int_0^{+\infty} t^2\partial_t G(n_0,t)dt = 2\int_0^{+\infty} tG(n_0,t)dt. \tag{19}$$

We integrate Eq. 17 with respect to $t$ to obtain

$$-1 = p_+T(n_0+1) + p_-T(n_0-1) - (p_+ + p_-)T(n_0) = p_+[T(n_0+1) - T(n_0)] + p_-[T(n_0-1) - T(n_0)]. \tag{20}$$

The boundary conditions imply $T(N) = 0$, $T(0) = T(-1)$. To solve this difference equation we introduce

$$U(n_0) = T(n_0) - T(n_0-1), \tag{21}$$

and substituting into Eq. 20 yields

$$p_+U(n_0+1) - p_-U(n_0) = -1. \tag{22}$$

Solving the above two difference equations recursively, we obtain (38)

$$T(n_0) = \sum_{n=n_0+1}^{N}\frac{1}{p_+}\sum_{n'=0}^{n-1}\left(\frac{p_-}{p_+}\right)^{n'}. \tag{23}$$

By setting $K = \mathrm{p}_-/\mathrm{p}_+$ and observing that $0 \leq K < 1$, we can write

$$T(n_0) = \frac{1}{p_+}\sum_{n=n_0+1}^{N}\frac{1-K^n}{1-K} = \frac{1}{p_+(1-K)}\left[N - n_0 - \frac{K^{n_0+1} - K^{N+1}}{1-K}\right]. \tag{24}$$

Finally by letting $n_0 = 0$, we obtain the mean elongation time

$$\mu = \frac{1}{p_+(1-K)}\left[N - \frac{K(1-K^N)}{1-K}\right]. \tag{25}$$

For the variance of the elongation time we carry out a similar derivational. Multiplying by $t$ and integrating Eq. 17 over $t$, we obtain

$$-2T(n_0) = p_+T_2(n_0+1) + p_-T_2(n_0-1) - (p_+ + p_-)T_2(n_0) = p_+[T_2(n_0+1) - T_2(n_0)] + p_-[T_2(n_0-1) - T_2(n_0)]. \tag{26}$$

Once again solving the above equation recursively leads to

$$T_2(n_0) = -\sum_{n=n_0+1}^{N}U(n), \tag{27}$$

where $U(n)$ is given by

$$U(n) = -\frac{2}{p_+}\sum_{i=0}^{n-1}K^{n-i-1}T(n). \tag{28}$$

For $n_0 = 0$, the second moment becomes

$$\langle T^2\rangle = \frac{(1-K+6K^{N+1})}{p_+^2(1-K)^3}\left[N + \frac{(1+K)}{(1-K+6K^{N+1})}N^2 - \frac{2K(1-K^N)(2+K^{N+1})}{(1-K)(1-K+6K^{N+1})}\right]. \tag{29}$$

Finally, the variance of the elongation time is given by

$$\sigma^2 = \langle T^2\rangle - \langle T\rangle^2 = \frac{(1+K+K^{1+N})}{p_+^2(1-K)^3}\left[N - \frac{K(1-K^N)(4+K+K^{1+N})}{(1-K)(1+K+4K^{1+N})}\right]. \tag{30}$$

In the limit $K \ll 1$ (polymerization is overwhelmingly favored over depolymerization) we can express the mean elongation time and variance up to first-order in $K$ (see Eq. 2). In this regime, both the mean and the variance of the elongation time depend linearly on the template length $N$. Also the mean elongation time and variance approach the mean and variance

of the sum of $N$ independent and identically distributed (i.i.d.) exponential steps. Since the sum of i.i.d. exponential random variables is $\gamma$-distributed we can assume that in the small $K$ limit the elongation time, $T$, follows a $\gamma$-distribution

$$G(T|\alpha,\beta)=\frac{T^{\alpha-1}e^{-\frac{T}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}. \tag{31}$$

The parameters $\alpha$ and $\beta$ can be calculated from the mean and variance using the relationships $\mu=\alpha\beta$ and $\sigma^2=\alpha\beta^2$:

$$\alpha=\frac{(N+KN-K)^2}{N+4KN-4K}, \tag{32a}$$

$$\beta=\frac{1}{p_+}\frac{N+4NK-4K}{N+NK-K}. \tag{32b}$$

In the limit of large $N$ the distribution of elongation times approaches a Gaussian with mean and variance given by Eqs. 2a and 2b, respectively, in agreement with the Central Limit Theorem.

## APPENDIX B: ELONGATION PAUSES AND BACKTRACKING

We model the dynamics of backtracking in terms of an unbiased random walk with rate $c$. For simplicity, we characterize backtracked states by $l=-m$ where $1\le l\le M$. The probability, $P(l,t)$, of finding the TEC in state $l$ at time $t$ given it was in state $l=1$ at $t=0$, follows the Master equation given in Eq. 3. By using the Laplace transform $\tilde{p}(l,s)=\int_0^\infty P(l,t)e^{-\mathrm{st}}dt$, we can eliminate the time derivative in Eq. 3 and obtain an algebraic difference equation,

$$s\tilde{p}(l,s)-\delta_{l,1}=c\tilde{p}(l-1,s)+c\tilde{p}(l+1,s)-2c\tilde{p}(l,s), \tag{33}$$

where $\delta_{l,1}$ is the Kronecker delta.

### Case 1: uninterrupted backtracking

In this case (see *schematic diagram* in Fig. 10 *a*), the boundary conditions for Eq. 3 are: $P(0,\ t)\ =\ 0$ (absorbing) and $cP(M,\ t)\ =\ cP(M\ +\ 1,\ t)$ (reflecting).

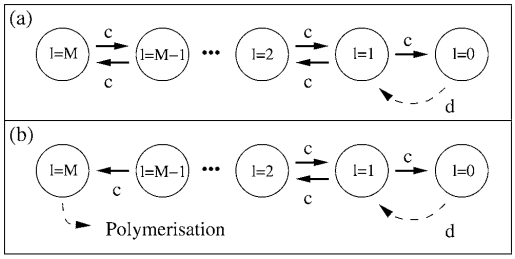



FIGURE 10 Schematic illustration of the two cases of restricted backtracking: (*a*) uninterrupted backtracking and (*b*) backtracking with transcript arrest. In both cases, variable $l$ denotes the number of nucleotides that the TEC has translocated backward. Translocation is possible up to $l=M$. A backtracking pause commences with the TEC at state $l=1$ (*dashed arrow*) and terminates when state $l=0$ is reached. For the case of backtracking with transcript arrest, the TEC is halted at state $l=M$ and can resume polymerization only with the aid of accessory factors (*left dashed arrow*).

We solve Eq. 33 (as described in (39)), with boundary conditions $\tilde{p}(0,s)=0\ c\tilde{p}(M,s)=c\tilde{p}(M+1,s)$, and obtain a closed formula for the Laplace transform of the probability flux to state $l=0$, $\tilde{F}(0,s)=c\tilde{p}(1,s)$,

$$\tilde{F}(0,s)=\frac{\sinh[M\phi(s)]-\sinh[(M-1)\phi(s)]}{\sinh[(M+1)\phi(s)]-\sinh[M\phi(s)]}, \tag{34}$$

where $\tanh\phi(s)=\sqrt{1-1/(s/2c+1)^2}$ The probability flux $F(0,\ t)$ is equivalent to the probability of exiting the pause at time $t$, and its Laplace transform, $\tilde{F}(0,s)$, evaluated at $s=0$, gives the probability of eventually exiting the pause (39). From Eq. 34, one obtains $\tilde{F}(0,s=0)=1$, i.e., the TEC will eventually exit the pause and resume elongation. $\tilde{F}(0,s)$ is also the moment-generating function containing all the positive integer moments of the exit time, as the coefficients of its power expansion in $s$ (39). We expand Eq. 34 to get

$$\tilde{F}(0,s)=1-\frac{M}{c}s+\mathrm{O}(s^2), \tag{35}$$

from which we obtain the mean pause duration $\langle t\rangle=\mathrm{M/c}$.

We can also use $\tilde{F}(0,s)$ to calculate the distribution of pauses. In the limit $t\gg 1/\mathrm{c}$, i.e., for times much longer than the time for a single step, Eq. 34 becomes

$$\tilde{F}(0,s)\approx\frac{\cosh\left[\sqrt{\frac{s}{c}}(M)\right]}{\cosh\left[\sqrt{\frac{s}{c}}(M+1)\right]}. \tag{36}$$

By inverting the above Laplace transform (58), we can express the distribution of pause duration, $\mathcal{P}(t)\equiv F(0,t)$ (for times $>1/c$) in terms of the Jacobi $\theta_1$ function,

$$\mathcal{P}(t)=a^{-1}\frac{\partial}{\partial\nu}\theta_1\left(\frac{1}{2}\nu a^{-1}\middle|ta^{-2}\right), \tag{37}$$

where $\nu=M/\sqrt{c}$, $a=(M+1)/\sqrt{c}$ and $\theta_1(z|t)$ can be expressed as the infinite series (58)

$$\theta_1(z|t)=\frac{1}{\sqrt{\pi t}}\sum_{\mathrm{n}=-\infty}^{\infty}(-1)^{\mathrm{n}}\exp[-(z+n-1/2)^2/t]. \tag{38}$$

Equation 37 leads to an expression for $\mathcal{P}(t)$. In particular, we obtain

$$\mathcal{P}(t)=\frac{-(M+1)}{\sqrt{\pi}\sqrt{ct^3}/2}\sum_{\mathrm{n}=-\infty}^{+\infty}(-1)^{\mathrm{n}}\exp\left[e^{-\frac{(1+M)^2}{ct}\left(n-\frac{1}{2(M+1)}\right)^2}\right]\left(n-\frac{1}{2(M+1)}\right). \tag{39}$$

Simpler expressions for $\mathcal{P}(t)$ can be obtained in the limits $t\ll M^2/c$ and $t\gg M^2/c$ (see Eq. 4 in main text). Plots of the analytic expression for $\mathcal{P}(t)$ along with the two asymptotic limits are shown in Fig. 11 *a*.

### Case 2: backtracking with transcript arrest

In this case (see *schematic diagram* in Fig. 10 *b*) the boundary conditions imposed on Eq. 3 are: $P(0,\ t)=P(M,\ t)=0$. Once again, we solve Eq. 33 with boundary conditions $\tilde{p}(0,s)=\tilde{p}(M,s)=0$ to obtain a closed expression for the Laplace transforms of the exit probabilities to either boundary,

$$\tilde{F}(0,s)=\frac{\sinh[(M-1)\phi(s)]}{\sinh[M\phi(s)]}, \tag{40a}$$

$$\tilde{F}(M,s)=\frac{\sinh[\phi(s)]}{\sinh[M\phi(s)]}, \tag{40b}$$

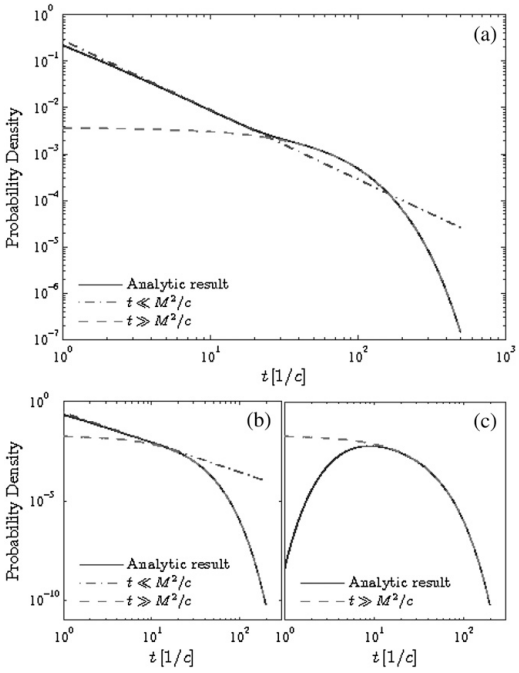



FIGURE 11 Analytic results for the duration of backtracking pauses, cases 1 and 2, for $M = 10$. (*a*) Case 1: restricted, uninterrupted backtracking. Probability distribution $\mathcal{P}(t)$ of exit time to absorbing boundary $l = 0$ in the presence of a reflecting boundary at $l = M$. Solid line corresponds to the analytic result Eq. 39, and dashed and dash-dotted lines to the two asymptotic limits in Eq. 4. (*b*, *c*). Case 2: restricted backtracking with transcript arrest. (*b*) Probability distribution $\mathcal{P}_0(t)$, of exit time to absorbing boundary $l = 0$ in the presence of an absorbing boundary at $l = M$. Solid line corresponds to the analytic result Eq. 45a, and dashed and dash-dotted lines to the two asymptotic limits in Eq. 5. (*c*) Probability distribution $\mathcal{P}_M(t)$ of exit time to absorbing boundary $l = M$ in the presence of an absorbing boundary at $l = 0$. Solid line corresponds to the analytic result Eq. 45b, and dashed line to the asymptotic limit in Eq. 5. In all cases, the initial state is assumed to be $l = 1$.

where $\tanh\phi(s) = \sqrt{1 - 1/(s/2c+1)^2}$. Evaluating the Laplace transforms at $s = 0$, we find that the TEC will eventually exit the pause either through state $l = M$ with probability $1/M$ or through state $l = 0$ with probability $1 - 1/M$. Once again, since Eq. 40a and Eq. 40b are generating functions, we can expand them in power series in $s$ to obtain the mean exit times to either boundary, $\langle t\rangle_0$ and $\langle t\rangle_M$:

$$\langle t\rangle_0 = \frac{2M-1}{6c}, \tag{41a}$$

$$\langle t\rangle_M = \frac{M^2-1}{6c}. \tag{41b}$$

In the presence of accessory factors the arrested transcript is cleaved and the TEC returns to a polymerization competent state. If we assume that the accessory factors act on relatively fast timescales, then the overall mean pause duration is just the weighted sum of $\langle t\rangle_0$ and $\langle t\rangle_M$, $\langle t\rangle = (M-1)/2c$.

We can also use $\tilde{F}(0,s)$ and $\tilde{F}(M,s)$ to calculate the full distribution for the exit times to either boundary. For times much longer than the time for a single step, $t \gg 1/c$, Eqs. 40a and 40b become

$$\tilde{F}(0,s) \approx \frac{\sinh\left[\sqrt{\frac{s}{c}}(M-1)\right]}{\sinh\left[\sqrt{\frac{s}{c}}M\right]}, \tag{42a}$$

$$\tilde{F}(M,s) \approx \frac{\sinh\left[\sqrt{\frac{s}{c}}\right]}{\sinh\left[\sqrt{\frac{s}{c}}M\right]}. \tag{42b}$$

By inverting the above Laplace transforms (58), the distribution of exit times to the boundaries at $l = 0$, $\mathcal{P}_0(t) \equiv F(0,t)$, and at $l = M$, $\mathcal{P}_M(t) \equiv F(M,t)$ (for times much greater than $1/c$) can be expressed in terms of the Jacobi $\theta_4$ function

$$\mathcal{P}_0(t) = a_0^{-1}\frac{\partial}{\partial\nu_0}\theta_4\left(\frac{1}{2}\nu_0 a_0^{-1}\middle| ta_0^{-2}\right), \tag{43a}$$

$$\mathcal{P}_M(t) = a_M^{-1}\frac{\partial}{\partial\nu_M}\theta_4\left(\frac{1}{2}\nu_M a_M^{-1}\middle| ta_M^{-2}\right), \tag{43b}$$

where $\nu_0 = (M-1)/\sqrt{c}$, $\nu_M = 1/\sqrt{c}$, $a_0 = a_M = M/\sqrt{c}$, and $\theta_4(z|t)$ can be expressed as the infinite series (58)

$$\theta_4(z|t) = \frac{1}{\sqrt{\pi t}}\sum_{n=-\infty}^{\infty}(-1)^n\exp[-(z+n+1/2)^2/t]. \tag{44}$$

Equations 43a and 43b lead to the following expressions for $\mathcal{P}_0(t)$ and $\mathcal{P}_M(t)$:

$$\mathcal{P}_0(t) = \frac{-M}{\sqrt{\pi}\sqrt{c}t^{3/2}}\sum_{n=-\infty}^{+\infty}\exp\left[e^{-\frac{M^2}{ct}\left(n-\frac{1}{2M}\right)^2}\right]\left(n-\frac{1}{2M}\right), \tag{45a}$$

$$\mathcal{P}_M(t) = \frac{-M}{\sqrt{\pi}\sqrt{c}t^{3/2}}\sum_{n=-\infty}^{+\infty}\exp\left[e^{-\frac{M^2}{ct}\left(n+\frac{M+1}{2M}\right)^2}\right]\left(n+\frac{M+1}{2M}\right). \tag{45b}$$

Simpler expressions for both $\mathcal{P}_0(t)$ and $\mathcal{P}_M(t)$ can be obtained in the limits $t \ll M^2/c$ and $t \gg M^2/c$ (see Eq. 5 in main text). Plots of the analytic expression for $\mathcal{P}_0(t)$ and $\mathcal{P}_M(t)$, along with the corresponding asymptotic limits are shown in Fig. 11, panels *b* and *c*, respectively.

## APPENDIX C: TRANSCRIPTION WITH RESTRICTED BACKTRACKING

Parameter $d$, the transition rate from translocation state $m = 0$ to $m = -1$ (see Fig. 2), determines the density of backtracking. If we assume rapid transition between the active transition states $m = 0$ and $m = 1$, then at each template position the TEC can 1), proceed with polymerization, with rate $p_+ = k_f(b/(a+b))$; 2), proceed with depolymerization, with rate $p_- = k_b(a/(a+b))$; or 3), enter state $m = -1$, with an effective rate $d' = d(a/(a+b))$ (see Fig. 2 *b*). Therefore, at a given position $n$, the TEC enters a pause with probability

$$P_{PAUSE} = \frac{d'}{d' + p_+ + p_-}. \tag{46}$$

Since we assume that a pause occurs independently at each template position, we can estimate the probability $P_{PAUSE}$ as the ratio of the expected number of pauses to the DNA template length i.e., $\delta/N = P_{PAUSE}$.

## Simulations

Simulated data were generated using standard Monte Carlo techniques (Gillespie algorithm) (59,60), implemented in ANSI-C. At each step a random, exponentially distributed, number was generated that was used as the time interval until the next transition. The parameter, $\lambda$, of the exponential distribution was set equal to the sum of the transition rates to all accessible states. To decide to which state the transition will occur, a state was picked randomly from all accessible states with a probability proportional to the corresponding transition rate. The total elapsed time and the state were updated accordingly and the process was repeated.

In the case of Model A and for each set of parameter values, data were generated by $10^3$ independent simulation runs. Since the values of parameters $a$ and $b$ are not known, arbitrary ones where used, which preserved the ratio found in the literature (see Table 1 of main text) and were higher than the rates of polymerization/depolymerization. In the case of the models of backtracking pauses and Model B, $10^5$ simulations were performed for each set of parameter values to accurately capture the shape of the distribution and the scaling behavior. The parameters for Model B were selected so as to yield the experimentally observed values (19,22). In particular, $a$, $b$, $k_f$, and $k_b$ were selected to yield an average velocity of 10 bp/s, while $d$ was chosen to yield 1 and 10 pauses/kb. For simulations of the integrated initiation/elongation/degradation model the parameters used were selected to match the ones observed in Golding et al. (11): $N = 4$ kb, $L = 100$ bp, $M = 10$ bp, $p_+ = 50\ s^{-1}$, $K = 0.01$, $c = 0.1\ s^{-1}$, $k_i = 0.02\ s^{-1}$, and $k_d = 310^{-4}\ s^{-1}$ and $d' = 0.05\ s^{-1}$ (yielding 1 pause/kb).

*Note added in proof:* After submission we became aware of the recent experimental work by Galburt et al. (57), which studies the distribution of durations of pauses of RNAP II and finds a $t^{-3/2}$ dependence as predicted by Eqs. 4 and 5.

This work was supported by the Engineering and Physical Sciences Research Council under grants No. EP/D003105 and EP/C011953/1 (N.C.), the Medical Research Council under grant No. G0300556 (N.C., C.M.P., T.B.L.), the Royal Society (T.B.L.), and the University of Leeds (M.V.).

## REFERENCES

1. Schrödinger, E. 1944. What is Life? Cambridge University Press, New York.
2. Paulsson, J. 2004. Summing up the noise in gene networks. *Nature*. 427:415–418.
3. Paulsson, J. 2005. Models of stochastic gene expression. *Phys. Life Rev.* 2:157–175.
4. Kærn, M., T. C. Elston, W. J. Blake, and J. J. Collins. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Gen.* 6:451–464.
5. Elowitz, M. B., A. J. Levine, E. D. Siggia, and P. S. Swain. 2002. Stochastic gene expression in a single cell. *Science*. 297:183–186.
6. Pedraza, J. M., and A. van Oudenaarden. 2005. Noise propagation in gene networks. *Science*. 307:1965–1969.
7. Yu, J., J. Xiao, X. Ren, K. Lao, and X. S. Xie. 2006. Probing gene expression in live cells, one protein molecule at a time. *Science*. 311: 1600–1603.
8. Greive, S. J., and P. H. von Hippel. 2005. Thinking quantitatively about transcriptional regulation. *Nat. Rev. Mol. Cell Biol.* 6:221–232.
9. von Hippel, P. H., and T. D. Yager. 1991. Transcription elongation and termination are competitive kinetic processes. *Proc. Natl. Acad. Sci. USA*. 88:2307–2311.
10. Swain, P. S., and A. Longtin. 2006. Noise in genetic and neural networks. *Chaos*. 16:026101.
11. Golding, I., J. Paulsson, S. M. Zawilski, and E. C. Cox. 2005. Real-time kinetics of gene activity in individual bacteria. *Cell*. 123:1025–1036.
12. Bai, L., T. Santangelo, and M. D. Wang. 2006. Single-molecule analysis of RNA polymerase transcription. *Annu. Rev. Biophys. Biomol. Struct.* 35:343–360.
13. Braslavsky, I., B. Hebert, E. Kartalov, and S. R. Quake. 2003. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. USA*. 100:3960–3964.
14. Malan, T. P., A. Kolb, H. Buc, and W. R. McClure. 1984. Mechanism of CRP-cAMP activation of *lac* operon transcription initiation activation of the P1 promoter. *J. Mol. Biol.* 180:881–909.
15. McClure, W. R. 1985. Mechanisms and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.* 54:171–204.
16. Zhang, X., T. Reeder, and R. Schleif. 1996. Transcription activation parameters at ara pBAD. *J. Mol. Biol.* 258:14–28.
17. Skinner, G. M., C. G. B. D. M. Quinn, J. E. Molloy, and J. G. Hoggett. 2004. Promoter binding, initiation, and elongation by bacteriophage T7 RNA polymerase. *J. Biol. Chem.* 279:3239–3244.
18. Saunders, A., L. J. Core, and J. T. Lis. 2006. Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.* 7:557–567.
19. Shaevitz, J. W., E. A. Abbondanzieri, R. Landick, and S. M. Block. 2003. Backtracking by single RNA polymerase molecules observed at near base pair resolution. *Nature*. 426:684–687.
20. Neuman, K. C., E. A. Abbondanzieri, R. Landick, J. Gelles, and S. M. Block. 2003. Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking. *Cell*. 115:437–447.
21. Herbert, K. M., A. L. Porta, B. J. Wong, R. A. Mooney, K. C. Neuman, R. Landick, and S. Block. 2006. Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell*. 125:1083–1094.
22. Forde, N. R., D. Izhaky, G. R. Woodcock, G. J. L. Wuite, and C. Bustamante. 2002. Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by *Escherichia coli* RNA polymerase. *Proc. Natl. Acad. Sci. USA*. 99:11682–11687.
23. Zenkin, N., Y. Yuzenkova, and K. Severinov. 2006. Transcript-assisted transcriptional proofreading. *Science*. 313:518–520.
24. Nickels, B. E., and A. Hochschild. 2004. Regulation of RNA polymerase through the secondary channel. *Cell*. 118:281–284.
25. Mote, J. J., and D. Reines. 1998. Recognition of a human arrest site is conserved between RNA polymerase II and prokaryotic RNA polymerases. *J. Biol. Chem.* 273:16843–16852.
26. Guajardo, R., and R. Sousa. 1997. A model for the mechanism of polymerase translocation. *J. Mol. Biol.* 256:8–19.
27. Julicher, F., and R. Bruinsma. 1998. Motion of RNA polymerase along DNA: a stochastic model. *Biophys. J.* 74:1169–1185.
28. Bai, L., A. Shundrovsky, and M. D. Wang. 2004. Sequence-dependent kinetic model for transcription elongation by RNA polymerase. *J. Mol. Biol.* 344:335–349.
29. Tadigotla, V. R., D. O. Maoiléidigh, A. M. Sengupta, V. Epshtein, R. H. Ebright, E. Nudler, and A. E. Ruckenstein. 2006. Thermodynamic and kinetic modeling of transcription pausing. *Proc. Natl. Acad. Sci. USA*. 103:4439–4444.
30. Raj, A., C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. 2006. Stochastic mRNA synthesis in mammalian cells. *PLOS Biol.* 4:1707–1719.
31. Chubb, J. R., T. Trcek, S. M. Shenoy, and R. H. Singer. 2006. Transcriptional pulsing of a developmental gene. *Curr. Biol.* 16:1018–1025.
32. Korzheva, N., A. Mustaev, M. Kozlov, A. Malhorta, V. Nikiforov, A. Goldfarb, and S. A. Darst. 2000. A structural model of transcription elongation. *Science*. 289:619–625.
33. Nudler, E., A. Mustaev, E. Lukhtanov, and A. Goldfarb. 1997. The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell*. 89:33–41.
34. Komissarova, N., and M. Kashlev. 1997. Transcriptional arrest: *Escherichia coli* RNA polymerase translocates backward, leaving the 3′ end of the RNA intact and extruded. *Proc. Natl. Acad. Sci. USA*. 279:1755–1760.
35. Borukhov, S., V. Sagitov, and A. Goldfarb. 1993. Transcript cleavage factor from *E. coli*. *Cell*. 72:459–466.

36. Fish, R. N., and C. M. Kane. 2002. Promoting elongation with transcript cleavage stimulatory factors. *Biochim. Biophys. Acta.* 1577:287–307.
37. van Kampen, N. G. 1992. Stochastic Processes in Physics and Chemistry. Elsevier, New York.
38. Gardiner, C. W. 2004. Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences, 3rd Ed. Springer-Verlag, Berlin, Germany.
39. Redner, S. 2001. A Guide to First-Passage Processes. Cambridge University Press, New York.
40. Gotta, S. L., O. L. Miller, and S. L. French. 1991. rRNA transcription rate in *Escherichia coli*. *J. Bacteriol.* 173:6647–6649.
41. Marko, J. F., and E. D. Siggia. 1995. Statistical mechanics of supercoiled DNA. *Phys. Rev. E.* 52:2912–2938.
42. Col, A. D., and T. B. Liverpool. 2004. Statistical mechanics of double-helical polymers. *Phys. Rev. E.* 69:61907–61911.
43. Derrida, B., S. A. Janowsky, J. L. Lebowitz, and E. R. Speer. 1993. Exact solution of the totally asymmetric simple exclusion process: shock profiles. *J. Stat. Phys.* 73:813–842.
44. Evans, M. R., D. P. Foster, C. Godrèche, and D. Mukamel. 1995. Spontaneous symmetry breaking in a one-dimensional driven diffusive system. *Phys. Rev. Lett.* 74:208–211.
45. Bremer, H., and M. Ehrenberg. 1995. Guanosine tetraphosphate as a global regulator of bacterial RNA synthesis: a model involving RNA polymerase pausing and queuing. *Biochim. Biophys. Acta.* 1262: 15–36.
46. Artsimovitch, I., and R. Landick. 2000. Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. *Proc. Natl. Acad. Sci. USA.* 97:7090–7095.
47. Li, B., M. Carey, and J. L. Workman. 2007. The role of chromatin during transcription. *Cell.* 128:707–719.
48. Carey, M., B. Li, and J. L. Workman. 2006. RSC Exploits histone acetylation to abrogate the nucleosomal block to RNA polymerase II elongation. *Mol. Cell.* 24:481–487.
49. van Zon, J. S., M. J. Morelli, S. T. Tanase, and P. R. ten Wolde. 2006. Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophys. J.* 91:4350–4367.
50. Blank, A., J. A. Gallant, R. R. Burgess, and L. A. Loeb. 1986. An RNA polymerase mutant with reduced accuracy of chain elongation. *Biochemistry.* 25:5920–5928.
51. Reference deleted in proof.
52. Rasmussen, E. B., and J. T. Lis. 1993. *In vivo* transcriptional pausing and cap formation on three *Drosophila* heat shock genes. *Proc. Natl. Acad. Sci. USA.* 90:7923–7927.
53. Adelman, K., M. T. Marr, J. Werner, A. Saunders, Z. Ni, E. D. Andrulis, and J. T. Lis. 2005. Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Mol. Cell.* 17: 103–112.
54. Lee, R. C., and V. Ambros. 2001. An extensive class of small RNAs in *Caenorhabditis elegans* complementarity to lin-14. *Science.* 295:862–864.
55. Lagos-Quintana, M., R. Rauhut, W. Lendeckel, and T. Tuschl. 2001. Identification of novel genes coding for small expressed RNAs. *Science.* 295:853–858.
56. Lau, N. C., L. P. Lim, E. G. Weinstein, and D. P. Bartel. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans. Science.* 295:858–862.
57. Galburt, E. A., S. W. Grill, A. Wiedmann, L. Lubkowska, J. Choy, E. Nogales, M. Kashlev, and C. Bustamante. 2007. Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner. *Nature.* 446:820–823.
58. Oberhettinger, F., and L. Badii. 1973. Tables of Laplace Transforms. Springer-Verlag, Berlin, Germany.
59. Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81:2340–2361.
60. Binder, K. 1986. Monte Carlo Methods in Statistical Physics, 2nd Ed. Springer-Verlag, Berlin, Germany.
61. Kingston, R. E., W. C. Nierman, and M. J. Chamberlin. 1981. A direct effect of guanosine tetraphosphate on pausing of *Escherichia coli* RNA polymerase during RNA chain elongation. *J. Biol. Chem.* 256:2787–2797.
62. Yin, H., M. D. Wang, K. Svodoba, R. Landick, and S. M. Block. 1995. Transcription against an applied force. *Science.* 270:1653–1657.
63. Schafer, D. A., J. Gelles, M. P. Sheetz, and R. Landick. 1991. Transcription by single molecules of RNA polymerase observed by light microscopy. *Nature.* 352:444–448.

PRL **102,** 258101 (2009) PHYSICAL REVIEW LETTERS week ending 26 JUNE 2009

# Backtracking and Proofreading in DNA Transcription

Margaritis Voliotis,[1,2] Netta Cohen,[1] Carmen Molina-París,[2] and Tanniemola B. Liverpool[3,*]
[1]*School of Computing, University of Leeds, Leeds, LS2 9JT, United Kingdom*
[2]*Department of Applied Mathematics, University of Leeds, Leeds, LS2 9JT, United Kingdom*
[3]*Department of Mathematics, University of Bristol, Bristol, BS8 1TW, United Kingdom*
(Received 13 June 2008; published 22 June 2009)

Biological cell function crucially relies on the accuracy of RNA sequences, transcribed from the DNA genetic code. To ensure sufficiently high fidelity in the face of high spontaneous error rates during transcription, error correction mechanisms must play an important role. A particular mechanism of transcriptional error correction involves backtracking of the RNA polymerase and RNA cleavage. Motivated by recent single molecule experiments characterizing the dynamics of backtracking, we present a microscopic model of this editing process. We show that such a mechanism can yield error frequencies that are in agreement with *in vivo* observations.

DOI: 10.1103/PhysRevLett.102.258101 PACS numbers: 87.15.rp, 82.39.Fk, 87.10.Mn, 36.20.Fz

The accuracy with which genetic information is processed is an essential factor in the survival and perpetuation of life. Efficient error correction mechanisms are therefore necessary for countering the frequent errors introduced by thermal fluctuations. For example, simple thermodynamic considerations suggest that during DNA transcription passive errors should occur with high propensity [$10^{-3}$–$10^{-2}$ errors/nucleotide (nt)]. Nevertheless, transcriptional error rates appear significantly lower ($10^{-5}$ errors/nt) [1]. Kinetic proofreading (KP) [2] provides a general phenomenological framework for understanding mechanisms that ensure low error rates and increased specificity in life processes [2]. To complement this general level of description, quantitative and predictive models that incorporate detailed information about specific biological processes are needed [3].

A particularly important example is the transcription of DNA into RNA. However, a comprehensive understanding of the mechanisms involved in transcriptional error correction is still lacking. Classical KP postulates the existence of a high energy intermediate along the polymerization pathway [2], acting as a fidelity checkpoint and enhancing the discriminatory power of the RNA polymerase (RNAP). Such an intermediate has indeed been suggested by recent structural studies of DNA transcription [4]. In addition, the RNAP's ability to induce cleavage of the RNA (or its so-called nuclease activity) suggests an alternative mode of transcriptional error correction, hereafter referred to as *nucleolytic proofreading*. This involves the backward sliding (*backtracking*) of the RNAP on the DNA template followed by *cleavage* of the nascent transcript [5]. In this manner previously misincorporated nucleotides can be discarded and repolymerized. The existence of these different proofreading mechanisms raises interesting questions regarding their relative roles in enhancing transcriptional fidelity. These can be answered by the construction of predictive models able to discriminate between the different processes.

During backtracking, the active site of the RNAP disengages from the $3'$ end of the transcript, and the transcription elongation complex (TEC), consisting of the RNAP and the DNA-RNA hybrid, steps backwards along the DNA [5]. The subsequent cleavage of the RNA chain is catalyzed by the active site of the polymerase and in certain cases accessory proteins are necessary to stimulate the reaction [6,7]. Recent single molecule experiments [8] provide support for nucleolytic proofreading by showing that (i) artificially induced misincorporation increases backtracking and (ii) cleavage factors reduce backtracking lifetimes.

In this Letter, we propose a stochastic, nonequilibrium model of transcription elongation involving polymerization of correct and incorrect nucleotides, backtracking, and RNA cleavage. We use the model to assess the role of nucleolytic proofreading in terms of the *error fraction*, defined as the ratio of probabilities of incorporating an incorrect as compared to a correct nucleotide at a given position of the transcript [2]. We study the problem both analytically, in different limits, and numerically, using stochastic simulations. Our results indicate that transcriptional error correction, involving backtracking by multiple nucleotides [8] and RNA cleavage, yields results consistent with multistep KP in the limit of high backtracking rates. More importantly, our results offer a quantitative understanding of nucleolytic proofreading by linking the observed error rate directly to the microscopic rates of the process. Finally, we suggest a number of experiments to test our model and clarify the role of nucleolytic proofreading in transcription.

Transcription elongation can be described in terms of two variables [9]. Let $n = 0, \ldots, N$ denote the length of the transcript or equivalently the template position of the last transcribed nucleotide [10]. Let $m = 0, \ldots, M$ denote the position of the TEC (specifically the RNAP's active site) relative to $n$ (i.e., the corresponding position of the active site along the DNA template is $n - m$). State $m = 0$ cor-

PRL **102,** 258101 (2009) PHYSICAL REVIEW LETTERS week ending 26 JUNE 2009

responds to a TEC in an active state, where polymerization of the next nucleotide can occur, while $m > 0$ corresponds to a TEC in a backtracked state [see Fig. 1(a)]. Extensive backtracking is often blocked by RNA secondary structures (e.g., hairpins) that are formed in the portion of the transcript outside the TEC [5]. Therefore, we assume that backtracking is restricted to a fixed distance $m = M$, which we take to be independent of $n$ [11]. The process starts with the TEC at $(n = 0, m = 0)$ and terminates at $(n = N, m = 0)$.

A schematic diagram of state transitions for the model is given in Fig. 1(b). Given a TEC in an active state $(n, m = 0)$, the TEC can either backtrack to state $(n, m = 1)$ with rate $k_b$ or polymerize the next nucleotide $(n + 1, m = 0)$. Polymerization of correct and incorrect nucleotides proceeds with effective rates $k_p$ and $\epsilon k_p$, respectively, yielding a spontaneous error fraction $\epsilon$. Once backtracked the TEC hops randomly between adjacent backtracked states $(n, 0 < m \leq M)$ at rate $c$. However, given an error at some position $n - l$ $(l \geq 0)$ transition of the TEC from state $(n, m = l + 1)$ to $(n, m = l)$ occurs at a slower rate $\bar{c}$. Finally, from each backtracked state cleavage can occur with rate $k_c$. Cleavage from any state $(n, m > l)$ ensures removal of the error.

The distinct hopping rate at an error site $(\bar{c} \ll c)$ is the key ingredient of this error correction process since it increases the likelihood of cleavage at states $(n, m > l)$. The ratio of the two hopping rates is given by $\bar{c}/c \approx e^{-\Delta G/kT}$ [12], where $\Delta G$ is the free energy increase due to the incorporation of an incorrect nucleotide in the RNA-DNA hybrid. The ratio of the polymerization rates for correct and incorrect nucleotides can also be approximated by $\Delta G$, i.e., $\epsilon \approx e^{-\Delta G/kT} \approx \bar{c}/c$ [2].

For the analytic treatment of the model we first consider the dynamics of the process at a fixed template position $n$

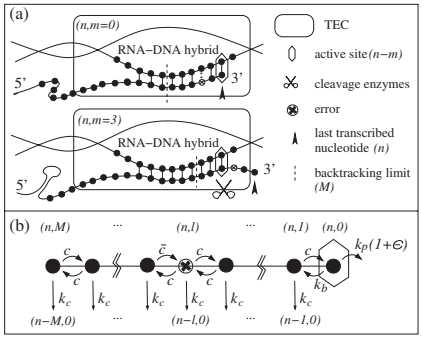



FIG. 1. (a) Schematic illustration of the model. The RNA is marked by $3'$ and $5'$. The transcription elongation complex (TEC) is depicted in the active $(n, m = 0)$ (top) and in a backtracked $(n, m = 3)$ (bottom) state, both with $M = 5$. (b) Schematic illustration of the TEC dynamics at a given position $n$. The TEC will eventually polymerize forward or cleave from one of the backtracked states.

which allows us to construct an effective model of the full elongation process. The master equation

$$\dot{\mathbf{P}}(t) = \mathbf{W}^{(s)} \cdot \mathbf{P}(t) \tag{1}$$

defines the stochastic dynamics of the TEC at a fixed position $n$. $\mathbf{P}$ is a column vector of size $(M + 1)$ with entries $P_m(t)$, the probabilities of finding the TEC at translocation state $m$ at time $t$, having started from $m = 0$ at $t = 0$. $\mathbf{W}^{(s)}$ is the $(M + 1) \times (M + 1)$ transition matrix. The transcription index $s$ is a binary list of 0's and 1's representing the sequence of correct (0) and incorrect (1) nucleotides along the entire transcript. In particular, $s \in S^n$ with $S \equiv \{0, 1\}$ (i.e., for an error at position $n - l$, $s_{n-l} = 1$). The general tridiagonal structure of $\mathbf{W}^{(s)}$ is given below. Along the main diagonal: $W^{(s)}_{j,j} = -[2c + s_{n-j+2}(\bar{c} - c) + k_c]$ except for $W^{(s)}_{1,1} = -[(1 + \epsilon)k_p + k_b]$ and $W^{(s)}_{M+1,M+1} = -[c + s_{n-M+1}(\bar{c} - c) + k_c]$. Along the first diagonal below the main: $W^{(s)}_{j+1,j} = c$, except for $W^{(s)}_{2,1} = k_b$. Along the first diagonal above the main: $W^{(s)}_{j,j+1} = c + s_{n-j+1}(\bar{c} - c)$. All other components are zero. Note that the form of the matrix depends only on the last $M$ elements of $s$.

The above formulation of $\mathbf{W}^{(s)}$ implies $M + 1$ absorbing boundaries, corresponding either to polymerization from state $m = 0$ or cleavage from each possible backtracked state $(1 \leq m \leq M)$. By applying the Laplace transform $\tilde{\mathbf{P}}(z) = \int_0^\infty e^{-zt}\mathbf{P}(t)dt$ to Eq. (1), we obtain a system of algebraic difference equations, which can be used to derive the splitting probabilities $p_m$ for eventually hitting boundary $m$ $(0 \leq m \leq M)$ and the corresponding conditional mean exit times, $t_m$ [13]. Note that both $p_m$ and $t_m$ depend on the sequence $s$.

We now use the splitting probabilities $p_m$ to construct an effective model for the elongation dynamics. Let $\Pi^{(s)}_n(t)$ be the probability of finding a transcript of length $n$ and index $s$ at time $t$. The transcript can either be extended by one nucleotide (through polymerization) or get shortened by up to $M$ nucleotides (through backtracking and cleavage). These transitions occur with rates $r_m$, proportional to the splitting probabilities obtained above, i.e., $r_m = p_m/\tau$ $(0 \leq m \leq M)$, where $\tau$ defines a sufficiently long time scale (i.e., $\tau \gg t_m$, $0 \leq m \leq M$). We note that all results given below depend only on the relative rates and hence do not depend on the exact definition of $\tau$. Summing over $s$, one obtains $\Pi_n(t) = \sum_{s \in S^n} \Pi^{(s)}_n(t)$, the probability of finding a transcript of length $n$ irrespective of its composition. The dynamics of $\Pi_n(t)$ can be expressed as

$$\frac{d\Pi_n}{dt} = \mathcal{J}_{n-1|0} - \mathcal{J}_{n|0} + \sum_{m=1}^{M} (\mathcal{J}_{n+m|m} - \mathcal{J}_{n|m}), \tag{2}$$

where $\mathcal{J}_{n|m} = \sum_{s \in S^n} r^{(s)}_m \Pi^{(s)}_n(t)$. For any specific $M$, Eq. (2) can be used to obtain an expression for $\mathcal{P}_n$ $(\bar{\mathcal{P}}_n)$, the probability of reaching the terminal position $N$, having

PRL **102,** 258101 (2009) PHYSICAL REVIEW LETTERS week ending 26 JUNE 2009

incorporated a correct (incorrect) nucleotide at position $n$. The error fraction for position $n$ is defined as $\mathcal{E} \equiv \bar{\mathcal{P}}_n/\mathcal{P}_n$. Given a large ensemble of completed transcripts, $\mathcal{E}$ gives the ratio of the number of transcripts with correct nucleotides to those with incorrect nucleotides at position $n$.

For simplicity, in most of the analysis below, we treat the case $M = 1$, where the TEC can backtrack by only one nucleotide. We introduce the following dimensionless quantities to characterize the competing processes in the model: $\alpha_1 \equiv k_c/c$ and $\alpha_2 \equiv k_c/\bar{c} = \alpha_1/\epsilon$ capture the efficiency of cleavage of correct and incorrect nucleotides, respectively, and $K \equiv k_p/k_b$ the tendency of the TEC to backtrack. The splitting probabilities, obtained from Eq. (1), are determined completely by the identity of the last incorporated nucleotide, $s_n$. We denote these splitting probabilities when $s_n = 0$ or 1 with $p_i$ and $\bar{p}_i$, respectively, where $i = 0$ corresponds to polymerization of $s_n$ and $i = 1$ to cleavage. The splitting probabilities take the form $p_0 = \kappa(\epsilon, \alpha_1)/[\kappa(\epsilon, \alpha_1) + \alpha_1]$, $p_1 = 1 - p_0$, $\bar{p}_0 = \kappa(\epsilon, \alpha_2)/[\kappa(\epsilon, \alpha_2) + \alpha_2]$, and $\bar{p}_1 = 1 - \bar{p}_0$, where $\kappa(\epsilon, a) = K(1 + \epsilon)(1 + a)$.

Given the above splitting probabilities, Eq. (2) can now be written for $M = 1$. Laplace transform techniques [13] then yield the termination probabilities $\mathcal{P}_n = \mathcal{N} p_0/(1 - A_n p_0)$ and $\bar{\mathcal{P}}_n = \mathcal{N}\epsilon\bar{p}_0/(1 - A_n\bar{p}_0)$. Here, $\mathcal{N}$ is the normalization constant (such that $\mathcal{P}_n + \bar{\mathcal{P}}_n = 1$), and in the limit $\epsilon \to 0$, one has $A_n \approx \beta(\beta^{N-n} - 1)/(\beta^{N-n+1} - 1)$, where $\beta = p_1/p_0$ [14]. Thus, the error fraction for $M = 1$ is

$$\mathcal{E} = \frac{\epsilon\bar{p}_0(1 - A_n p_0)}{p_0\,(1 - A_n\bar{p}_0)}. \tag{3}$$

Figure 2 (top panel) shows the error fraction $\mathcal{E}$ for different positions $n$ as a function of $K$.

We next consider two limits where $\mathcal{E}$ attains a constant value independent of position $n$. In the limit $K \gg 1$, one expects that the rare backtracking can hardly improve the error fraction. Indeed, in this limit Eq. (3) reduces to $\mathcal{E} \approx \epsilon$. On the other hand, in the limit $K \ll \alpha_1 \ll \epsilon$, cleavage events dominate the process, and Eq. (3) reduces to $\mathcal{E} \approx \epsilon\bar{p}_0/p_0$, or, in terms of the microscopic rate parameters, $\mathcal{E} \approx \epsilon\bar{c}/c$. Hence, the error fraction depends only on $\epsilon$ and the ratio of hopping rates. Since we take these two quantities to be approximately equal, we obtain the limiting error fraction for $M = 1$ to be $\mathcal{E} \approx \epsilon^2$. These two limits are illustrated in Fig. 2 (bottom panel). Numerical data were generated using stochastic simulations [15] of the full elongation model.

In the more general case of $1 \le M \ll 1/\epsilon$ (i.e., with at most one error occurring in a region of $M$ nucleotides), it can similarly be shown that in the same limit ($K \ll \alpha_1 \ll \epsilon$) the error fraction is

$$\mathcal{E} \approx \epsilon^{M+1}\frac{M^M}{\Gamma(M + 1)}, \tag{4}$$

where $\Gamma$ denotes the Gamma function. Thus, nucleolytic proofreading can result in error fractions that scale exponentially with the maximum backtracking distance $M$. We note that the error fraction attained by KP has a similar dependence on the number of intermediate states [2].

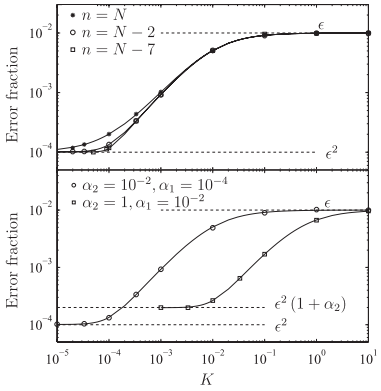



FIG. 2. The error fraction as a function of $K$ ($M = 1$ case). Analytic results [Eq. (3)] are plotted as solid lines, while markers show results obtained from stochastic simulations of the elongation model. Top: The error fraction for different positions with $\alpha_1 = 10^{-4}$, $\alpha_2 = 10^{-2}$, $\epsilon = 10^{-2}$, and $N = 9$. Bottom: The error fraction for different cleavage efficiencies with $\epsilon = 10^{-2}$, $n = N - 2$, and $N = 4$. Dashed lines show limits discussed in text.

So far we have assumed a constant backtracking rate. However, the presence of an error in the RNA-DNA hybrid could destabilize the TEC, causing more frequent backtracks. A simple model capturing this has backtracking rate $\bar{k}_b$ if an error is within $M$ nucleotides from the 3′ RNA end, and $k_b$ otherwise ($k_b < \bar{k}_b$). This can be approximated by an effective backtracking rate $k_b^* = M\epsilon\bar{k}_b + k_b$, giving rise to an effective $K^* = k_p/k_b^* = K/[\epsilon/\epsilon^* + 1]$, where $K = k_p/k_b$ and $\epsilon^* = k_b/(\bar{k}_b M)$. Furthermore, a reasonable assumption is that the TEC rarely backtracks when no errors are present, i.e., $K \gg 1$. Parameter $\epsilon^*$ is an intrinsic error scale: When $\epsilon/\epsilon^* \ll 1$ the high $K^*$ regime is obtained, whereas for $\epsilon/\epsilon^* \gg 1$ the behavior of the model is shifted towards the low $K^*$ regime [16].

Let us now estimate the error fractions implied by our model taking into account information from experimental studies. The spontaneous error fraction $\epsilon$ can be calculated from the free energy difference due to a misincorporated nucleotide ($\Delta G \approx 4$–$7kT$), i.e., $\epsilon \approx e^{-\Delta G/kT} \approx 10^{-3}$–$10^{-2}$ [1]. An estimate of the cleavage rate (for bacterial RNAP in the presence of saturating concentrations of accessory cleavage factors) based on biochemical experiments is $k_c \approx 0.1$–$1\ \mathrm{s}^{-1}$ [17]. Finally, single molecule experiments have suggested that the TEC hops between backtracked states with rate $c \approx 1$–$10\ \mathrm{s}^{-1}$ [8]. Using estimates of the maximum error $\epsilon \approx 0.01$, slowest cleavage rates $k_c \approx 0.1\ \mathrm{s}^{-1}$ and fastest hopping rate $c \approx 10\ \mathrm{s}^{-1}$ we can obtain estimates of the lower bounds on the

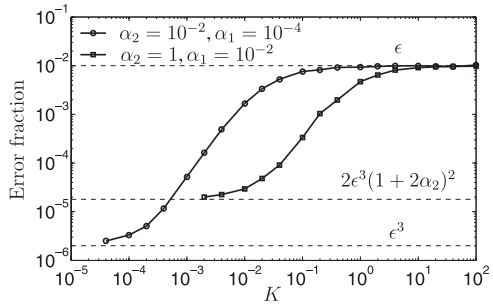



FIG. 3. Error fraction as a function of $K$ ($M = 2$ case). Results were obtained using stochastic simulations of the model for $N = 4$, $\epsilon = 10^{-2}$, and $\alpha_1 = 10^{-2}, 10^{-4}$.

"cleavage efficiencies" $\alpha_1 \sim 0.01$ and $\alpha_2 \sim 1$. These estimates yield error fractions comparable to the ones observed *in vivo*, even for $M = 1$ but sufficiently low values of $K$ (see Fig. 2, bottom panel). Most importantly, however, low error fractions can be obtained in our model even well away from the limiting regime with small $M$ (see Fig. 3 for the $M = 2$ case).

In summary, we have presented a microscopic model of a transcription editing mechanism, involving backtracking and RNA cleavage. Our work extends the existing qualitative description of the process by linking the observed error rates directly to microscopic rate parameters. Backtracking by more than one nucleotide provides a multiple-checking reaction, which probes the fidelity of the last few nucleotides before the next polymerization step. We find, in accordance with the KP scheme, that the greater the delay introduced by this step, the greater the accuracy of the process [2]. Consistent with this picture, the minimum error fraction is obtained in the limit where backtracking and cleavage dynamics dominate the process. In this limit, the error fraction scales exponentially with the maximum backtracking distance $M$.

Recent experiments have provided support for at least two mechanisms of transcriptional error correction [4,8,18,19]. The first one involves a fidelity checkpoint during the nucleotide addition cycle [20], whereas the second involves backtracking of the RNAP and RNA cleavage. Our model suggests experiments that would provide the quantitative details required to discriminate between these mechanisms and elucidate their relative roles in transcriptional proofreading.

A particular prediction of our model is the strong dependence of transcriptional fidelity on backtracking *rates*. For example, guanine-cytosine–rich domains that lead to lower backtracking rates (due to the increased stability of the RNA-DNA hybrid) [21] should reduce the efficiency of error correction. More importantly, single molecule manipulation techniques can be used to vary backtracking rates in a controlled manner and validate our model. In particular, applying a load is expected to strongly affect nucleolytic proofreading since the TEC moves a distance $\sim M\delta x$ (where $\delta x = 3.4$ Å) during the backtracking phase. In contrast, minor effects are expected for proofreading mechanisms along the polymerization pathway, since they should only involve small movements ($\ll \delta x$) of the enzyme. Finally, experimental studies have already revealed that specific mutations in the sequence of RNAP can have a profound effect on transcriptional fidelity [22]. By precisely studying the effects of the mutations on backtracking rates, single molecule experiments with such mutant RNAPs can be used to assess whether nucleolytic proofreading can compensate for such deficiencies.

T. B. L. acknowledges the hospitality of the Curie Institute in Paris.

*t.liverpool@bristol.ac.uk

[1] A. Blank *et al.*, Biochemistry **25**, 5920 (1986).
[2] J. J. Hopfield, Proc. Natl. Acad. Sci. U.S.A. **71**, 4135 (1974); J. Ninio, Biochimie **57**, 587 (1975).
[3] T. McKeithan, Proc. Natl. Acad. Sci. U.S.A. **92**, 5042 (1995); J. Yan, M. O. Magnasco, and J. F. Marko, Nature (London) **401**, 932 (1999); P. S. Swain and E. D. Siggia, Biophys. J. **82**, 2928 (2002).
[4] D. G. Vassylyev *et al.*, Nature (London) **448**, 163 (2007).
[5] S. J. Greive and P. H. von Hippel, Nat. Rev. Mol. Cell Biol. **6**, 221 (2005).
[6] M. J. Thomas, A. A. Platas, and D. K. Hawley, Cell **93**, 627 (1998).
[7] R. N. Fish and C. M. Kane, Biochim. Biophys. Acta, Gene Struct. Expr. **1577**, 287 (2002).
[8] J. W. Shaevitz *et al.*, Nature (London) **426**, 684 (2003); E. A. Galburt *et al.*, Nature (London) **446**, 820 (2007).
[9] M. Voliotis *et al.*, Biophys. J. **94**, 334 (2008).
[10] We define $n = 0$ to be the position at which the elongation phase is entered, a few (8–10) nucleotides downstream of the actual transcriptional starting point.
[11] For positions $n < M$, backtracking is restricted to $m = n$.
[12] J. Howard, *Mechanics of Motor Proteins and the Cytoskeleton* (Sinauer, Sunderland, MA, 2001).
[13] S. Redner, *A Guide to First-Passage Processes* (Cambridge University Press, Cambridge, England, 2001).
[14] $A_n$ can be understood as the probability that starting from position $n + 1$ cleavage to position $n$ will occur (before the terminal position $N$ has been reached).
[15] D. T. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[16] Error correction is attempted more often when the spontaneous error rate is high.
[17] E. Sosunova *et al.*, Proc. Natl. Acad. Sci. U.S.A. **100**, 15 469 (2003).
[18] N. Zenkin, Y. Yuzenkova, and K. Severinov, Science **313**, 518 (2006).
[19] N. Alic *et al.*, Proc. Natl. Acad. Sci. U.S.A. **104**, 10 400 (2007).
[20] Such a mechanism can be described by a model similar to our model with $M = 1$ (with a minimum error fraction limited to $\epsilon^2$).
[21] T. Ambjörnsson *et al.*, Phys. Rev. Lett. **97**, 128105 (2006).
[22] S. F. Holmes *et al.*, J. Biol. Chem. **281**, 18 677 (2006).

# Bibliography


[1] E. A. Abbondanzieri, W. J. Greenleaf, J. W. Shaevitz, R. Landick, and S. M. Block. Direct observation of base-pair stapping by RNA polymerase. *Nature*, 438:460–465, 2005.

[2] P. K. Ajikumar, K. Tyo, S. Carlsen, O. Mucha, T. H. Phon, and G. Stephanopoulos. Terpenoids: Opportunities for biosynthesis of natural product drugs using engineered microorganisms. *Mol. Pharm.*, 5(2):167–190, 2008.

[3] B. Alberts, A. Johnson, J. Lewis, Raff M., K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, 4nd edition, 2002.

[4] N. Alic, N. Ayoub, E. Landrieux, E. Favry, P. Baudouin-Cornu, M. Riva, and C. Carles. Selectivity and proofreading both contribute significantly to the fidelity of RNA polymerase III transcription. *Proc. Natl. Acad. Sci. USA*, 104:10400–10405, 2007.

[5] T. Ambjörnsson, S.K. Banik, O. Krrichevsky, and R. Metzler. Sequence sensitivity of breathing dynamics in heteropolymer DNA. *Phys. Rev. Lett.*, 97:128105, 2006.

[6] J. C. Anderson, E. J. Clarke, A. P. Arkin, and C. A. Voigt. Environmentally controlled invasion of cancer cells by engineered bacteria. *J. Mol. Biol.*, 355(4):619–627, 2006.

[7] I. Artsimovitch and R. Landick. Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. *Proc. Natl. Acad. Sci. USA*, 97:7090–7095, 2000.

[8] M. R. Atkinson, M. A. Savageau, J. T. Myers, and A. J. Ninfa. Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in escherichia coli. *Cell*, 113(5):597–607, 2003.

[9] D. W. Austin, M. S. Allen, J. M. McCollum, R. D. Dar, J. R. Wilgus, G. S. Sayler, N. F. Samatova, C. D. Cox, and M. L. Simpson. Gene network shaping of inherent noise spectra. *Nature*, 439(7076):608–611, 2006.

[10] L. Bai, A. Shundrovsky, and M. D. Wang. Sequence-dependent kinetic model for transcritpion elongation by RNA polymarese. *J. Mol. Biol.*, 344:335–349, 2004.

[11] F. K. Balagadde, H. Song, J. Ozaki, C. H. Collins, M. Barnet, F. H. Arnold, S. R. Quake, and L. C. You. A synthetic escherichia coli predator-prey ecosystem. *Mol. Syst. Biol.*, 4:8, 2008.

[12] S. Basu, Y. Gerchman, C. H. Collins, F. H. Arnold, and R. Weiss. A synthetic multicellular system for programmed pattern formation. *Nature*, 434(7037):1130–1134, 2005.

[13] S. Basu, R. Mehreja, S. Thiberge, M. T. Chen, and R. Weiss. Spatiotemporal control of gene expression with pulse-generating networks. *Proc. Natl. Acad. Sci. USA*, 101(17):6355–6360, 2004.

[14] A. Becskei and L. Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590–593, 2000.

[15] B. N. Belintsev, S. K. Zavriev, and M. F. Shemyakin. On the promoter complex-formation rate of *escherichia coli* RNA-polymerases with T7-phage DNA. *Nucleic Acids Res.*, 8(6):1391–1404, 1980.

[16] E. Ben-Jacob, I. Cohen, and H. Levine. Cooperative self-organization of microorganisms. *Adv. Phys.*, 49(4):395–554, 2000.

[17] K. Binder. *Monte Carlo methods in Statistical Physics*. Springer-Verlag, Berlin, Germany, 2nd edition, 1986.

[18] A. Blank, J. A. Gallant, R. R. Burgess, and L. A. Loeb. An RNA polymerase mutant with reduced accuracy of chain elongation. *Biochemistry*, 25:5920–5928, 1986.

[19] R. Blossey and H. Schiessel. Kinetic proofreading of gene activation by chromatin remodeling. *HSFP J.*, 2:167–170, 2008.

[20] S. Borukhov, V. Sagitov, and A. Goldfarb. Transcript cleavage factor from *E. coli*. *Cell*, 72:459–466, 1993.

[21] S. J. Bray. Notch signalling: a simple pathway becomes complex. *Nat. Rev. Mol. Cell Biol.*, 7(9):678–689, 2006.

[22] K. Brenner, D. K. Karig, R. Weiss, and F. H. Arnold. Engineered bidirectional communication mediates a consensus in a microbial biofilm consortium. *Proc. Natl. Acad. Sci. USA*, 104(44):17300–17304, 2007.

[23] J. G. Cao and E. A. Meighen. Purification and structural identification of an autoinducer for the luminescence system of *vibrio harveyi*. *J. Biol. Chem.*, 264(36):21670–21676, 1989.

[24] M. Carey, B. Li, and J. L. Workman. RSC exploits histone acetylation to abrogate the nucleosomal block to RNA polymerase II elongation. *Mol. Cell*, 24:481–487, 2006.

[25] P. J. Choi, L. Cai, K. Frieda, and S. Xie. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322(5900):442–446, 2008.

[26] Greenberg E. P. Choi S. H. Genetic evidence for multimerization of LuxR, the transcriptional activator of *vibrio fischeri* luminescence. *Mol. Mar. Biol. Biotheconol.*, 1:408–413, 1992.

[27] J. R. Chubb, T. Trcek, S. M. Shenoy, and R. H. Singer. Transcriptional pulsing of a developmental gene. *Curr. Biol.*, 16:1018–1025, 2006.

[28] A. De Col and T. B. Liverpool. Statistical mechanics of double-helical polymers. *Phys. Rev. E*, 69:61907–61911, 2004.

[29] L. Core and M. Perego. TPR-mediated interaction of RapC with ComA inhibits response regulator-DNA binding for competence development in *bacillus subtilis*. *Mol. Microbiol.*, 49(6):1509–1522, 2003.

[30] P. Cramer. Self-correcting messages. *Science*, 313(5786):447–448, 2006.

[31] P. Cramer, K. J. Armache, S. Baumli, S. Benkert, E. Brueckner, C. Buchen, G. E. Damsma, S. Dengl, S. R. Geiger, A. J. Jaslak, A. Jawhari, S. Jennebach, T. Kamenski, H. Kettenberger, C. D. Kuhn, E. Lehmann, K. Leike, J. E. Sydow, and A. Vannini. Structure of eukaryotic RNA polymerases. *Annu. Rev. Biophys.*, 37:337–352, 2008.

[32] X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, and R. H. Singer. In vivo dynamics of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.*, 14(9):796–806, 2007.

[33] X. Darzacq, J. Yao, D. R. Larson, S. Z. Causse, L. Bosanac, V. de Turris, V. M. Ruda, T. Lionnet, D. Zenklusen, B. Guglielmi, R. Tjian, and R. H. Singer. Imaging transcription in living cells. *Annu. Rev. Biophys.*, 38:173–196, 2009.

[34] N. G. de Bruijn. *Asymptotic methods in analysis*. Dover, New York, 1981.

[35] M. Depken, E. A. Galburt, and S. W. Grill. The origin of short transcriptional pauses. *Biophys. J.*, 96(6):2189–2193, 2009.

[36] B. Derrida, S. A. Janowsky, J. L. Lebowitz, and E. R. Speer. Exact solution of the totally asymmetric simple exclusion process: Shock profiles. *J. Stat. Phys.*, 73:813–842, 1993.

[37] M. Dobrzynski and F. J. Bruggeman. Elongation dynamics shape bursty transcription and translation. *Proc. Natl. Acad. Sci. USA*, 106(8):2583–2588, 2009.

[38] Y. H. Dong, L. H. Wang, J. L. Xu, H. B. Zhang, X. F. Zhang, and L. H. Zhang. Quenching quorum-sensing-dependent bacterial infection by an N-acyl homoserine lactonase. *Nature*, 411(6839):813–817, 2001.

[39] P. Dufour, S. Jarraud, F. Vandenesch, T. Greenland, R. P. Novick, M. Bes, J. Etienne, and G. Lina. High genetic variability of the *agr* locus in *staphylococcus* species. *J. Bacteriol.*, 184(4):1180–1186, 2002.

[40] T. Ellis, X. Wang, and J. J. Collins. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nat. Biotechnol.*, 27(5):465–471, 2009.

[41] M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

[42] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297:183–186, 2002.

[43] M. R. Evans, D. P. Foster, C. Godrèche, and D. Mukamel. Spontaneous symmetry breaking in a one dimensional driven diffusive system. *Phys. Rev. Lett.*, 74:208–211, 1995.

[44] W. Feller. *An introduction to probability theory and its applications, Vol. I.* Willey, New York, 2nd edition, 1971.

[45] R. N. Fish and C. M. Kane. Promoting elongation with transcript cleavage stimulatory factors. *Biochim. Biophys. Acta*, 1577:287–307, 2002.

[46] N. R. Forde, D. Izhaky, G. R. Woodcock, G. J. L. Wuite, and C. Bustamante. Using mechanical force to probe the mechanism of pausing and arrest during continuous elongation by *escherichia coli* RNA polymerase. *Proc. Natl. Acad. Sci. USA*, 99:11682–11687, 2002.

[47] E. A. Galburt, S. W. Grill, A. Wiedmann, L. Lubkowska, J. Choy, E. Nogales, M. Kashlev, and C. Bustamante. Backtracking determines the force sensitivity of RNAP II in a factor-dependent manner. *Nature*, 446:820–823, 2007.

[48] L. Gammaitoni, P. Hanggi, P. Jung, and F. Marchesoni. Stochastic resonance. *Rev. Mod. Phys.*, 70(1):223–287, 1998.

[49] J. Garcia-Ojalvo, M. B. Elowitz, and S. H. Strogatz. Modeling a synthetic multicellular clock: Repressilators coupled by quorum sensing. *Proc. Natl. Acad. Sci. USA*, 101(30):10955–10960, 2004.

[50] C. W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Springer-Verlag, Berlin, Germany, 3nd edition, 2004.

[51] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *escherichia coli*. *Nature*, 403(6767):339–342, 2000.

[52] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340–2361, 1977.

[53] K. Glover-Cutter, S. Kim, J. Espinosa, and D. L. Bentley. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat. Struct. Mol. Biol.*, 15(1):71–78, 2008.

[54] K. I. Goh, B. Kahng, and K. H. Cho. Sustained oscillations in extended genetic oscillatory systems. *Biophys. J.*, 94(11):4270–4276, 2008.

[55] I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123:1025–1036, 2005.

[56] A. B. Goryacgev, D. J. Toh, and T. Lee. Systems analyasis of a quorum sensing netwrok: Design constrains imposed by the functional requirements, network topology and kinetic constants. *BioSystems*, 83:178, 2006.

[57] S. L. Gotta, O. L. Miller, and S. L. French. rRNA transcription rate in *Escherichia coli*. *J. Bacteriol.*, 173:6647–6649, 1991.

[58] S. J. Greive and P. H. von Hippel. Thinking quantitatively about transcripptional regulation. *Nat. Rev. Mol. Cell Biol.*, 6:221–232, 2005.

[59] A. D. Grossman. Genetic networks controlling the initiation of sporulation and the development of genetic competence in *bacillus subtilis*. *Annu. Rev. Genet.*, 29:477–508, 1995.

[60] R. Guajardo and R. Sousa. A model for the mechanism of polymerase translocation. *J. Mol. Biol.*, 256:8–19, 1997.

[61] B. L. Hanzelka and E. P. Greenberg. Evidence that the N-terminal region of the *vibrio fischeri* LUXR protein constitutes an autoinducer-binding domain. *J. Bacteriol.*, 177(3):815–817, 1995.

[62] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761):C47–C52, 1999.

[63] K. M. Herbert, W. J. Greenleaf, and S. M. Block. Single-molecule studies of RNA polymerase: Motoring along. *Annu. Rev. Biochem.*, 77:149–176, 2008.

[64] K. M. Herbert, A. La Porta, B. J. Wong, R. A. Mooney, K. C. Neuman, R. Landick, and S.M. Block. Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell*, 125:1083–1094, 2006.

[65] C. Hodges, L. Bintu, L. Lubkowska, M. Kashlev, and C. Bustamante. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science*, 325(5940):626–628, 2009.

[66] S.F. Holmes, T.J. Santangelo, C.K. Cunningham, J.W. Roberts, and D.A. Erie. Kinetic investigation of escherichia coli RNA polymerase mutants that influence nucleotide discrimination and transcription fidelity. *J. Biol. Chem.*, 281:18677–18683, 2006.

[67] S. Hooshangi, S. Thiberge, and R. Weiss. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proc. Natl. Acad. Sci. USA*, 102(10):3581–3586, 2005.

[68] J. J. Hopfield. Kinetic proofreading: A new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. USA*, 71(10):4135–4139, October 1974.

[69] J. Howard. *Mechanics of Motor Proteins and the Cytoskeleton*. Sinauer Associates, 2001.

[70] C. Joo, H. Balci, Y. Ishitsuka, C. Buranachai, and T. Ha. Advances in single-molecule fluorescence methods for molecular biology. *Annu. Rev. Biochem.*, 77:51–76, 2008.

[71] F. Julicher and R. Bruinsma. Motion of RNA polymerase along DNA: A stochastic model. *Biophys. J.*, 74:1169–1185, 1998.

[72] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Gen.*, 6:451–464, 2005.

[73] A. N. Kapanidis, E. Margeat, S. O. Ho, E. Kortkhonjia, S. Weiss, and R. H. Ebright. Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science*, 314(5802):1144–1147, 2006.

[74] R. E. Kingston, W. C. Nierman, and M. J. Chamberlin. A direct effect of guanosine tetraphosphate on pausing of *escherichia coli* RNA polymerase during RNA chain elongation. *J. Biol. Chem.*, 256:2787–2797, 1981.

[75] H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002.

[76] N. Korzheva, A. Mustaev, M. Kozlov, A. Malhorta, V. Nikiforov, A. Goldfarb, and S. A. Darst. A structural model of transcription elongation. *Science*, 289:619–625, 2000.

[77] B. P. Kramer, A. U. Viretta, M. D. El Baba, D. Aubel, W. Weber, and M. Fussenegger. An engineered epigenetic transgene switch in mammalian cells. *Nat. Biotechnol.*, 22(7):867–870, 2004.

[78] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 295:853–858, 2001.

[79] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in *caenorhabditis elegans*. *Science*, 295:858–862, 2001.

[80] R. C. Lee and V. Ambros. An extensive class of small RNAs in *caenorhabditis elegans* complementarity to lin-14. *Science*, 295:862–864, 2001.

[81] R. C. Lee, R. L. Feinbaum, and V. Ambros. The *c. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854, 1993.

[82] B. Li, M. Carey, and J. L. Workman. The role of chromatin during transcritpion. *Cell*, 128:707–719, 2007.

[83] R. Losick and C. Desplan. Stochasticity and cell fate. *Science*, 320(5872):65–68, 2008. Losick, Richard Desplan, Claude.

[84] G. J. Lyon, J. S. Wright, A. Christopoulos, R. P. Novick, and T. W. Muir. Reversible and specific extracellular antagonism of receptor-histidine kinase signaling. *J. Biol. Chem.*, 277(8):6247–6253, 2002.

[85] T. P. Malan, A. Kolb, H. Buc, and W. R. McClure. Mechanism of CRP-cAMP activation of *lac* operon transcription initiation activation of the P1 promoter. *J. Mol. Biol.*, 180:881–909, 1984.

[86] P. C. Maloney and B. Rotman. Distribution of suboptimally induced beta-d-galactosidase in *Escherichia coli*. the enzyme content of individual cell. *J. Mol. Biol.*, 73:77–91, 1973.

[87] M. Manefield and S. L. Turner. Quorum sensing in context: Out of molecular biology and into microbial ecology. *Microbiology*, 148:3762–3764, 2002.

[88] J. F. Marko and E. D. Siggia. Statistical mechanics of supercoiled DNA. *Phys. Rev. E*, 52:2912–2938, 1995.

[89] W. R. McClure. Mechanisms and control of transcription initiation in prokaryotes. *Ann. Rev. Biochem.*, 54:171–204, 1985.

[90] T. W. McKeithan. Kinetic proofreading in t-cell receptor signal transduction. *Proc. Natl. Acad. Sci. USA*, 92:5042–5046, 1995.

[91] D. McMillen, N. Kopell, J. Hasty, and J. J. Collins. Synchronizing genetic relaxation oscillators by intercell signaling. *Proc. Natl. Acad. Sci. USA*, 99(2):679–684, 2002.

[92] Y. X. Mejia, H. B. Mao, N. R. Forde, and C. Bustamante. Thermal probing of *e. coli* RNA polymerase off-pathway mechanisms. *J. Mol. Biol.*, 382(3):628–637, 2008.

[93] F. Monod and F. Jacob. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, 1961.

[94] J. J. Mote and D. Reines. Recognition of a human arrest site is conserved between RNA polymerase II and prokaryotic RNA polymerases. *J. Biol. Chem.*, 273:16843–16852, 1998.

[95] M. J. Munoz, M. S. P. Santangelo, M. P. Paronetto, M. de la Mata, F. Pelisch, S. Boireau, K. Glover-Cutter, C. Ben-Dov, M. Blaustein, J. J. Lozano, G. Bird, D. Bentley, E. Bertrand, and A. R. Komblihtt. Dna damage regulates alternative splicing through inhibition of RNA polymerase II elongation. *Cell*, 137(4):708–720, 2009.

[96] N. Muzyczka, M. J. Bessman, and R. L. Poland. Studies on biochemical basis of spontaneous mutation .1. comparison of deoxyribonucleic acid polymerases of mutator, antimutator, and wild-type strains of bacteriophage-T4. *J. Biol. Chem.*, 247(22):7116, 1972.

[97] M. M. Nakano and P. Zuber. The primary role of comA in establishment of the competent state in *bacillus subtilis* is to activate expression of srfA. *J. Bacteriol.*, 173(22):7269–7274, 1991.

[98] K. H. Nealson and J. W. Hastings. Bacterial bioluminescence - its control and ecological significance. *Microbiol. Rev.*, 43(4):496–518, 1979.

[99] K. C. Neuman, E. A. Abbondanzieri, R. Landick, J. Gelles, and S. M. Block. Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking. *Cell*, 115:437–447, 2003.

[100] B. E. Nickels and A. Hochschild. Regulation of RNA polymerase through the secondary channel. *Cell*, 118:281–284, 2004.

[101] J. Ninio. Kinetic amplification of enzyme discrimination. *Biochimie*, 57:587–595, 1975.

[102] A. Novick and M. Weiner. Enzyme induction as an all-or-none phenomenon. *Proc. Natl. Acad. Sci. USA*, 43:553–566, 1957.

[103] R. P. Novick. Autoinduction and signal transduction in the regulation of staphylococcal virulence. *Mol. Microbiol.*, 48(6):1429–1449, 2003.

[104] E. Nudler, A. Mustaev, E. Lukhtanov, and A. Goldfarb. The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell*, 89:33–41, 1997.

[105] F. Oberhettinger and L. Badii. *Tables of Laplace transforms*. Springer-Verlag, Berlin, Germany, 1973.

[106] A. Papoulis. *Probability, random variables, and stochastic Processes*. McGraw-Hill, New York, 1965.

[107] J. Paulsson. Summing up the noise in gene networks. *Nature*, 427:415–418, 2004.

[108] J. Paulsson. Models of stochastic gene expression. *Phys. Life Rev.*, 2:157–175, 2005.

[109] J. M. Pedraza and A. van Oudenaarden. Noise propagation in gene networks. *Science*, 307:1965–1969, 2005.

[110] G. Pesole. What is a gene? An updated operational definition. *Gene*, 417(1-2):1–4, 2008.

[111] A. Pikovsky, A. Zaikin, and M. A. de la Casa. System size resonance in coupled noisy systems and in the Ising model. *Phys. Rev. Lett.*, 88(5):4, 2002.

[112] M. Ptashne. *A genetic switch*. Cold spring harbor laboratory press, Cold spring harbor, New York, 3nd edition, 2004.

[113] P. E. M. Purnick and R Weiss. The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.*, 10(6):410–422, 2009.

[114] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and Tyagi S. Stochastic mRNA synthesis in mammalian cells. *PLOS Biol.*, 4:1707–1719, 2006.

[115] A. Raj and A. van Oudenaarden. Single-molecule approaches to stochastic gene expression. *Annu. Rev. Biophys*, 38:255–270, 2009.

[116] S. Redner. *A Guide to First-Passage Processes*. Cambridge University Press, 2001.

[117] K. Rinaudo, L. Bleris, R. Maddamsetti, S. Subramanian, R. Weiss, and Y. Benenson. A universal RNAi-based logic evaluator that operates in mammalian cells. *Nat. Biotechnol.*, 25(7):795–801, 2007.

[118] H. Risken. *The Fokker-Planck equation: methods of solutions and applications*. Springer-Verlag, Berlin, Germany, 2nd edition, 1989.

[119] A. Saunders, L. J. Core, and J. T. Lis. Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.*, 7(8):557–567, 2006.

[120] D. F. Savage, J. Way, and P. A. Silver. Defossiling fuel: How synthetic biology can transform biofuel production. *ACS Chem. Biol.*, 3(1):13–16, 2008.

[121] D. A. Schafer, J. Gelles, Michael P. Sheetz, and R. Landick. Transcription by single molecules of RNA polymerase observed by light microscopy. *Nature*, 352:444–448, 1991.

[122] E. Schrödinger. *What is Life?* Cambridge University Press, 1944.

[123] P. C. Seed, L. Passador, and B. H. Iglewski. Activation of the *pseudomonas aeruginosa lasi* gene by LasR and the *pseudomonas* autoinducer PAI: An autoinduction regulatory hierarchy. *J. Bacteriol.*, 177(3):654–659, 1995.

[124] J. W. Shaevitz, E. A. Abbondanzieri, R Landick, and S. M. Block. Backtracking by single RNA polymerase molecules observed at near base pair resolution. *Nature*, 426:684–687, 2003.

[125] W. Y. Shou, S. Ram, and J. M. G. Vilar. Synthetic cooperation in engineered yeast populations. *Proc. Natl. Acad. Sci. USA*, 104(6):1877–1882, 2007.

[126] D. M. Sitnikov, J. B. Schineller, and T. O. Baldwin. Transcriptional regulation of bioluminesence genes from *vibrio fischeri*. *Mol. Microbiol.*, 17(5):801–812, 1995. 84.

[127] G. M. Skinner, C. G. Baumannand D. M. Quinn, J. E. Molloy, and J. G. Hoggett. Promoter binding, initiation, and elongation by bacteriophage T7 RNA polymerase. *J. Biol. Chem.*, 279:3239–3244, 2004.

[128] Ekaterina Sosunova, Vasily Sosunov, Maxim Kozlov, Vadim Nikiforov, Alex Goldfarb, and Arkady Mustaev. Donation of catalytic residues to RNA polymerase active center by transcription factor Gre. *Proc. Natl. Acad. Sci. USA*, 100(26):15469–15474, December 2003.

[129] A. M. Stevens, K. M. Dolan, and E. P. Greenberg. Synergistic binding of the *vibrio fischeri* LuxR transcriptional activator domain and RNA polymerase to the *lux* promoter region. *Proc. Natl. Acad. Sci. USA*, 91(26):12619–12623, 1994.

[130] J. Stricker, S. Cookson, M. R. Bennett, W. H. Mather, L. S. Tsimring, and J. Hasty. A fast, robust and tunable synthetic gene oscillator. *Nature*, 456(7221):516–U39, 2008.

[131] L. Strmecki, D. M. Greene, and C. J. Pears. Developmental decisions in *dictyostelium discoideum*. *Dev. Biol.*, 284(1):25–36, 2005.

[132] S. H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Westview Press, Cambridge, MA, 2000.

[133] P. S. Swain, M. B. Elowitz, and E. D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. USA*, 99(20):12795–12800, 2002.

[134] P. S. Swain and E. D. Siggia. The role of proofreading in signal transduction pathways. *Biophys. J.*, 82:2928–2933, 2002.

[135] I. A. Swinburne, D. G. Miguez, D. Landgraf, and P. A. Silver. Intron length increases oscillatory periods of gene expression in animal cells. *Genes Dev.*, 22(17):2342–2346, 2008.

[136] J. F. Sydow, F. Brueckner, A. C. M. Cheung, G. E. Damsma, S. Dengl, E. Lehmann, D. Vassylyev, and P. Cramer. Structural basis of transcription: Mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. *Mol. Cell*, 34(6):710–721, 2009.

[137] V. R. Tadigotla, D. O. Maoiléidigh, A. M. Sengupta, V. Epshtein, R. H. Ebright, E. Nudler, and A. E. Ruckenstein. Thermodynamic and kinetic modeling of transcription pausing. *Proc. Natl. Acad. Sci. USA*, 103:4439–4444, 2006.

[138] M. J. Thomas, A. A. Platas, and D. K. Hawley. Transcriptional fidelity and proofreading by RNA polymerase II. *Cell*, 93:627–637, 1998.

[139] M. Tigges, T. T. Marquez-Lago, J. Stelling, and M. Fussenegger. A tunable synthetic mammalian oscillator. *Nature*, 457(7227):309–312, 2009.

[140] E. Ullner, A. Zaikin, E. I. Volkov, and J. Garcia-Ojalvo. Multistability and clustering in a population of synthetic genetic oscillators via phase-repulsive cell-to-cell communication. *Phys. Rev. Lett.*, 99(14):4, 2007.

[141] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 1992.

[142] J. S. van Zon, M. J. Morelli, S. T. Tanase, and P. Rein ten Wolde. Diffusion of transciption factors can drastically enhance the noise in gene expression. *Biophys. J.*, 91:4350–4367, 2006.

[143] D. G. Vassylyev, M. N. Vassylyeva, J. W. Zhang, M. Palangat, I. Artsimovitch, and R. Landick. Structural basis for substrate loading in bacterial RNA polymerase. *Nature*, 448(7150):163, 2007.

[144] K. L. Visick, J. Foster, J. Doino, M. McFall-Ngai, and E. G. Ruby. *Vibrio fischeri* lux genes play an important role in colonization and development of the host light organ. *J. Bacteriol.*, 182(16):4578–4586, 2000.

[145] P. H. Von Hippel and O. G. Berg. Facilitated target location in biological-systems. *J. Biol. Chem.*, 264(2):675–678, 1989.

[146] P. H. von Hippel and T. D. Yager. Transcription elongation and termination are competitive kinetic processes. *Proc. Natl. Acad. Sci. USA*, 88:2307–2311, 1991.

[147] D. Wang, D. A. Bushnell, X. H. Huang, K. D. Westover, M. Levitt, and R. D. Kornberg. Structural basis of transcription: Backtracked RNA polymerase II at $3.4$ angstrom resolution. *Science*, 324(5931):1203–1206, 2009.

[148] J. W. Wang, J. J. Zhang, Z. J. Yuan, and T. S. Zhou. Noise-induced switches in network systems of the genetic toggle switch. *BMC Syst. Biol.*, 1:14, 2007.

[149] W. D. Wang, Z. T. Chen, B. G. Kang, and R. Li. Construction of an artificial intercellular communication network using the nitric oxide signaling elements in mammalian cells. *Exp. Cell Res.*, 314(4):699–706, 2008.

[150] C. M. Waters and B. L. Bassler. Quorum sensing: Cell-to-cell communication in bacteria. *Annu. Rev. Cell Dev. Biol.*, 21:319–346, 2005.

[151] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids - a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[152] A. T. Willingham, A. P. Orth, S. Batalov, E. C. Peters, B. G. Wen, P. Aza-Blanc, J. B. Hogenesch, and P. G. Schultz. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science*, 309(5740):1570–1573, 2005.

[153] M. N. Win and C. D. Smolke. Higher-order cellular information processing with synthetic RNA devices. *Science*, 322(5900):456–460, 2008.

[154] X. S. Xie, P. J. Choi, G. W. Li, N. K. Lee, and G. Lia. Single-molecule approach to molecular biology in living bacterial cells. *Annu. Rev. Biophys*, 37:417–444, 2008.

[155] Jie Yan, Marcelo O. Magnasco, and John F. Marko. A kinetic proofreading mechanism for disentanglement of DNA by topoisomerases. *Nature*, 401:932–935, 1999.

[156] H. Yin, M. D. Wang, K. Svodoba, R. Landick, and S. M. Block. Transcription against an applied force. *Science*, 270:1653–1657, 1995.

[157] L. C. You, R. S. Cox, R. Weiss, and F. H. Arnold. Programmed population control by cell-cell communication and regulated killing. *Nature*, 428(6985):868–871, 2004.

[158] J. Yu, J. Xiao, X. Ren, K. Lao, and X. S. Xie. Probing gene expression in live cells, one protein molecule at a time. *Sience*, 311:1600–1603, 2006.

[159] N. Zenkin, Y. Yuzenkova, and K. Severinov. Transcript-assisted transcriptional proofreading. *Science*, 313:518–520, 2006.

[160] X. Zhang, T. Reeder, and R. Schleif. Transcription activation parameters at ara pBAD. *J. Mol. Biol.*, 258:14–28, 1996.